日本語表記ゆれ辞書「たんし」

Ver. 1.1.0

はじめに

この表記ゆれ辞書「たんし」は、NAIST Japanese Dictionary[i](mecab-naist-jdic-0.4.3-20080917) を、表記ゆれ辞書として使用できるように変換したものである。単語、その読み、その品詞、活用形ならびに表記ゆれと見なされる単語群で構成されている。

辞書の形式

文字コード:UTF-8(BOM なし)
改行コード:CR/LF

辞書はタブ区切りのテキストファイル(TSV)で保存してある。各項目は国立国語研究所の「表記統合辞書[ii]」に準ずる。

- 第 1 フィールド:『NAIST Japanese Dictionary』の「見出し語」
- 第 2 フィールド:『NAIST Japanese Dictionary』の「当該見出し語の読み」
- 第 3 フィールド:『NAIST Japanese Dictionary』の「当該見出し語の発音」
- 第 4 フィールド:『NAIST Japanese Dictionary』の「当該見出し語の品詞名」
- 第 5 フィールド:『NAIST Japanese Dictionary』の「当該見出し語の活用型」
- 第 6 フィールド: 当該の「見出し語」に対して，同語と判断された「見出し語」のリスト

ただし、第 4 フィールドは省略する場合がある。また、以下の部分が「表記統合辞書」とは異なる。

- 辞書の並びが見出し語の五十音順ではない。
- 「表記統合辞書」では、

  行きあう　　{イ/ユ}キアウ

  のように読みがほぼ同一と思われる部分を統合してあるが、「たんし」では分けている。

表記ゆれと見なす基準

「表記統合辞書」に準ずる。

『NAIST Japanese Dictionary』のうち、

- 品詞、活用形が同一
- 発音が同一
- 記号、地名と組織以外の固有名詞を除いたもの

であり、以下の項目のいずれかに合致するものを表記ゆれとした。

- 送り仮名による違い

- 送り仮名の有無

  割付　ワリツケ　　名詞-一般　　割付/わりつけ/割り付け/割付け

- 促音，撥音の有無

  真裸　マッパダカ　　名詞-一般　　真裸/真っ裸/まっ裸

- 字種による違い
  - ひらがな，カタカナ，漢字

    鬱金香　　ウッコンコウ　名詞-一般　　鬱金香/ウッコンコウ/うっこんこう

  - 一般名詞，および，数詞における漢数字，アラビア数字，ローマ数字

    1　　イチ　名詞-数　　1/一/1/いち

  - アルファベット表記とカタカナ表記

    総 t 数 ソウトンスウ　名詞-一般　　総 t 数/総トン数

  - 「ヵ」，「か」，「カ」，「ヶ」，「ケ」，「箇」，「個」

    ヵ村　カソン 名詞-接尾-助数詞　　ヵ村/カ村/か村/ケ村/ヶ村

  - アルファベットの大文字・小文字(2文字以上の形態素の場合)

    エヌジー　　エヌジー　　名詞-一般　　エヌジー/NG

  - 名詞-接尾-助数詞で同一の単位における字種

    キロメートル　キロメートル　名詞-接尾-助数詞　　キロメートル/km/粁/kメートル

- 記号類による違い
  - 読点・中黒の違い，読点・中黒の有無

    小・中学校　ショウチュウガッコウ　名詞-一般　　小・中学校/小、中学校/小中学校

  - 「々」，「ゝ」などの踊り字の種類，有無

    シバシバ　　シバシバ　　副詞-一般　　シバシバ/屡/屡々/しばしば

TODO

- カタカナ語の異表記ゆれ

  例えば、
  - 長音の有無(例:「サーバー」,「サーバ」)

  がある。しかし、「長音の有無」を安易に「表記ゆれと見なす基準」に含めた場合、以下の語が表記ゆれと見なされてしまう。
  - 「ウェブ(「World Wide Web」の略語)、「ウェーブ(波)」[iii]
  - 「シチュー(料理)」、「シチュ(「シチュエーション」の略語)」

  よって長音に関しては、表記的な基準でなく意味的な基準を設ける必要があると考える。

- 漢字の旧字, 異体字(例:暁/曉, 虱/蝨)
- 別語とするべきものの扱い

著作権

ライセンスに関する注記

更新履歴

Ver. 1.1.0　(2009/3/21)

- 発音を辞書に含めることにより、フィールド数を 5 から 6 に変更
- 表記ゆれと見なす基準を変更
  - 「表記が同一」を、「発音が同一」
  - 地名と組織の固有名詞を辞書に含める
- 繰り返し記号のバグを修正

(見出し語数 299,214 語, 表記ゆれ候補数 688,429 語)

Ver. 1.0.0　(2009/3/20)

- 公開開始

(見出し語数 295,921 語, 表記ゆれ候補数 671,519 語)

Koumei_S
koumeism@gmail.com

i 浅原正幸, 松本裕治(奈良先端科学技術大学院大学)
NAIST Japanese Dictionary version 0.4.0 ユーザーズマニュアル
ii 山口昌也, 桐生りか, 田中牧郎(国立国語研究所)
表記統合辞書 利用マニュアル
iii Takeshi Masuyama†, Satoshi Sekine‡, Hiroshi Nakagawa†
Automatic Construction of Japanese KATAKANA Variant List from Large Corpus
COLING 2004 (Proceedings of the 20th International Conference on Computational Linguistics), pp. 1214-1219, Geneva, Switzerland, August 2004

(†University of Tokyo, ‡New York University)