2019/03/03

コーパスの解析単位（短単位・長単位）

国立国語研究所
音声言語研究領域　　柏野　和佳子

はじめに

- 『中納言』の短単位検索
  - 「英語」「電車」「魚屋」を検索　→ 検索結果が表示
  - 「日本語」「自転車」「パン屋」を検索 → 検索結果0件

【理由】「英語」「電車」「魚屋」は1 語になっているが，「日本語」「自転車」「パン屋」は，2 語に分割されている。

|日本|語|　　|自転|車|　　|パン|屋|

《語》

- コーパスでは，全てのテキストを《語》に分割し，それぞれの語に対して，見出し・品詞・活用型・活用形・語種等の情報を付与している。
- 検索：「国語」を検索した際に，検索結果に「外国語」「韓国語」等の意図しない結果も含まれるという問題を回避できる。
- 分析：あるテキストの延べ語数・異なり語数，品詞比率，語種比率，ある語の使用率を調べることができる。
- 《語》の定義は様々であり，一定しない。
  - 1 語 |国立国語研究所|
  - 2 語 |国立|国語研究所|
  - 3 語 |国立|国語|研究所|
  - 4 語 |国立|国語|研究|所|

言語単位の設計

- 「国立国語研究所」を単語に区切る場合，４通りの区切り方が考えられるが，どれか一つが正しいというものではない。
- 言語単位の認定に関する基準を立て，その基準に基づいて，コーパス全体を不統一のないように語（言語単位）に分割することが重要。
- 「不統一がない」とは，

① 同じ語が常に同じに区切られていること。

【例】　|日本|語|　|にほん|ご|

|話し=合う|　|話し=あう|　|話=合う|

② 同じ構造の語が常に同じに区切られていること。

【例】　|日本|語|　|韓国|語|　|中国|語|

|自転|車|　|自動|車|　|人力|車|　|動力|車|

|国立|国語|研究|所|　|国立|科学|博物|館|

言語単位の設計

- 「不統一がある」場合の問題

① 語数を正確に把握することができない。

【例】　|オレンジ|色| 2 語　|オレンジ色| 1 語

→ 語種の計量に影響

|オレンジ|色| 外来語1,　和語1

|オレンジ色| 混種語1

|国立|国語|研究|所| 4 語

|国立|科学|博物館| 3 語（国立|科学|博物|館| 4 語）

→ 品詞の計量に影響

|国立|国語|研究|所| 名詞3,　接尾辞1

|国立|科学|博物館| 名詞3（国立|科学|博物|館| 名詞4 ）

② 全用例を収集するために複数の検索を試みる必要がある。

【例】　|日本|語|　|日本語|

→「日本」＋「語」と「日本語」の検索が必要

言語単位の設計

- 「中納言」では２種類の言語単位を使用
  - **短単位　(short unit word)**
    - 用例収集を目的
    - 形態的側面に着目
    - 「最小単位」を結合して認定
  - **長単位　(long unit word)**
    - 言語的特徴の解明を目的
    - 構文的機能に着目　　※文脈に依存
    - 文節内を自立語部分と付属語部分に分割して認定

意味を持つと考えられる最小の単位。

最小単位

- 現代において意味を持つ最小の単位。

  - 和 語 ： /豊か/な/暮らし/に/つい/て/
    /大雨/が/降っ/た/の/で/
  - 漢 語 ： /国/立/国/語/研/究/所/
  - 外来語 ： /レーザー/プリンター/
    /オレンジ/色/
  - 記 号 ： /図/Ａ/ /Ｊ Ｒ/
  - 人 名 ： /和田/豊/　/マット/・/マートン/　/林/威助/
  - 地 名 ： /大阪/府/豊中/市/待兼山町/　/六甲/山/

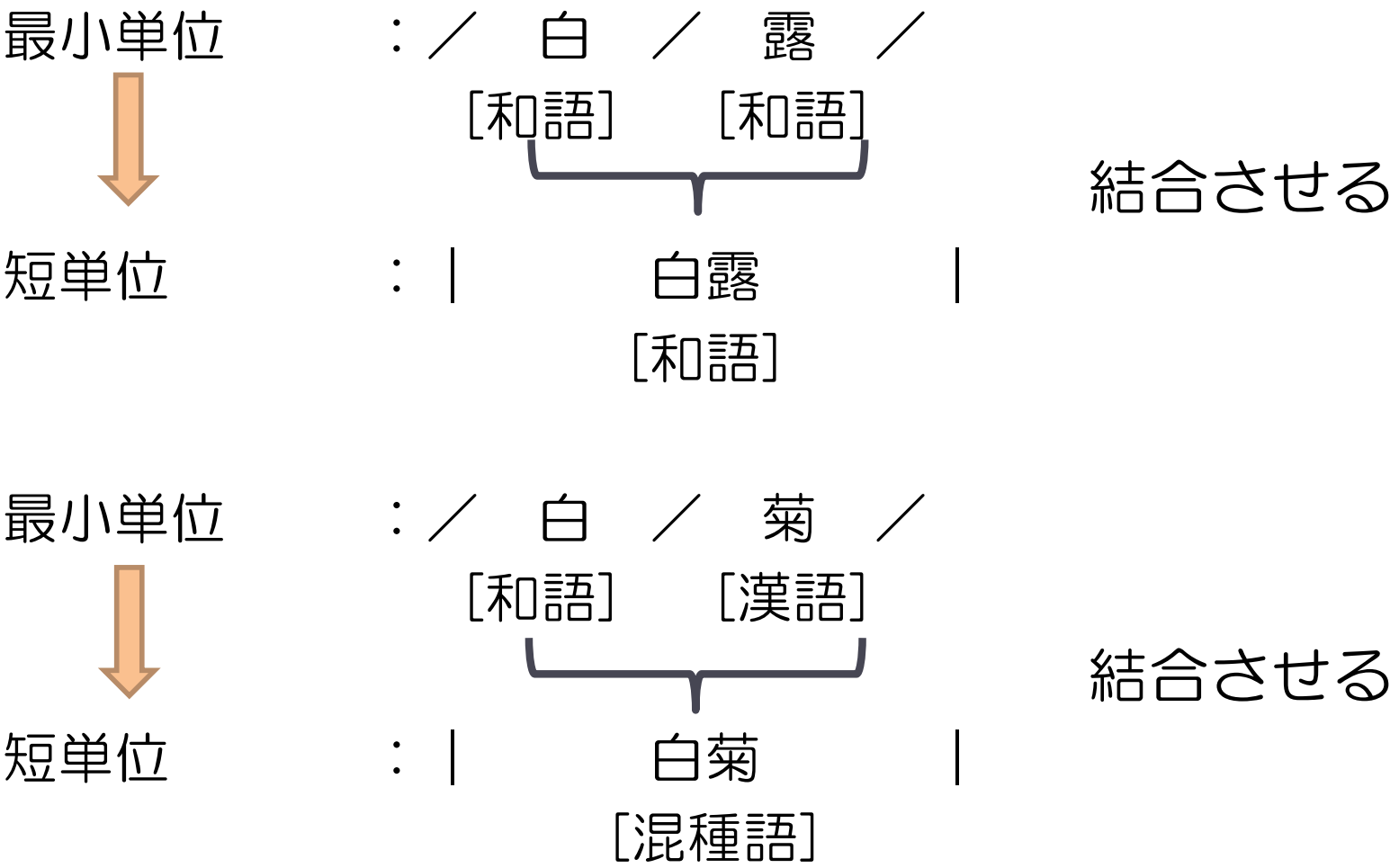
短単位

- 言語の形態的側面に着目して規定した言語単位
- 最小単位を結合させてつくる
- 最小単位の分類ごとに結合方法が決まっている
  - 和語・漢語　　：最小単位を２つまで結合
    （結合させなくてもいい）
  - 外来語 ：１最小単位＝１短単位　など

短単位認定のための最小単位の分類

分類		例	結合方法
一般	和語	山　川　白い　話す　言葉	２つまで結合可
	漢語	社　会　用　研　究　所	２つまで結合可
	外来語	オレンジ　ボックス	結合不可
付属要素	接頭的要素	相　御（お）（ご）　各	結合不可
	接尾的要素	合う　致す　っぽい　性	結合不可
その他	記号	、　・　。　「　」Ａ　Ｂ	結合不可
	数	一　二　十　幾　数	「数」同士結合可
	固有名・人名	星野　仙一　ジェフ	結合不可
	地名	大阪　待兼山町　六甲	結合不可
	助詞・助動詞	た　です　ます　か　から	結合不可

短単位の例



短単位の例

最小単位　　：　／　右　／　大　／　将　／

[漢語]　[漢語]　[漢語]

３最小単位の結合となるので不可

短単位　　：　｜　右　｜　大　　将　｜

[漢語]　　[漢語]

長単位

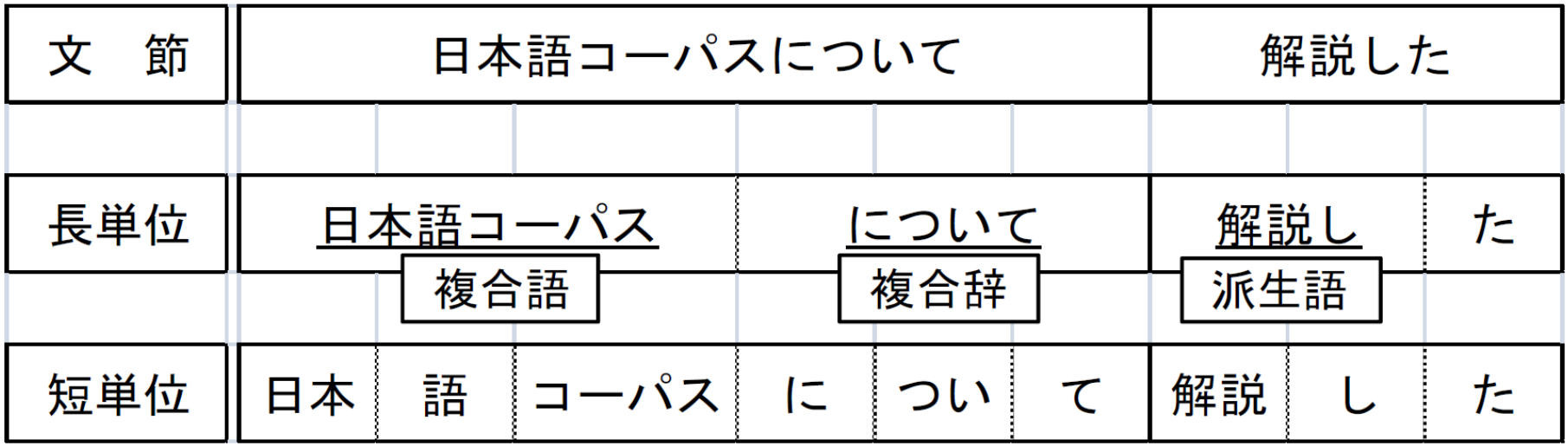
- 言語の構文的な機能に着目して規定した言語単位
- テキストをまず文節に区切る
- さらに、文節内を自立語部分と付属語部分に分割して認定

- 短単位との差
  - 接頭辞・接尾辞
  - 複合語、動詞連続など
  - 連語（トワズガタリ、オモイノホカetc.）
  - 複合辞（デゴザル、テクレル）

リスト化

長単位：短単位との違い

- 合成語を構成要素に分割しない。
- 複合辞や連語を1 長単位として認める。



長単位：注意点

- 長単位認定規程に従って分割された全ての語が長単位である。
  - 複合語・派生語・複合辞のみが長単位ではなく，1短単位から成る長単位もある（長単位と短単位とが一致する場合がある）。

文　節	今日は		いい	天気だ	
長単位	今日	は	いい	天気	だ
短単位	今日	は	いい	天気	だ

冒頭の例

例）ここは国立国語研究所です。

短単位：

｜国立｜国語｜研究｜所｜　　4語

長単位：

｜国立国語研究所｜　　1語

長単位の特徴

品詞性が明確

- 「うれしさ」

短単位：[形容詞語幹] ＋ [接尾辞-名詞的]

長単位：名詞

- 「哀れ」

短単位：[名詞-普通名詞-形状詞可能]

長単位：形状詞　or　名詞

用法に基づき
どちらかに決定

その子も哀れなような気がする。⇒　形状詞

滑稽な哀れさ！　⇒　名詞

見出し語を使った検索

- 短単位で認定された語を検索する。

例１）
書く
書いた
書いている
書かない

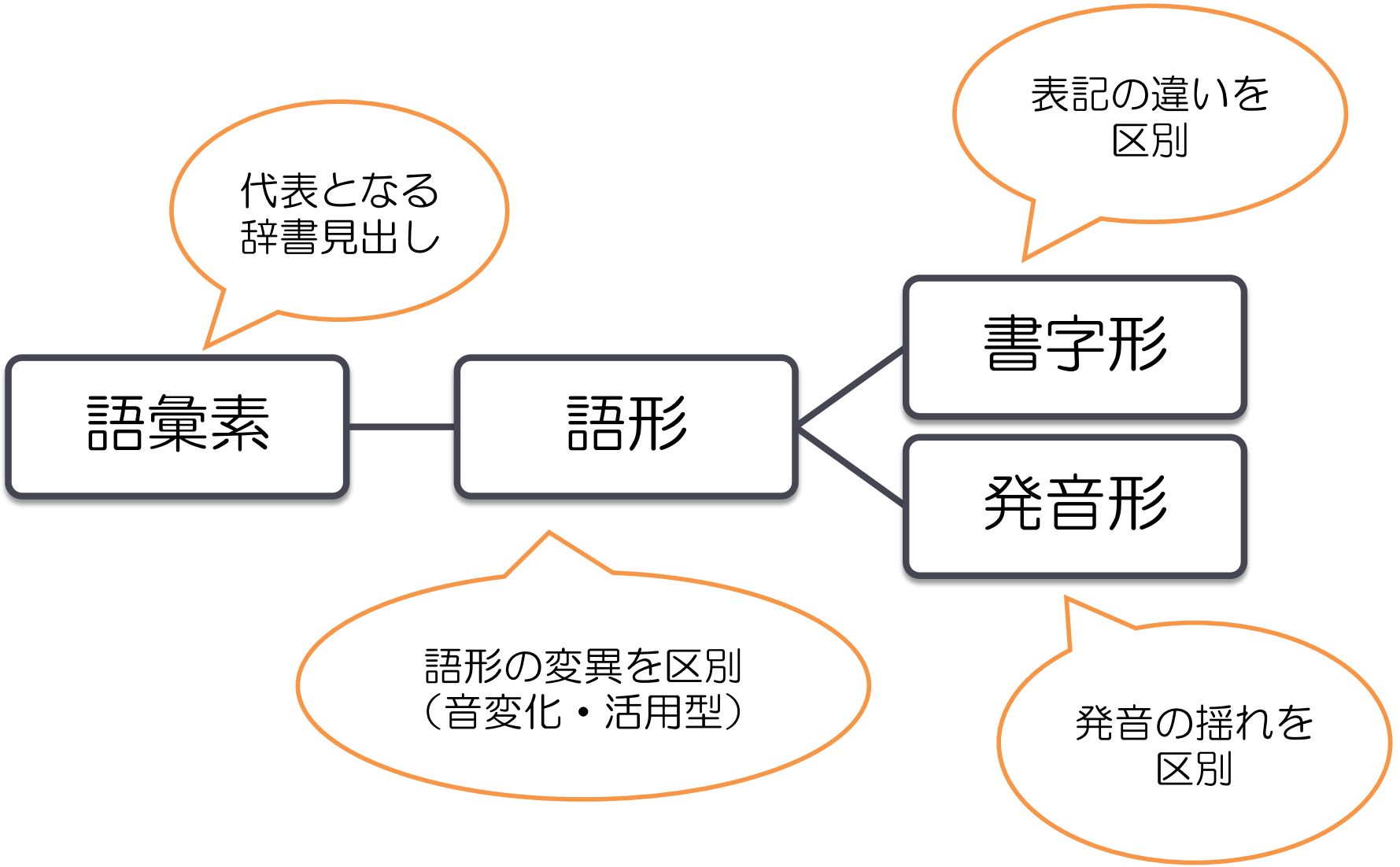
活用形によらず、まとめて検索したい

例２）
ほととぎす
郭公
時鳥

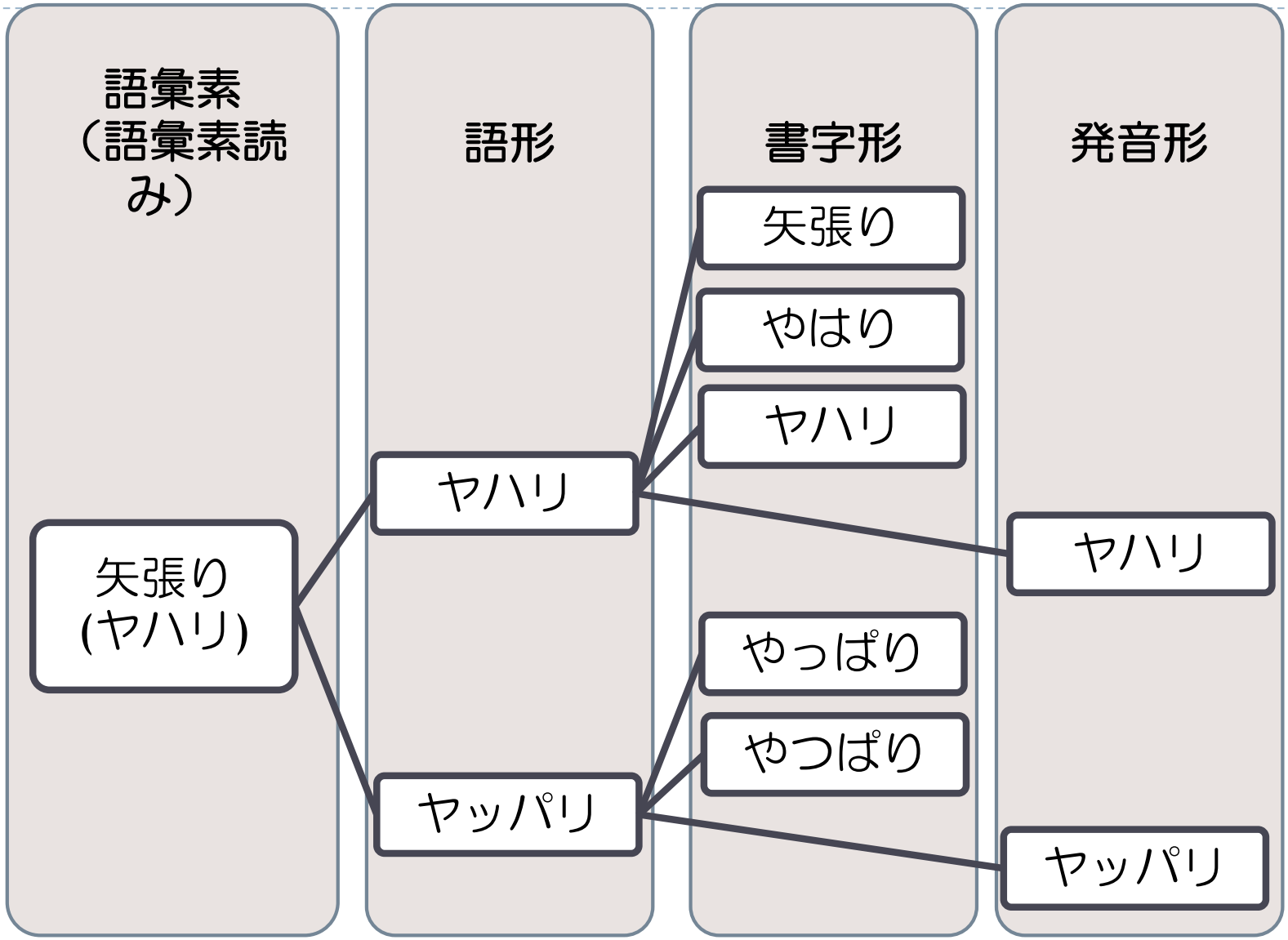
表記によらず、まとめて検索したい

- 形態論情報の見出し語を利用した検索
  ⇒一括して検索可能

見出し語の階層構造


代表となる
辞書見出し
表記の違いを
区別
語彙素
語形
書字形
発音形
語形の変異を区別
（音変化・活用型）
発音の揺れを
区別

「矢張り(ヤハリ)」の階層構造



基本形と出現形

- 基本形：活用語の終止形での形、語頭変化・語末変化していない形
- 出現形：コーパスの中での実際の形

例１）文を書いて
　　書字形基本形：書く
　　書字形出現形：書い
例２）三匹の犬
　　発音形基本形：ヒキ
　　発音形出現形：ビキ

書字形と発音形は「出現形」での検索が可能

見出し語の情報の付与

文字列を短単位or長単位の「語」に区切り、
それぞれの「語」に対して見出し語の情報
（語彙素／語形／書字形／発音形）を付与。

⇒表記や語形の揺れ（バリエーション）の研究が可能

検索の注意１

（ １ ） 単位境界を理解する。

検索したい語を，まず［ 文字列検索］ で検索し，短単位・長単位でどのように分割されているのかを理解する。

（ ２ ） 同じ文字列でも前後の文字によって短単位境界が変わる場合がある。

【例】 「市内」 |大阪|市|内|観光|　　|ＮＴＴ|市内|通話|料金|

「課長」 |経理|課|長|　　|本省|課長|級|職員|

検索の注意２

（ ３ ）語彙素・語彙素読みの一方を指定しただけでは，検索対象が一つに定まらない場合がある。

【例】

《語彙素読み》 アウ 会う／ 合う　 ヤサシイ 易しい／ 優しい

ミミザワリ 耳障り／ 耳触り

《語彙素》 難い ガタイ（ 接尾） ／ ニクイ（ 形容）

居る イル（ 動） ／ オル（ 動）

がら空き ガラアキ（ 体） ／ ガラスキ（ 体）

※ 語彙素の漢字表記は，　国語辞典等を参照して決定したため，あまり目にしない表記となっている場合がある。

【例】 迚も（ トテモ） 愈（ イヨイヨ） 齎す（ モタラス）