「分類語彙表番号－UniDic語彙素番号対応表」の構築　―コーパスへの網羅的・体系的な語義情報付与のために―

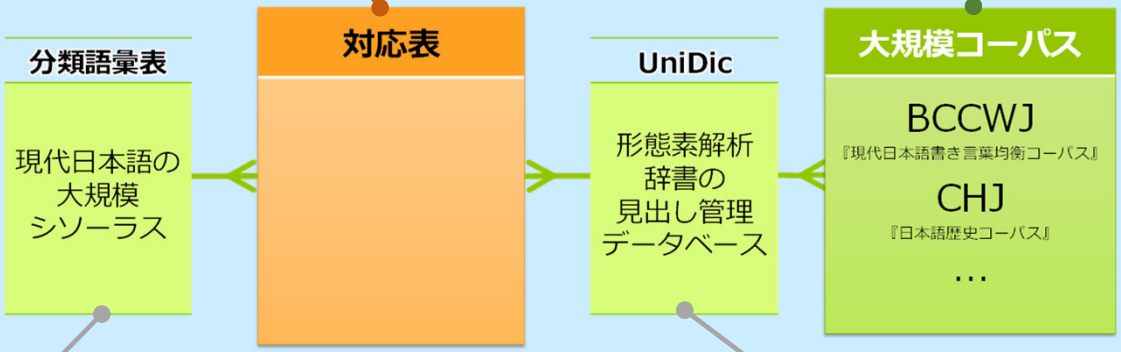
近藤明日子（国立国語研究所 コーパス開発センター）

1.　「分類語彙表番号－UniDic語彙素番号対応表」とは

分類語彙表とUniDicという2種類の語彙表に登録された見出しの間で、同語関係にあるものの多対多の対応を表す表（中間テーブル）です。

2.　対応表構築の目的

対応表の利用により、BCCWJ・CHJ等のUniDicによって形態論情報が付与された日本語の大規模コーパスに対して、分類語彙表による体系的な語義情報を網羅的に付与することを目指します。



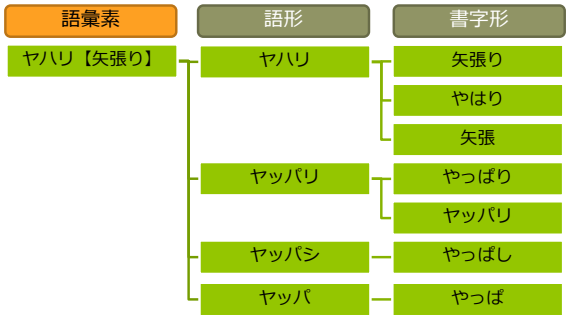
3.　対応表の規模と対応表の表す同語関係の例

対応表は全65,043レコードからなり、分類語彙表の64,759見出しとUniDicの50,795語彙素との間の同語関係を表しています。

分類語彙表見出し				UniDic語彙素			
分類語彙表番号	見出し本体	読み	類	語彙素ID	語彙素	語彙素読み	類
1.1000-03-01-01	事（こと）	こと	体	12836	事	コト	体
3.1030-09-01-01	正しい	ただしい	相	22353	正しい	タダシイ	相
2.3064-02-01-03	読む［数を～］	よむ	用				
2.3066-08-02-02	読む［相手の心を～］	よむ	用				
2.3100-20-03-02	詠む	よむ	用	39484	読む	ヨム	用
2.3150-02-01-01	読む	よむ	用				
2.3200-08-01-01	詠む	よむ	用				
1.5606-02-01-04	骨（こつ）	こつ	体	12809	骨	コツ	体
				12810	骨	コツ	接尾体
				12811	骨	コツ	接頭

分類語彙表…国立国語研究所で編纂された、現代日本語の大規模なシソーラスです。
分類番号と呼ばれる番号によって、日本語の意味世界を体系的に分類し、各見出しを分類番号により分類・配列しています。

UniDic…形態素解析辞書の見出し語を管理するデータベース、およびそこから作成した形態素解析辞書の名称です。国立国語研究所の構築する大規模コーパスの形態論情報付与に利用されています。
UniDicでは、語が階層化した形で登録されており、最上位の階層「語彙素」が、同じ意味を持つ複数の異語形や異表記をまとめあげています。



4.　対応表の活用例

- BCCWJの347,094 語を対象に分類番号による語義情報を付与するデータの構築
- 形態素解析結果に分類番号を付与する機能を実装した形態素解析ツール「ChaMame」の開発

5.　対応表のダウンロード

以下のウェブページのリンクからダウンロードして御利用いただけます。

https://pj.ninjal.ac.jp/corpus_center/goihyo.html

参考文献　近藤明日子・田中牧郎（2020）「「分類語彙表番号－UniDic語彙素番号対応表」の構築」『国立国語研究所論集』18, pp.77-91, http://doi.org/10.15084/00002542