日本語 WordNet 同義対データベース 説明書 ver.1.0

1. 概要

本データベースには、日本語の概念辞書である「日本語 WordNet」(version1.1)において、同じ synset(同じ概念を共有する語のまとまり)に掲載されている語を組み合わせて語対とし、人手で同義関係にあると判定された 11,753 対が収録されています。例えば、日本語 WordNet には、synset 00623862-n として、以下のような語が収録されています。

いざこざ、困り者、災い、ご苦労、故障、困難、波乱、もやくや、いさくさ、トラブル

ここから、以下のような語対を作成します。

いざこざ、困り者
いざこざ、災い
いざこざ、ご苦労
いざこざ、故障
いざこざ、困難
…

作成された語対のうち、人手で同義関係にあると判定された語対が本データベースに掲載されています。なお、ALAGIN フォーラムにて公開されている言語資源に含まれている語対、及び、約6億ページの Web データにおいて頻度が 1,500 以下の語などは、同義関係の判定対象としておりません。

2. 同義関係の判定と範囲

本データベースの作成においては、ある文で当該の語が対となる語と言い換えることが可能であると判定される場合、語対は同義関係にあると判定しました。例えば、「ごはん」と「めし」では、「ごはんを食べる」を「めしを食べる」と言い換えることができます。このような関係が語対に成り立つ場合、同義対と判定されています。なお、日本語 WordNet では、異表記や略記にも異なる ID(wordid)が付与されているため、表1に記載されたものも、同義対として本データベースに含めています。

語対の関係	具体例
異表記	お米、おこめ/サーバー、サーバ
略記	メールアドレス,メアド
丁寧語・敬語・接辞(人称) の違い	御本,本/高覧,見る/お母さん,母
方言の違い	かわず,カエル
一般用語と専門用語の違い	リンドウ、ゲンティアナ・ベルナ
日本語と外来語の違い	案内, ガイダンス
旧名称と現名称の違い	江戸,東京
別称	ラフカディオハーン,小泉八雲
都市名等の異称	ヴェネチア,ベニス
片方が他方の比喩	犬,スパイ

表1: 同義対に含まれるデータの範囲

3. フォーマット

タブをフィールド区切りとし、4 フィールド 1 レコードのデータとしています。フィールドの説明を表 2 に示します。なお、データにはフィールド名は含まれていません。

フィールド番号	フィールド名	説明
1	word1ID	日本語 WordNet の word id。複数ある場合は"@"で区切られる。
2	word1	word1ID を持つ日本語 WordNet の語(lemma)。
3	word2ID	日本語 WordNet の word id。複数ある場合は"@"で区切られる。
4	word2	word2ID を持つ日本語 WordNet の語(lemma)。

表2: フィールドの説明

以下に例を示します。

<データ例>

217548	いさかい	190577	ロげんか
157025	うす茶色	228576	ライトブラウン
248196	おしまい	199808	終了
232728	ほうき星	183633	コメット
178252	やや	225897	赤ちゃん
207752	やり方	187525	スタイル
182109	イラスト	222835	挿絵
179506	クラス	202454	科目
179506	クラス	219478	等級
174913	シネマ	248877	映画
199908	上天気	207086	晴天

4. ライセンス

本データベースのライセンスは、以下の日本語 WordNet のライセンスに準拠します。

5. 問い合わせ先

独立行政法人情報通信研究機構 ユニバーサルコミュニケーション研究所
情報分析研究室

日本語 WordNet に関する問い合わせ先は以下のメールアドレスにお願いします。

Email: jwordnet@gmail.com