

## Noun.xls

1. タグ付き KY コーパスの全レベルから(ナ形容詞を含む)名詞を抽出し、様々な言語情報を付与したデータ
2. レコードの説明
  - A列:ID番号
  - B列:単語の発音。カタカナで表記
  - C列:単語の表記。カナ交じり表記
  - D列:拍
  - E列:語種情報。国立国語研究所の「日本語コーパス」言語政策班が作成した語彙データ「Lexicon2008\_BookCorpus」から自動抽出したもの。自動抽出できなかったものは、人手で入力。
  - F～G列:茶筌に基づく品詞情報
  - H列:対象語彙がもっとも最初に発せられたレベル(例えば、「弟」は初級、「踊り子」は中級など)
  - I～M列:各レベルにおける正用としての延べ頻度
  - N～R列:各レベルにおける誤用としての延べ頻度
  - S～U列:各語彙の単語親密度<sup>1</sup>。単語親密度はNTTが開発したNTTデータベースシリーズ「日本語の語彙特性」から自動で抽出している。詳細は(<http://www.kecl.ntt.co.jp/mtg/goitokusei/>)。例えば、超級で出現している「反駁」という単語の単語親密度は 2.438 となっており、認知度が低い難易語であることがわかる。一方、初級で出てくる「食べ物」や「父」などは単語親密度が 6.6 以上であり、認知度が高い(日本人なら誰もが知っている単語である)。
  - V列:元データは言語政策班が作成した「Lexicon2008\_BookCorpus」から対象の単語の頻度を調査したもの<sup>2</sup>。
  - W列以降:「分類語彙表<sup>3</sup>」(<http://www.kokken.go.jp/kanko/goihyo/>)から語義を書きだしたもの。
    1. W列:語義の数。分類語彙表に登録されている語義の数。1 は語義が 1 つしかない語(単義語)であることを意味し、2 以上は複数の語義があること(多義語)であることを意味する。
    2. X列:最初の語義の部門。生産物や主体など。
    3. Y列:分類語彙表・中項目。X列をより細かく分類したもの
    4. Z列:分類語彙表・分類項目。Y列をさらに細かく分類したもの
    5. AA列:XからZを一つのセルにまとめたもの
    6. AB～AK列:語義が二つある場合は、二つ目以降の語義を各セルに羅列。ついでに、もっとも語義が多い語は「手」で全部で 11 の語義がある。

<sup>1</sup> 単語親密度(word familiarity)とは、ある単語がどの程度なじみがあると感じられるかを表した指標です。単語親密度は、ある単語を複数の人に見せたり聞かせたりして、そのなじみの程度を 1 から 7 までの数字(1:なじみがない ― 7:なじみがある)で答えてもらい、その平均をとって求めます。(<http://www.kecl.ntt.co.jp/mtg/goitokusei/example-exps.html>)

<sup>2</sup> 元データは「日本語コーパス」のモニター版で 6000 万語規模のコーパス。書籍 (13587 サンプル)、白書 (1500 サンプル)、Yahoo!知恵袋 (45725)、国会会議録 (159)

<sup>3</sup> 分類語彙表とは、「語を意味によって分類・整理したシソーラス(類義語集)です。昭和 39 年(1964 年)に出版された初版『分類語彙表』(現在は絶版)は、現代日本語の本格的なシソーラスとして幅広く活用されてきました。この度、収録語数を増やした『分類語彙表―増補改訂版―』が刊行されましたが、研究開発用にそのデータベース版を用意しました。本データベース版は、書籍版の『分類語彙表―増補改訂版―』の元となったデータを加工したものです。データベースソフトに取り込めるよう CSV 形式になっています。レコード総数は、101,070 件です(この中には、見出しの併記を分割してできたレコード及び分類項目内の意味的区切りを示すレコードを含みます)。レコードを構成する項目は、次のとおりです。(<http://www.kokken.go.jp/kanko/goihyo/>)

レコードID番号／見出し番号／レコード種別／類／部門／中項目／分類項目

分類番号／段落番号／小段落番号／語番号／見出し／見出し本体／読み／逆読み

001946,01838,A,体,関係,存在,成立,1.1220,14,01,03,国立,国立,こくりつ,つりくこ

030548,29140,A,体,活動,言語,言語,1.3101,03,01,01,国語,国語,こくご,ごっこ

022620,21486,A,体,主体,社会,社寺・学校,1.2630,15,01,01,研究所,研究所,けんきゅうじょ,よじょうゆきんけ

## Verb-Adjective.xls

1. タグ付き KY コーパスの全レベルから動詞と形容詞を抽出し、様々な言語情報を付与したデータ

### 2. レコードの説明

- A列:ID番号
- B列:単語の発音。カタカナで表記
- C列:単語の表記。カナ交じり表記
- D列:KYコーパス内で用いられる異なる表記例
- E列:拍
- F列:茶釜に基づく品詞情報
- G列:対象語彙がもつとも最初に発せられたレベル
- H～L列:各レベルにおける正用としての延べ頻度
- M～Q列:各レベルにおける誤用としての延べ頻度
- R列:各動詞の正答率を計算したもの。計算方法:正用の合計/正用の合計+誤用の合計\*100
- S～U列:各語彙の単語親密度。単語親密度はNTTが開発したNTTデータベースシリーズ「日本語の語彙特性」から自動で抽出している。詳細は(<http://www.kecl.ntt.co.jp/mtg/goitokusei/>)。
- V列:元データは言語政策班が作成した「Lexicon2008\_BookCorpus」から対象の単語の頻度を調査したもの。
- W列以降:「分類語彙表」(<http://www.kokken.go.jp/kanko/goihyo/>)から語義を書きだしたもの。
  1. W列:語義の数。分類語彙表に登録されている語義の数。1 は語義が 1 つしかない語(単義語)であることを意味し、2 以上は複数の語義があること(多義語)であることを意味する。
  2. X列:最初の語義の類
  3. Y列:語義の部門。生産物や主体など。
  4. Z列:分類語彙表・中項目。X列をより細かく分類したもの
  5. AA列:分類語彙表・分類項目。Y列をさらに細かく分類したもの
  6. AB列:XからZを一つのセルにまとめたもの
  7. AC～AL列:語義が二つある場合は、二つ目以降の語義を各セルに羅列。

- **お願い**

1. 本データベースを論文や研究などで使用された際には「李在鎬・浅尾仁彦(2008) KY コーパスの語彙リスト」を使用したことを明記していただければ幸いです。
2. 個人的な利用に限り、改変および複製していただいてもかまいません。ただし、改変したものを再配布することはご遠慮いただきたいと思います。
3. 利用に関して、疑問およびご意見などは李在鎬(jhlee.n@gmail.com)までご連絡ください。

- **謝辞:**

本データベースの公開を許可してくださった鎌田修先生(南山大学)および山内博之先生(実践女子大学)に感謝致します。なお、本データベースの作成においては、次の三つの研究助成を受けております。博報『第二回ことばと文化・教育』研究助成 グループ研究(2007)「定量的手法に基づく日本語の記述的研究:教育的応用の観点から」(研究代表者:李在鎬)、科学研究費補助金(若手・2007~2008)「コーパス分析に基づく認知言語学的構文研究と日本語教育文法へ」(研究代表者:李在鎬)、特定領域研究「日本語コーパス」(研究代表者:前川喜久雄)