短単位・長単位データマニュアル

ver.1.0 (2004-03-24)

山口昌也，小椋秀樹，西川賢哉，石塚京子，木村睦子(国立国語研究所)
内元清貴(情報通信研究機構)

目次

---

1. はじめに

本マニュアルでは，短単位・長単位データベースについて解説する。短単位・長単位データベースは，転記テキストを短単位，長単位に分割し，それらに対して形態論的な情報を付与したものである。ここでは，主として，データ形式の説明を行う。転記テキストの仕様，および，形態論的情報の分析方法については，次のマニュアルを参照のこと。

- 短単位・長単位：「『日本語話し言葉コーパス』の形態論情報の概要」(pos.pdf)
- 転記テキスト：「転記テキストの仕様」(transcription.pdf)

なお，本データベースの内容は，XML 形式の『日本語話し言葉コーパス』から形態論情報を取り出し，KWIC 付きのタブ区切りテキストに整形したものである。音声関連のデータなど，本データベースに含まれていないデータと組み合わせて利用したい場合は，XML 形式の『日本語話し言葉コーパス』をお使いいただきたい。詳細は，「『日本語話し言葉コーパス』XML文書について」(xml.pdf) を参照のこと。

↑

2. 収録内容

- 人手解析データ
  - 人手で短単位・長単位解析を行ったデータ
  - 品詞体系は，「『日本語話し言葉コーパス』の形態論情報の概要」に準ずる。
- 自動解析データ
  - 自動で短単位・長単位解析を行い，部分的に人手修正したデータ
  - 品詞体系は，「『日本語話し言葉コーパス』の形態論情報の概要」の品詞体系に対して，活用の種類，活用形が細分化されている。詳細は，本マニュアル5節を参照のこと。

3. データ形式

3.1 ファイル形式とファイル名

- ファイル形式
  - タブ区切りのテキストファイル
  - 漢字コード: シフト JIS
  - 改行コード: CR + LF
- ファイル名
  - ファイル名は，本体8文字，拡張子3文字からなる。
  - 同名のファイル名本体を持つ転記テキストと対応する。
  - 拡張子
    - sdb : 短単位・長単位混合形式
    - ldb : 長単位形式

3.2 短単位・長単位混合形式

概要

- 1行が1短単位の情報を含む。
- 長単位の情報は，長単位の先頭を構成する短単位に付随する。先頭以外の短単位には，長単位の情報は付かず，空欄となる。次の例は，「日本語の品詞体系」に対して，短単位と長単位の情報を付与したものである。

短単位の情報			長単位の情報		
代表表記[短]	品詞[短]	その他の情報1[短]	代表表記[長]	品詞[長]	その他の情報1[長]
日本	名詞	固有名詞	日本語	名詞	
語	名詞				
の	助詞	格助詞	の	助詞	格助詞
文法	名詞		文法体系	名詞	
体系	名詞				
は	助詞	係助詞	は	助詞	係助詞

実例

- 「音響モデルを」に対する情報付与例
- 短単位は「音響|モデル|を」の三つに，長単位は「音響モデル|を」の二つに分割される。
- 実際のデータでは，各フィールドはタブで区切られるが，便宜上 / で表記した。また，# で始まる行は，本マニュアルにおけるコメントである。

```
# 「音響」
00000001/00000002/A01M0065/0017 00041.518-00044.572 L:-005-001/
まし た (F え) 最後 に まとめ ます 〈雑音〉 (F えー) 高 精度 で (F えー) 頑健 な/
音響/
モデル を (F おー) 目標 と し て (F えー) 音響 モデル の 研究 行なわ れ て/
音響/オンキョウ/音響/オンキョー
名詞//////
名詞//////
オンキョウモデル/音響モデル

# 「モデル」
00000002/00000003/A01M0065/0017 00041.518-00044.572 L:-005-005/
た (F え) 最後 に まとめ ます 〈雑音〉 (F えー) 高 精度 で (F えー) 頑健 な 音響/
モデル/
を (F おー) 目標 と し て (F えー) 音響 モデル の 研究 行なわ れ て い/
モデル/モデル/モデル/モデル/
名詞//////
//////
//

# 「を」
00000003/00000004/A01M0065/0018 00045.060-00047.337 L:-001-001/
(F え) 最後 に まとめ ます 〈雑音〉 (F えー) 高 精度 で (F えー) 頑健 な 音響 モデル/
を/
(F おー) 目標 と し て (F えー) 音響 モデル の 研究 行なわ れ て い ます/
を/ヲ/を/オ/
助詞///格助詞///
助詞///格助詞///
ヲ/を/
```

各フィールドの説明

フィールド番号	フィールド名	内容
1	ID	当該短単位の通し番号(8桁)
2	後続ID	後続する短単位のID(後続する短単位が存在しない場合は，00000000)
3	講演ID	当該短単位が収録されている転記テキストの講演ID
4	転記情報	当該短単位を含む転記単位のタイムスタンプなど(4.1 節参照)
5	前文脈	当該単位に先行する文脈(最大15短単位)
6	出現形	当該短単位の転記テキスト(基本形)における出現語形
7	後文脈	当該単位に後続する文脈(最大15短単位)
8	タグなし出現形	出現形から転記テキストのタグを取り除いたもの
9	代表形	出現形の標準的な語形(国語辞典の見出しに相当)
10	代表表記	代表形を漢字，仮名などで表記したもの
11	発音形	当該短単位の発音形(転記テキストの発音形に相当)
12	品詞	当該短単位の品詞
13	活用の種類	当該短単位の活用の種類(「カ行五段」等)
14	活用形	当該短単位の活用形(「連用形」等)
15	その他の情報1	品詞の下位分類(「助詞」の下位分類として「格助詞」等)
16	その他の情報2	語形の情報(「促音便」等)
17	その他の情報3	「言いよどみ」「メタ」等の補足情報(複数情報がある場合は，全角スペースで区切る)

フィールド番号	フィールド名	内容
18	品詞[長]	長単位の品詞
19	活用の種類[長]	長単位の活用の種類
20	活用形[長]	長単位の活用形
21	その他の情報1[長]	長単位のその他の情報1
22	その他の情報2[長]	長単位のその他の情報2
23	その他の情報3[長]	長単位のその他の情報3
24	代表形[長]	長単位の代表形
25	代表表記[長]	長単位の代表表記

3.3 長単位形式

概要

- 1行が1長単位の情報を含む。
- 長単位形式データに含まれる長単位情報は，短単位・長単位混合形式における長単位情報と同一である。
- 長単位形式は，主として，長単位に則した前文脈，後文脈を参照できるように用意したものである。

実例

- 「音響モデルを」に対する情報付与例
- 「音響モデル|を」の二つの長単位に分割される。
- 実際のデータでは，各フィールドはタブで区切られるが，便宜上 / で表記した。また，# で始まる行は，本マニュアルにおけるコメントである。

```
# 「音響モデル」
00000001/00000002/A01M0065/0017 00041.518-00044.572 L:-005-001/
比較し ました （F え） 最後 に まとめ ます〈雑音〉（F えー） 高精度 で （F えー） 頑健 な/
音響モデル/
を （F おー） 目標 と し て （F えー） 音響モデル の 研究 行なわ れ てい ます/
音響モデル/オンキョウモデル/音響モデル/オンキョーモデル
名詞//////

# 「を」
00000002/00000003/A01M0065/0018 00045.060-00047.337 L:-001-001/
ました （F え） 最後 に まとめ ます〈雑音〉（F えー） 高精度 で （F えー） 頑健 な 音響モデル/
を/
（F おー） 目標 と し て （F えー） 音響モデル の 研究 行なわ れ てい ます が/
を/ヲ/を/オ/
助詞///格助詞///
```

各フィールドの説明

フィールド番号	フィールド名	内容
1	ID	当該長単位の通し番号(8桁)
2	後続ID	後続する長単位のID(後続する長単位が存在しない場合は，00000000)
3	講演ID	当該長単位が収録されている転記テキストの講演ID
4	転記情報	当該長単位を含む転記単位のタイムスタンプなど(4.1 節参照)
5	前文脈	当該単位に先行する文脈(最大15短単位)

フィールド番号	フィールド名	内容
6	出現形	当該長単位の転記テキスト(基本形)における出現語形
7	後文脈	当該単位に後続する文脈(最大15短単位)
8	タグなし出現形	出現形から転記テキストのタグを取り除いたもの
9	代表形	出現形の標準的な語形(国語辞典の見出しに相当)
10	代表表記	代表形を漢字，仮名などで表記したもの
11	発音形	当該長単位の発音形(転記テキストの発音形に相当)
12	品詞	当該長単位の品詞
13	活用の種類	当該長単位の活用の種類(「カ行五段」等)
14	活用形	当該長単位の活用形(「連用形」等)
15	その他の情報1	品詞の下位分類(「助詞」の下位分類として「格助詞」等)
16	その他の情報2	語形の情報(「促音便」等)
17	その他の情報3	「言いよどみ」「メタ」等の補足情報(複数情報がある場合は，全角スペースで区切る)

4. 転記テキストとの関係

4.1 「転記情報」フィールド

- 当該単位と転記テキストは，「講演ID」フィールドと「転記情報」フィールドの組合せで対応づけられる。
- 「転記情報」フィールドは，転記テキストにおけるタイムスタンプに，短単位の位置情報を付加したものである。

- 形式

```
発話ID タイムスタンプ 単位位置情報
```

  - 発話ID: 当該短・長単位を含む転記基本単位の通し番号
  - タイムスタンプ: その転記基本単位の開始時刻・終了時刻
  - 短単位位置情報: 転記基本単位の先頭からの行数，および，各行における先頭からのバイト数(転記テキストの基本形を基準とする。文字コードは，シフトJIS)

- 実例
  - 転記テキスト

```
0017 00051.048-00056.945 L:
日本語の                        & ニホンゴノ
文法は                          & ブンポーワ
0018 00057.439-00061.747 L:
従来の                          & ジューライノ
```

  - 短単位・長単位混合形式(転記情報と出現形)

```
0017 00051.048-00056.945 L:-001-001     日本
0017 00051.048-00056.945 L:-001-005     語
0017 00051.048-00056.945 L:-001-007     の
0017 00051.048-00056.945 L:-002-001     文法
0017 00051.048-00056.945 L:-002-005     は
0018 00057.439-00061.747 L:-001-001     従来
0018 00057.439-00061.747 L:-001-005     の
```

4.2 非言語音タグ(＜雑音＞，＜ベル＞など)の扱い

- 非言語音タグのうち，＜雑音＞，＜ベル＞など，一つの転記単位全体に対して付与されているものは，便宜上，一つの短単位として扱う。
- ただし，「出現形」，「発音形」フィールドに当該タグが入るだけで，「代表形」，「代表表記」，「品詞」など，短単位，長単位に関する情報は，付与しない。
- 非転記タグ ＜雑音＞ の例
    - 転記テキスト

```
0202 00498.324-00501.003 L:
コーパスの                    & コーパスノ
0203 00501.163-00502.587 L:〈雑音〉
0204 00503.031-00503.812 L:
内容は                        & ナイヨーワ
```

    - 短単位・長単位混合形式

転記情報	出現形	発音形	品詞	その他の情報1
0202 00500.324-00501.003 L:-001-001	コーパス	コーパス	名詞	
0202 00500.324-00501.003 L:-001-009	の	ノ	助詞	格助詞
0203 00501.163-00502.587 L:-001-001	＜雑音＞	＜雑音＞		
0204 00503.031-00503.812 L:-001-001	内容	ナイヨー	名詞	
0204 00503.031-00503.812 L:-001-005	は	ワ	助詞	係助詞

4.3 R タグの扱い

個人名，差別語，誹謗中傷などにマークアップされる R タグ(例：(R ××)さん，(R ×××)教授)は，短単位・長単位データベース中では，次のように扱われる。なお，R タグの詳細は，「転記テキストの仕様」(transcription.pdf)のタグ (R) を参照のこと。

- 出現形，発音形は，転記テキストに準ずる。
- 出現形に R タグを含む短単位は，次のフィールドを「×」で伏字する。
    - タグなし出現形
    - 代表形
    - 代表表記
- 伏字処理をされた短単位を構成要素として持つ長単位は，代表形，代表表記を次のように伏字処理する。
    - 人手解析データの場合：伏字処理されている短単位に該当する部分を「×」で伏字する。
    - 自動解析データの場合：代表形，代表表記全体を「×」で伏字する。

例：「(R 山田)さん」の場合

	代表形	代表表記
人手解析データ	×××サン	××さん
自動解析データ	×××××	××××
伏字前のデータ	ヤマダサン	山田さん

5. 人手解析データと自動解析データとの相違

人手解析データと自動解析データの品詞体系の違いは，次のとおりである。

- 短単位の活用の種類と活用形
    - 自動解析データにおける，短単位の活用の種類，活用形は，「短単位辞書マニュアル」3節の活用表に従って細分類されている。ただし，上一段活用に関しては，人手解析データとの違いはない。
    - 細分類の内容は，次のとおりである。
        - 後続する短単位により，未然形，連用形を細分化した。「未然形1」のように1～4の数字で細分化を表示する。この数字を除去したものが人手解析データの活用形に対応する。
        - 活用型は，「カ行五段1」，「カ行五段2」といった形式で，細分化している。活用形と同様，末尾の数字を除去したものが人手解析データの活用型に対応する(なお，人手解析データでも細分化されている「文語形容詞型1～3」は除く)。
- 長単位の R タグの伏字処理(4.3 節参照)。

自動解析データでは，人手修正済データに比べて，次のような誤りが多い。

- 連体形と終止形の間違い
- 助動詞と助詞の間違い(主に，「で」と「に」)
- 格助詞と準体助詞の間違い(主に，「の」)
- F タグで囲まれた短単位の解析誤り(主に，F タグ内が複数の短単位から構成される場合)
- D2 タグで囲まれた短単位の解析誤り

自動解析の方法については，次の文献を参照されたい。

- 内元，高岡，野畑，山田，関根，井佐原：“『日本語話し言葉コーパス』への形態素情報付与”，第3回話し言葉の科学と工学ワークショップ講演予稿集， pp.39-46 (2004)
