Webデータに基づく複合動詞用例データベースの構築と活用

山口昌也（国立国語研究所）

明らかにしたいこと（データベースを作る目的）

- 複合動詞が構成されるとき，周辺の語の分布にどのような変化があるか

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
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

背景

- 共同研究プロジェクト
  「文脈情報に基づく複合的言語要素の合成的意味記述」

- 目的

  単語周辺の分布情報から，複合的な言語要素の意味記述を合成的に記述する方法を探る（分布意味論的なアプローチ（Pado, Lapata2007））

嘆く ＋ 悲しむ ⇔ 嘆き悲しむ

どのような関係か？

LCS1 LCS2 ⇒ LCS1 AND LCS2

由本(2005)などの理論的研究と関連付けたい

本日の内容

- 複合動詞用例データベースの紹介
  - 収録内容
  - Web 上の検索システムデモ
- データベースの活用例

データベースに対する要求と現状

- 要求
  - 複合動詞，構成動詞の用例を大量に保持すること
  - 格要素がすぐ取り出せること

- 既存の資料の状況
  - 野村，石井　「複合動詞資料集」(1987)
  - 各種形態素解析用辞書
    - 複合動詞は網羅的に登録されていない

⇨ 複合動詞のリストを作りつつ，用例も収集する

複合動詞用例データベース
http://csd.ninjal.ac.jp/comp

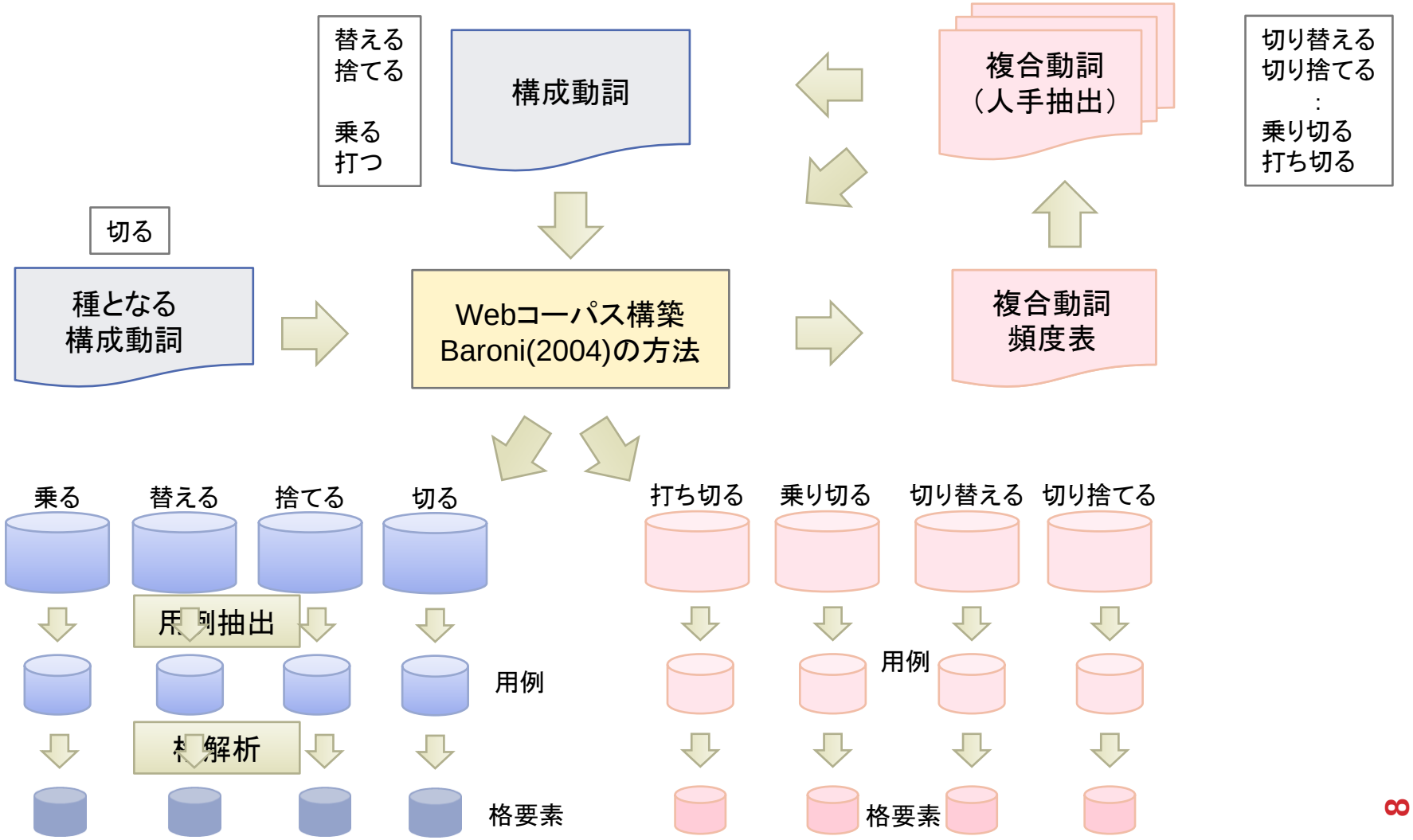
- 特徴
  - Web データに基づいて構築
    - Web 上の使用頻度に基づき，収録対象の複合動詞を決定
    - Web 上の用例を収集
  - 構成動詞の情報も収録

- 収録している情報
  - 語構成
  - 用例（格解析結果付き）

- 規模（2012-09-24現在）
  - 複合動詞　3362語（主として，語彙的複合動詞（影山1992））
  - 構成動詞　1040語

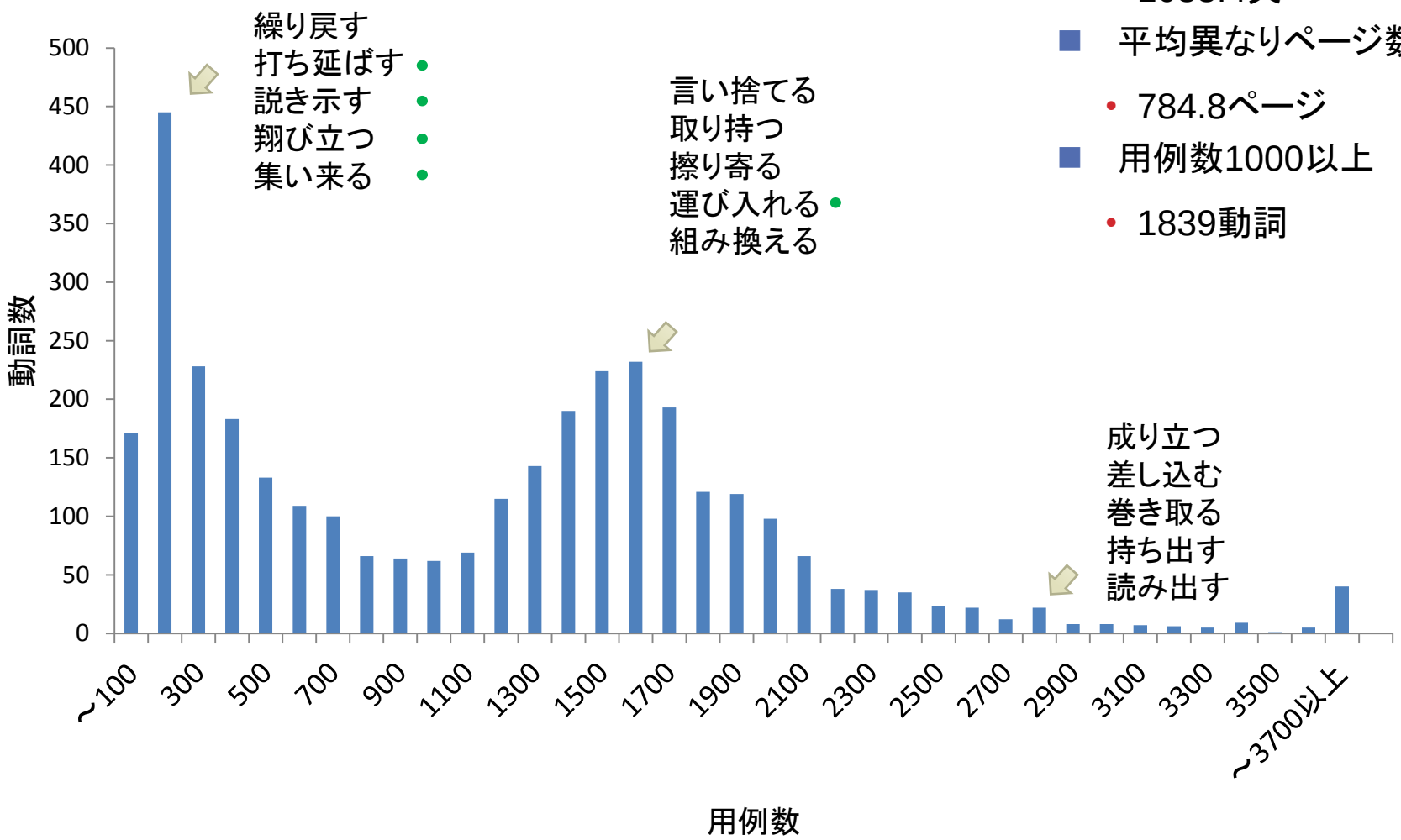
複合動詞用例データベースのデモ

データベース構築の流れ

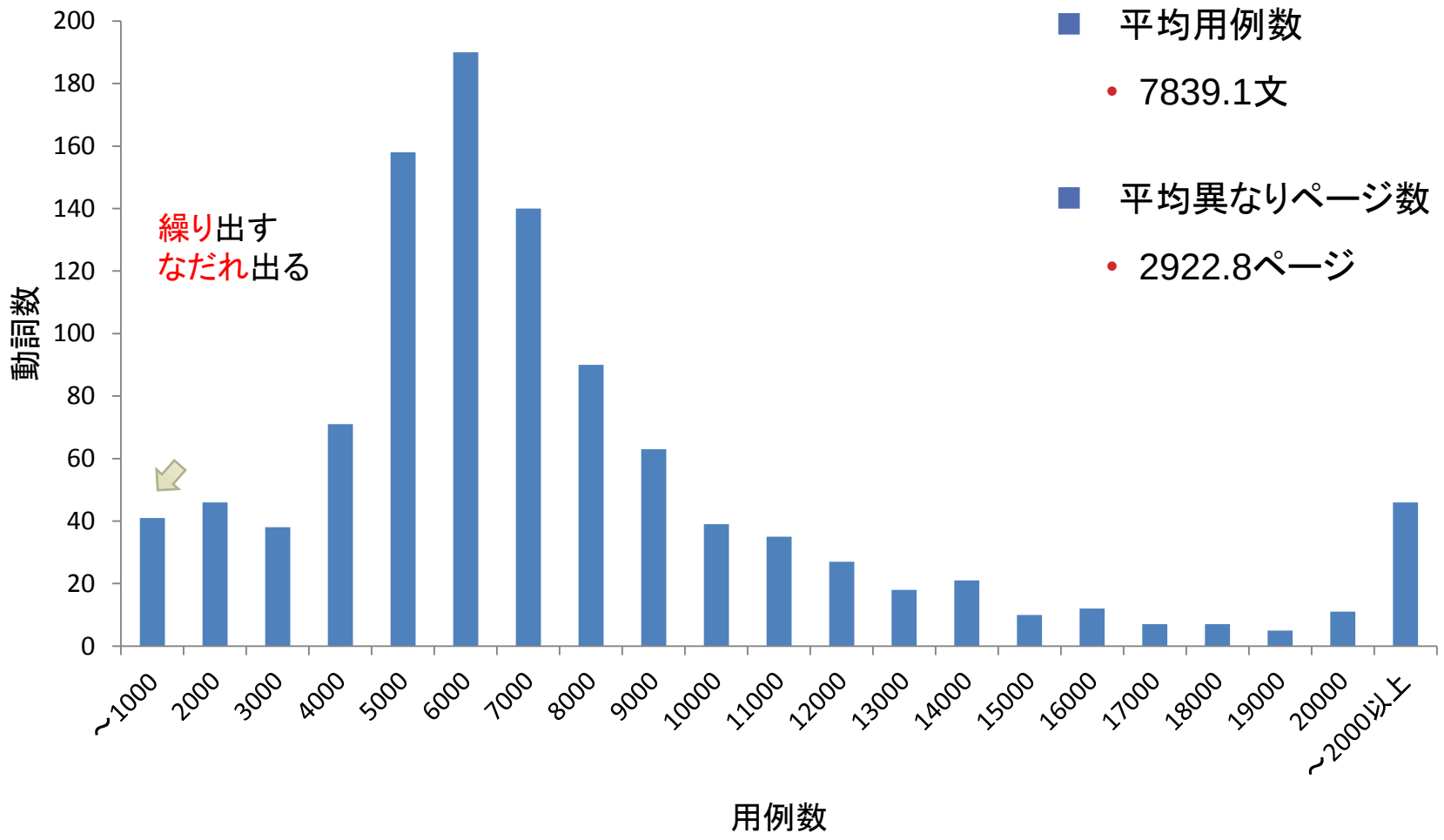


収録用例数（複合動詞）

- 平均用例数
  - 1088.4文
- 平均異なりページ数
  - 784.8ページ
- 用例数1000以上
  - 1839動詞



収録用例数（構成動詞）


200
180
160
140
120
100
80
60
40
20
0
動詞数
繰り出す
なだれ出る
～1000
2000
3000
4000
5000
6000
7000
8000
9000
10000
11000
12000
13000
14000
15000
16000
17000
18000
19000
20000
～2000以上
用例数
平均用例数
7839.1文
平均異なりページ数
2922.8ページ

活用例

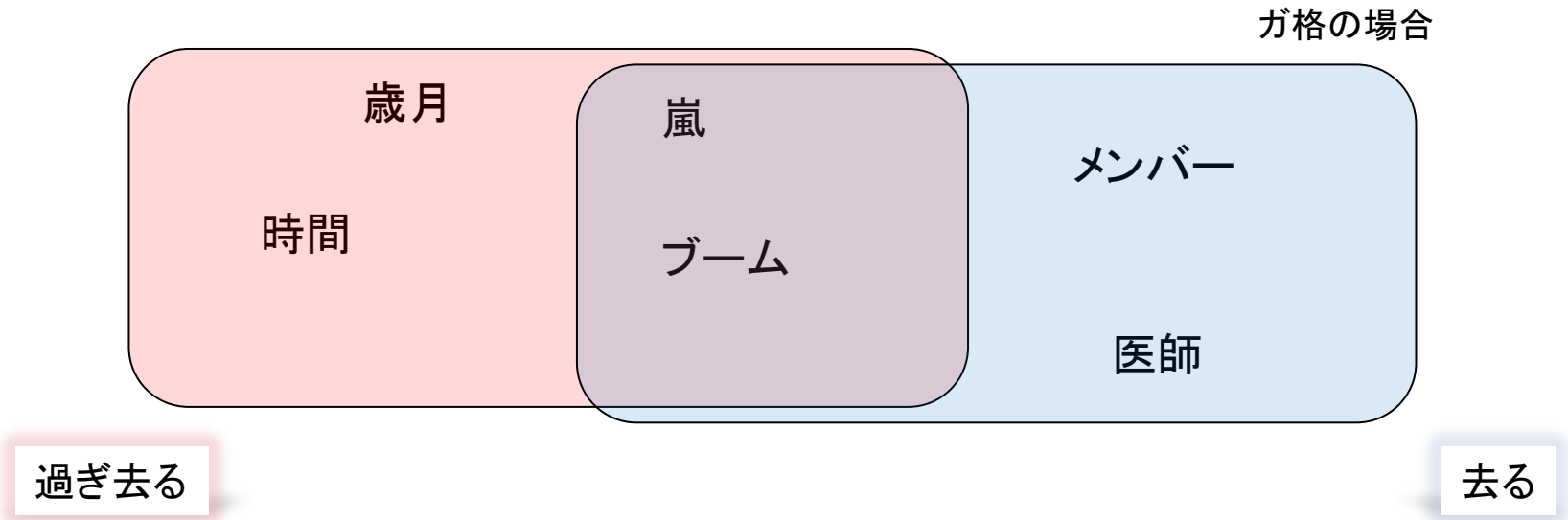
- 複合動詞と構成動詞における，周辺の語の分布の違いを分析
  - 格要素の「重複率」を測定
    （共通して用いられる格要素の割合）

- 活用例1：　複合動詞と前項・後項動詞との関係を俯瞰する
  - ヲ格を持つ複合動詞を対象に，重複率の分布を見てみる

- 活用例2：　重複率低下の原因を探る
  - 「～込む」タイプの複合動詞を対象とする

格要素の重複率

- 複合動詞の格要素のうち，構成動詞でも使用される格要素の割合

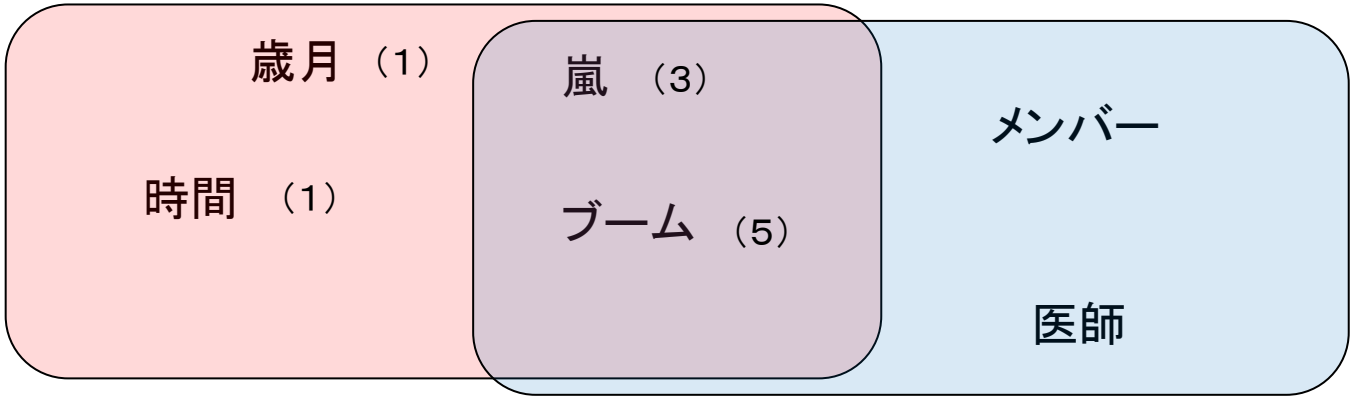


$$OV_i = \sum_{w_a \in E_{ci} \cap E_{si}} n(w_a) \ / \sum_{w_b \in E_{ci}} n(w_b)$$

「過ぎ去る」と「去る」の重複率

格要素の重複率

カッコ内は，「過ぎ去る」側の用例数



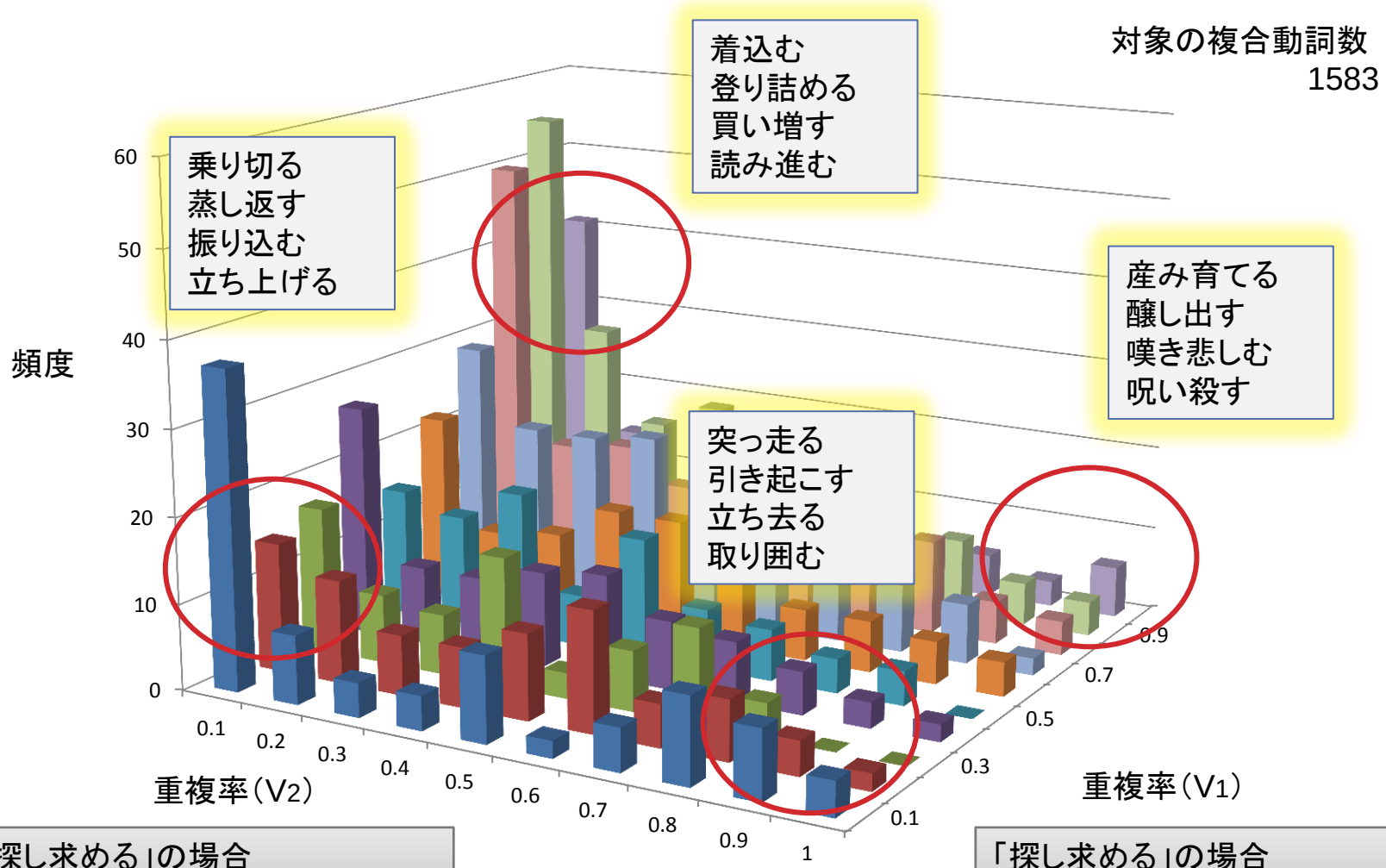
過ぎ去る

去る

$$OV_{ガ} = (3+5) / (1 + 1 + 3 + 5)$$
$$= 0.80$$

例1：ヲ格の重複率の分布

$F_{複,ヲ格} \geqq 50$, $F_{複} \geqq 1000$, $F_{単} \geqq 2000$



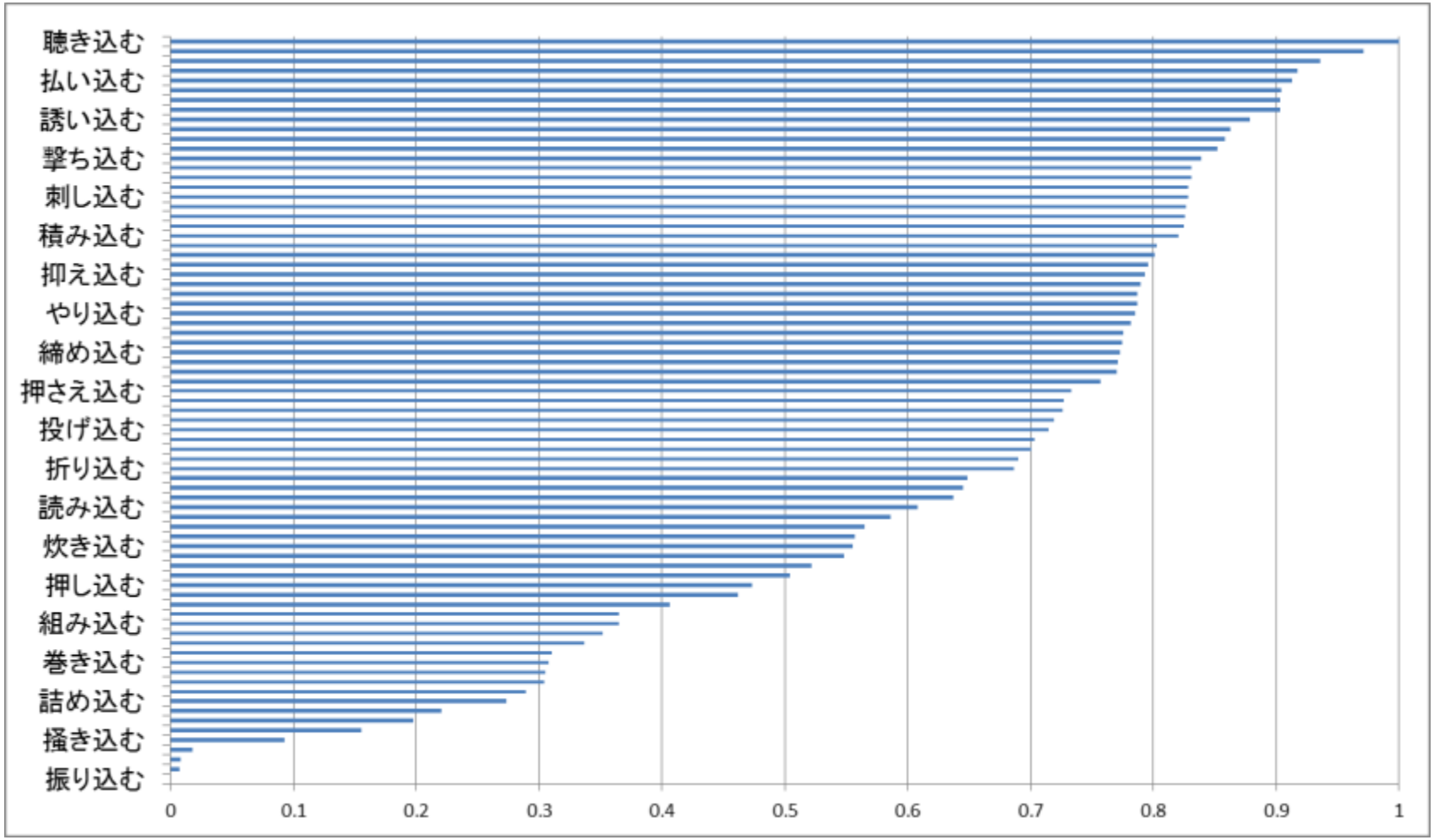
「探し求める」の場合
「求める」と「探し求める」の重複率

「探し求める」の場合
「探す」と「探し求める」の重複率

例2:「～込む」と前項動詞の重複率の分布

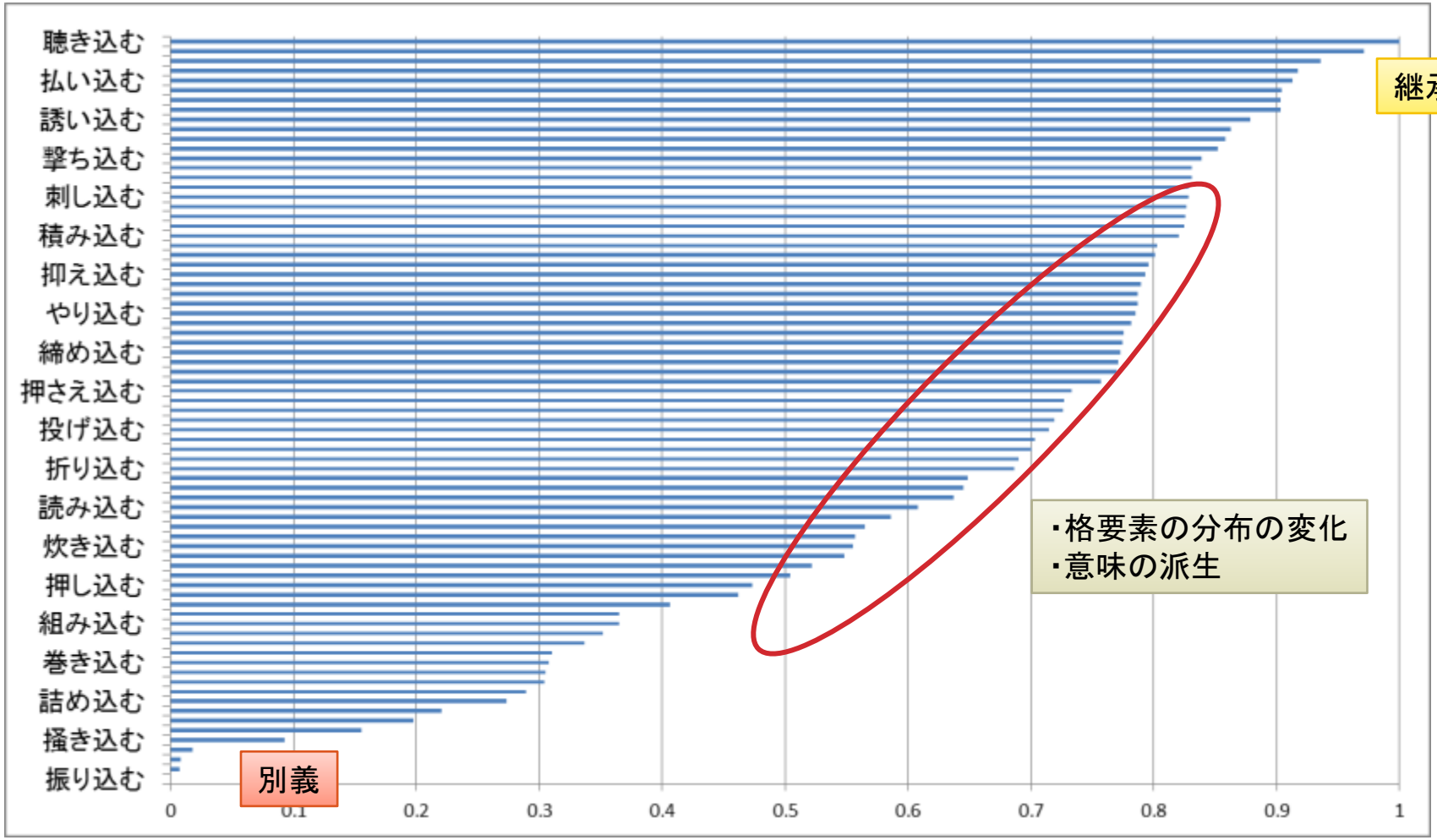
$F_{複} \geqq 1000$

（ヲ格）　対象語132語



重複率から見た動詞間関係

（ヲ格）　対象語132語



派生

- 構成動詞では，用いない格要素の使用
  - 新しい語義となる
  - 比喻的な用法が見受けられた

- 実例
  - 「織り込む」（ $OV_{ヲ} = 0.29$ ）
    「糸を織り込む」　⇔ 「糸を織る」
    「情報を資料に織り込む」 ⇔ *「情報を織る」

1. 織物で，地とは異なる糸などを織物の中へまぜて織り，模様などを作る。「金糸を－・む」
2. 一つの物事の中に，他の物事をふくめる。盛り込む。「その費用は予算に－・んである」

三省堂：大辞林（http://www.sanseido.net）

出現確率の変化

- 構成動詞では，あまり用いない格要素の使用

  複合動詞　＞＞　構成動詞

- 実例
  - 「流し込む」（$OV_{ヲ}$ = 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

  - 「読み込む」（$OV_{ヲ}$ = 0.61）

    「画像をコンピュータに読み込む」

    ⇔「画像を読む」

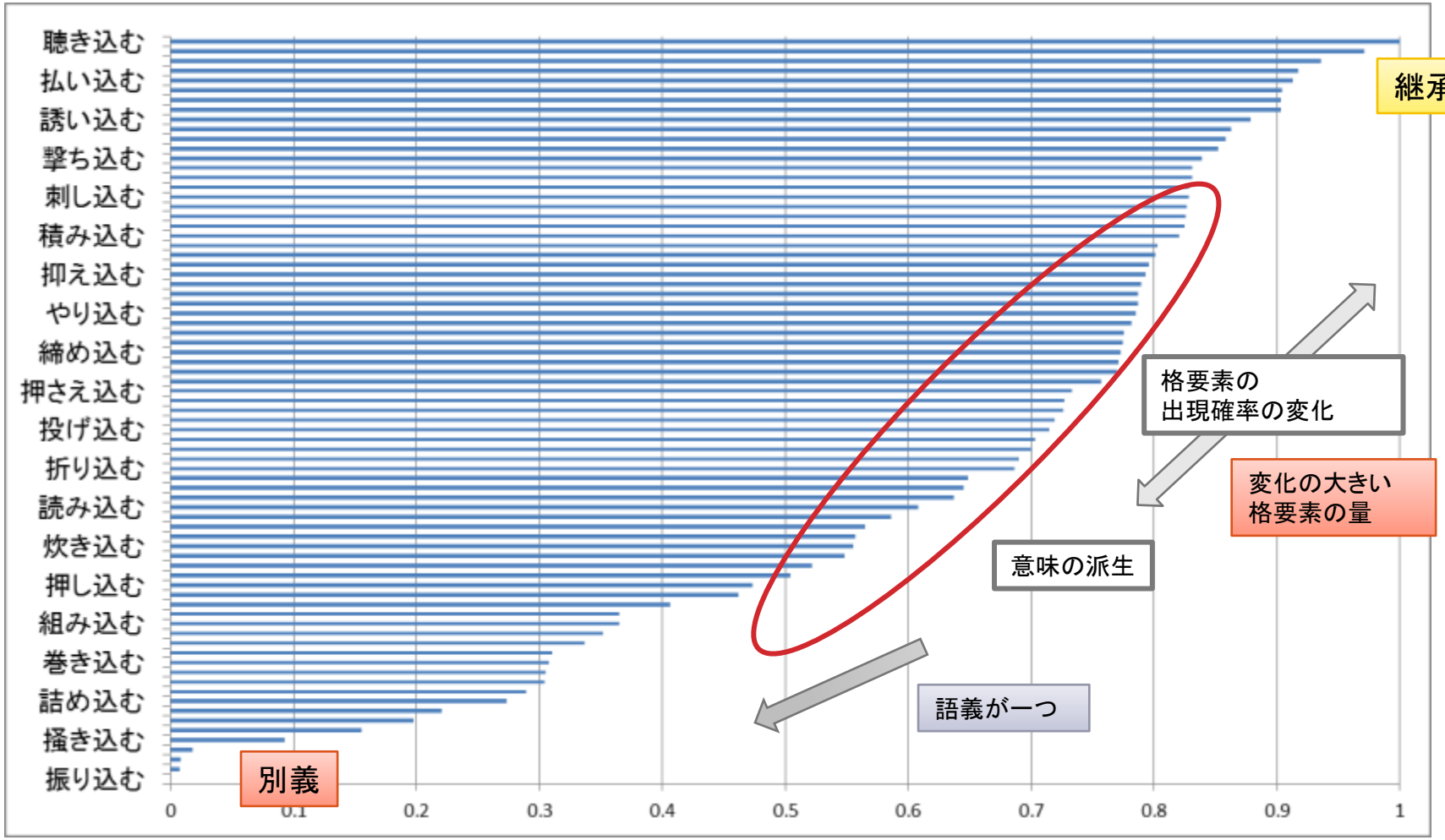
    「ファイルをコンピュータに読み込む」

    ⇔「ファイルを読む」

格要素	複合動詞	構成動詞
画像	41	0
写真	16	0
設定	15	0
ディスク	14	0

重複率から見た動詞間関係

（ヲ格）　対象語132語



まとめ

- データベース作成の目的

  複合動詞と構成動詞の周辺分布の変化を明らかにすること

- 複合動詞データベースの構築
  - 構築方法，結果を示した
  - 複合動詞・構成動詞間の関係を客観的に分析するための道具

- データベースの活用例
  - ヲ格の重複率の分布
  - 重複率低下の原因分析（「～込む」タイプの複合動詞を対象に）

参考文献

- ［影山1993］ 文法と語形成，ひつじ書房（1993）
- ［由本2005］ 複合動詞・派生動詞の意味と統語，ひつじ書房（2005）
- ［野村・石井1987］ 複合動詞資料集，科研費特定研究（1）言語データの収集と処理の研究 (1987)
- [Pado, Lapata2007] Sebastian Padó, Mirella Lapata. Dependency-Based Construction of Semantic Space Models, Computational Linguistics Vol.33, No.2, pp.161-199 (2007)
- [Baroni2004] M. Baroni and S. Bernardini. BootCaT: Bootstrapping corpora and terms from the web. Proceedings of LREC 2004.
- 大辞林，三省堂，http://dic.yahoo.co.jp