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

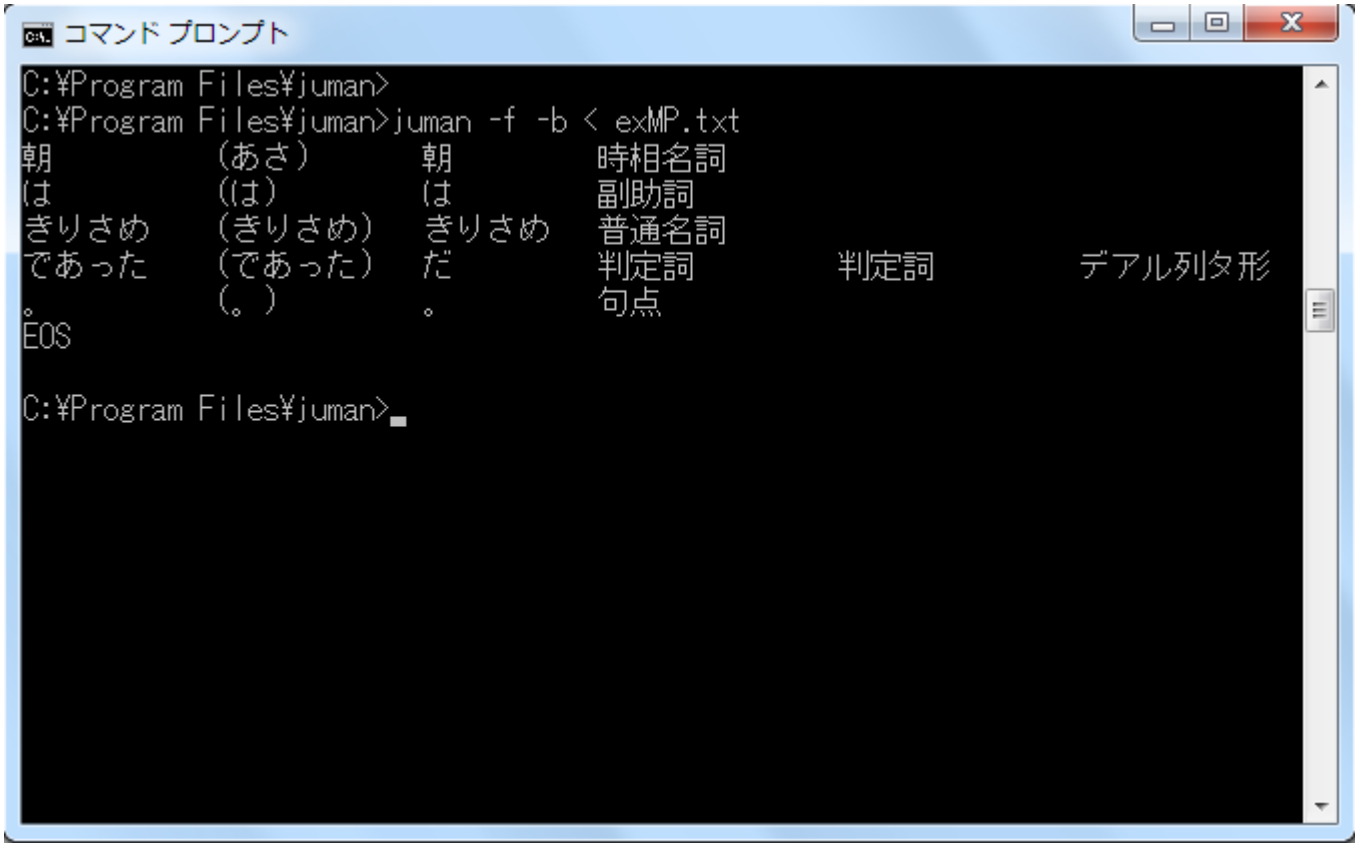
- 形態素解析
  - 文を単語ごとに分割し、品詞情報などを付け加える作業

例：花が咲く

「花」→名詞、「が」→助詞、「咲く」→動詞

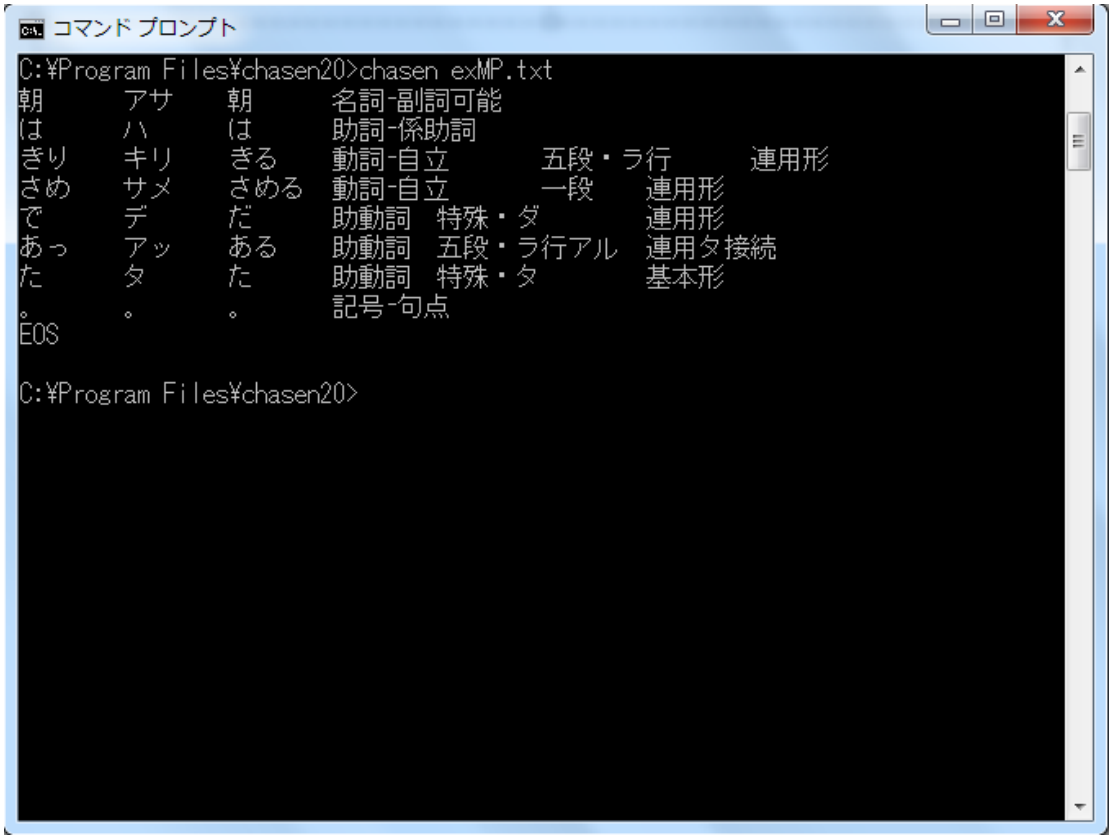
JUMAN

- 長尾真氏の研究室を中心に開発され、1992年に公開された



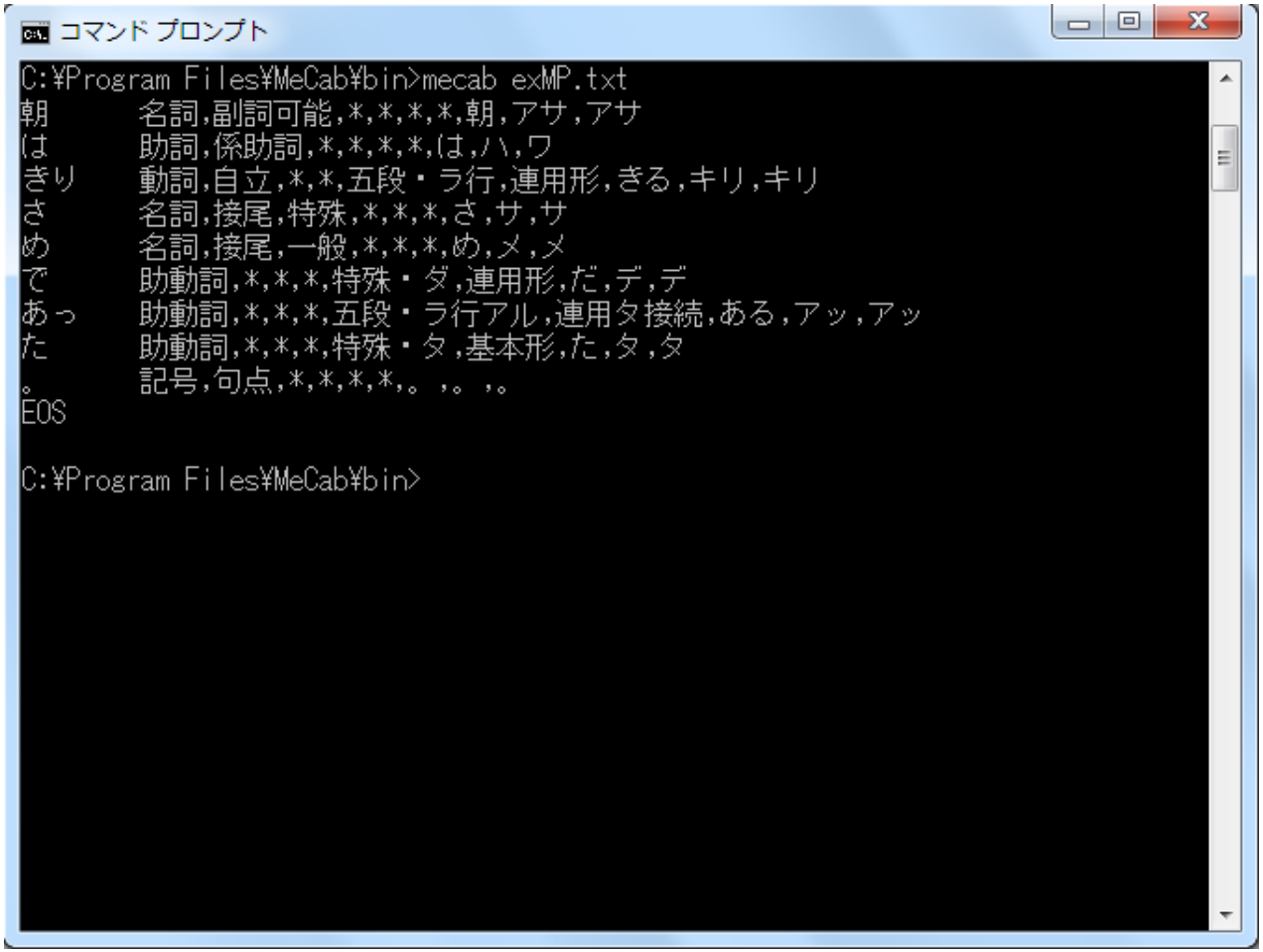
茶筌（ChaSen）

- 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



MeCab

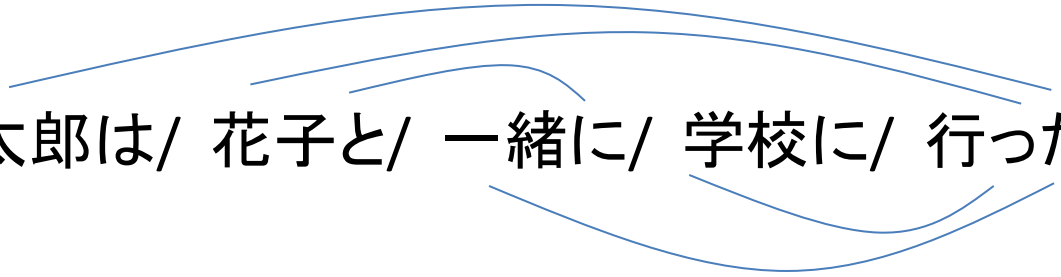
- 工藤拓氏が開発し、2002年に公開



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

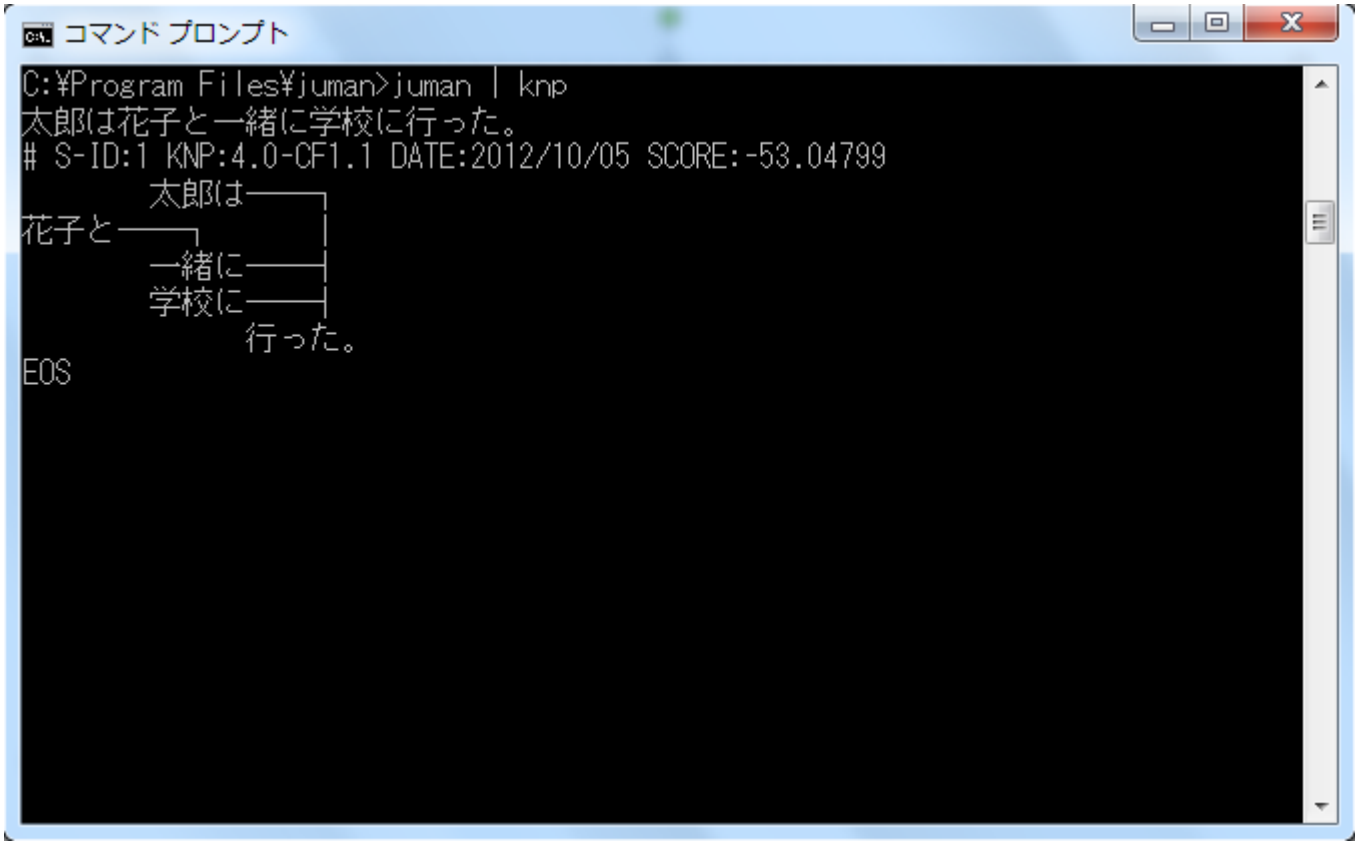
例：太郎は花子と一緒に学校に行った。



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



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