UniDic : コーパスアノテーションのための電子化辞書

UniDic ―コーパスアノテーションのための電子化辞書―

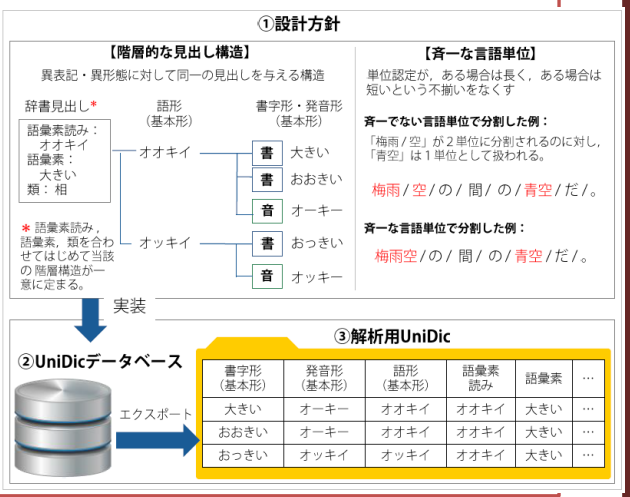
岡 照晃（コーパス開発センター・特任助教）

UniDicとは？

国語研の規定した斉一な言語単位（短単位）と、階層的見出し構造に基づく電子化辞書の
■　設計方針
およびその実装としてのリレーショナルデータベース
■　『UniDicDB』
と、そのデータベースからエクスポートした短単位を
エントリとする形態素解析器MeCab用の解析用辞書
■　『解析用UniDic』
の総称。

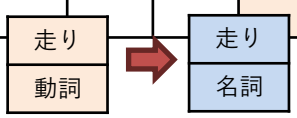
国語研におけるコーパスアノテーション
（生文を短単位に分かち書きし、形態論情報を付ける作業）
の支援を目的に整備されている。

実際のアノテーション作業では、
1．コーパス化したい生文を『解析用UniDic』で自動解析
2．『UniDicDB』を参照するツール『UniDicExplorer』と
コーパスデータベースを参照するツール『大納言』で
自動解析結果の誤りを修正している。



解析用UniDicを使いUniDicDBに格納された形態論情報 自動付与のイメージ

	走りに夢を、自転車の新しい走りへ										
分かち書き（書字形出現形）	走り	に	夢	を	、	自転	車	の	新た	な	走り	へ
品詞	名詞-普通名詞-一般	助詞-格助詞	名詞-普通名詞-一般	助詞-格助詞	補助記号-読点	名詞-普通名詞-サ変可能	接尾辞-名詞的-一般	助詞-格助詞	形状詞-一般	助動詞	動詞-一般	助詞-格助詞
活用型										助動詞-ダ	五段-ラ行	
活用形										連体形-一般	連用形-一般	
語彙素	ハシリ-走り-体	に-ニ-助動	ユメ-夢-体	ヲ-を-体	補助記号-読点-補助	ジテン-自転-体	シャ-車-接尾体	ノ-の-格助	アラタ-新た-相	ダ-だ-助動	ハシル-走る-用	へ-へ-格助
発音形出現形	ハシリ	ニ	ユメ	オ		ジテン	シャ	ノ	アラタ	ナ	ハシリ	エ
仮名形出現形	ハシリ	に	ユメ	ヲ		ジテン	シャ	ノ	アラタ	ナ	ハシリ	へ
語種	和	和	和	和	記号	漢	漢	和	和	和	和	和
書字形基本形	走り	に	夢	を	、	自転	車	の	新た	だ	走る	へ
発音形基本形	ハシリ	ニ	ユメ	オ		ジテン	シャ	ノ	アラタ	ダ	ハシル	エ
仮名形基本形	ハシリ	ニ	ユメ	ヲ		ジテン	シャ	ノ	アラタ	ダ	ハシル	へ
…												



自動解析結果の精度は100%ではないため、人手で修正する。
この際、分かち書きの誤りや、明らかな形態論情報付与誤り以外にも、
上の例のような形態論情報の揺れの斉一化も施していく。
また最新の解析用UniDicでは形態論情報自動付与の揺れが小さくなるように改良が施されているため、
実際には上記の揺れは起きない。

■岡 照晃: 言語研究のための電子化辞書, 講座日本語コーパス 7 「コーパスと辞書」第1章, 朝倉書店(2019) ■岡 照晃:CRF素性テンプレートの見直しによるモデルサイズを軽量化した解析用UniDic-unidic-cwj-2.2.0とunidic-csj-2.2.0-, LRW2017予稿集(2017)