Webデータに基づく複合動詞データベースの評価

山口昌也

（国立国語研究所）

概要

構築の背景

■ 複合動詞と構成動詞の関係分析

投げる

ヲ格	ニ格
ボール	相手
球	どこ
石	遠く
疑問	実際
身	時
物	中
ルアー	海
質問	上
言葉	人
速球	ところ

すべてが
継承される
わけではない

すべてが
構成動詞に
由来するのか？

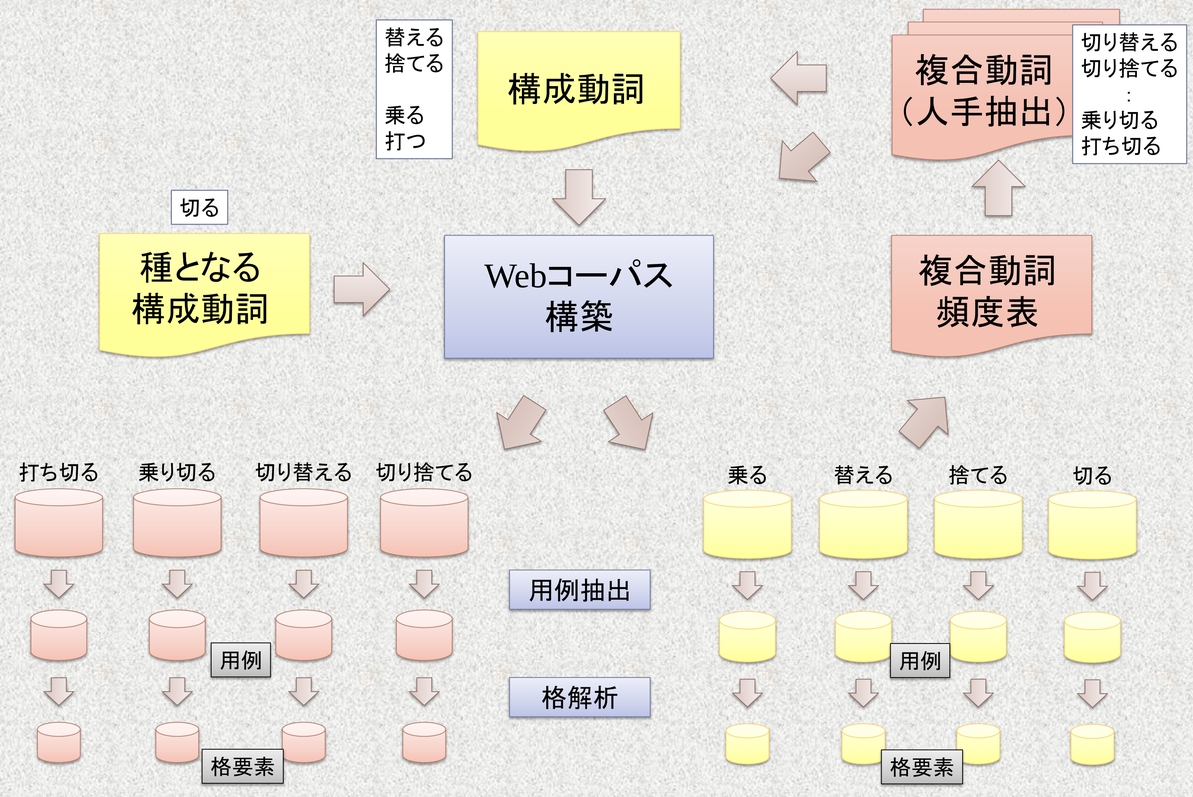
投げ込む

ヲ格	ニ格
ボール	中
球	川
速球	海
直球	池
石	そこ
ストレート	口
手榴弾	山
スライダー	水面
瓶	ポスト
ルアー	客席

構築の目的

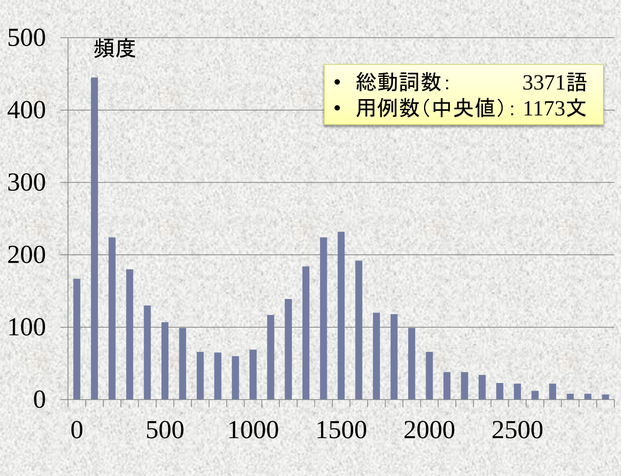
■ 実例に基づいた，複合動詞・構成動詞間の関係分析をすること
- ⇒ 一定数以上の用例を収集可能な複合動詞を収録する
- ⇒ Webデータの特性に合わせて，用例を収集・格解析する

構築方法

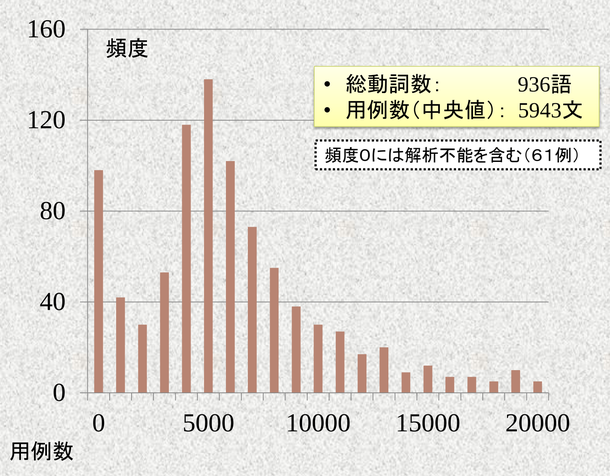


構築結果 （http://csd.ninjal.ac.jp/comp）

収集結果（複合動詞）



収集結果（単一動詞）

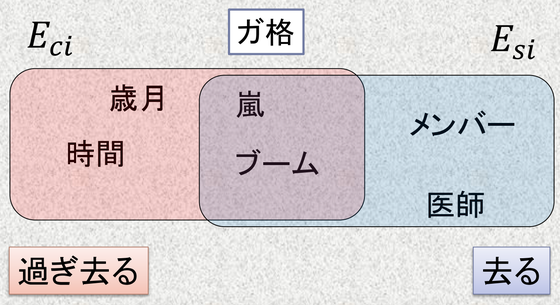


分析例：「～込む」と前項動詞との関係分析

重複度

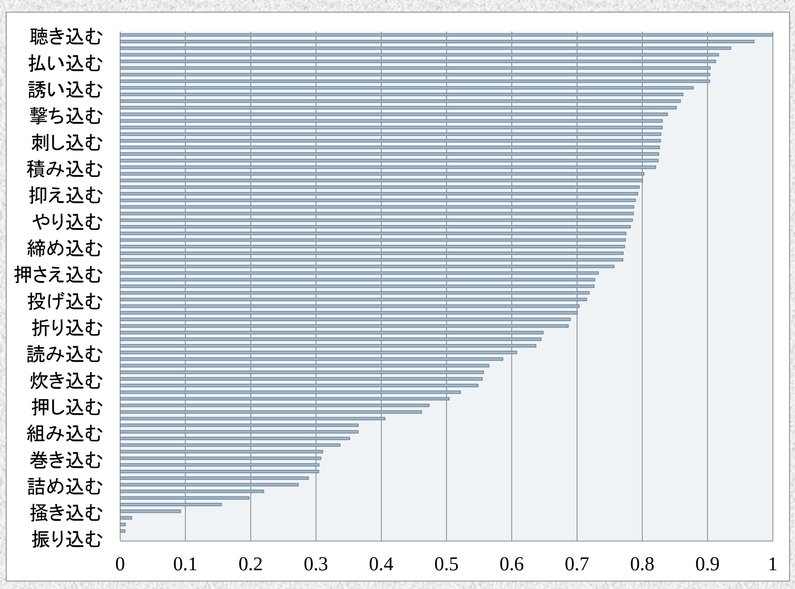
■ 複合動詞の格 i の取りうる格要素が，構成要素の格 i の格要素となりうる割合

$$OV_i = \sum_{w_a \in E_{ci} \cap E_{si}} n(w_a) \;/\; \sum_{w_b \in E_{ci}} n(w_b)$$



重複度の測定結果（ヲ格，$F_{複,ヲ格} \geqq 50$，$F_{複} \geqq 1000$，99語）と関係分類

■ 継承 「聴き込む」(1.0)，「着込む」(0.97)



■ 別義 「振り込む」(0.0)，「申し込む」

■ 派生 … 別義が混在
「織り込む」(0.29)
「糸を織り込む」
「情報を資料に織り込む」

■ 分布変化
構成動詞では，あまり用いない格要素

「流し込む」(0.46)
「鉄を鋳型に流し込む」
⇔ ?「鉄を流す」

格要素	複合動詞	構成動詞
モルタル	19	0
樹脂	17	0
ビール	17	0
金属	16	0
セメント	15	0
鉄	11	0

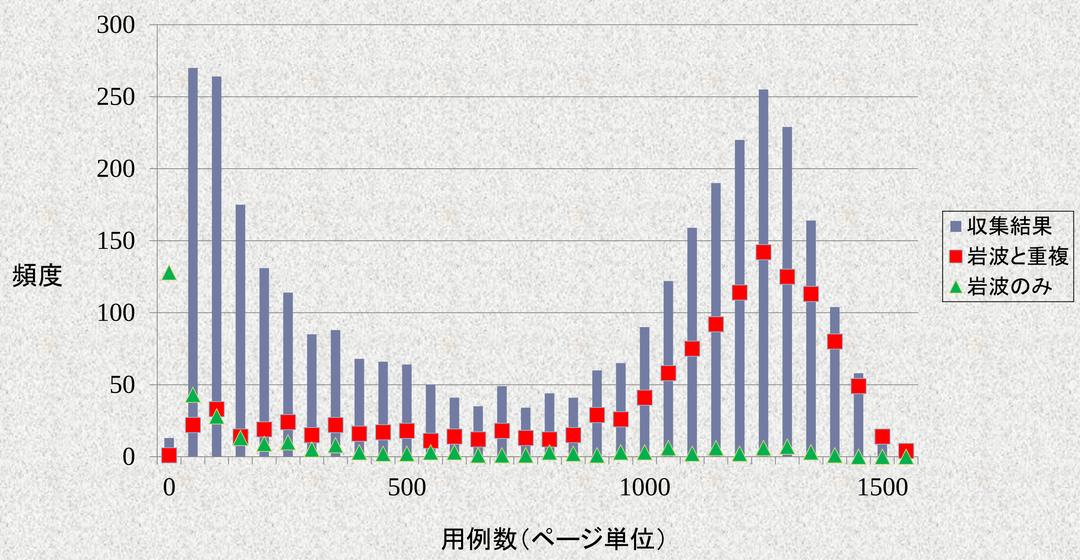
評価

評価の観点

■ 収録された複合動詞の性質
- ⇒ 岩波国語辞典との比較

■ 収録された格要素の分布の性質
- ⇒ ランダムサンプリングWebコーパスとの比較

岩波国語辞典（第5版）との比較



■ 「岩波」収録の複合動詞の77.2%を収録
- 高頻度側の山は，「岩波」収録語を中心に収録
- 低頻度側の山は，定着していない語や専門用語か？

■ 「岩波」のみに収録されている複合動詞（371語，うち解析不能66語）
- 低頻度語が多い（中央値74p）
- 解析不能語には，古語，音便形が含まれる場合が多い
（例：噛みしだく，突っ立てる）

ランダムサンプリングWebコーパスとの比較

■ ランダムサンプリングWebコーパス
- 単一動詞（用例数1000以上）のWebコーパスから重複を取り除いた448万ページがソース
- 各ページからランダムに3文抽出し，コーパスとした（約2.1億語）

■ 共起語ベクトルのCOS類似度の比較
- 対象
ランダムサンプリングWebコーパス，複合動詞用例データベースともに用例数1000以上の複合動詞（115），単一動詞（395）
- COS類似度（0以上1以下）の計算結果
単一動詞： 0．86
複合動詞： 0．83