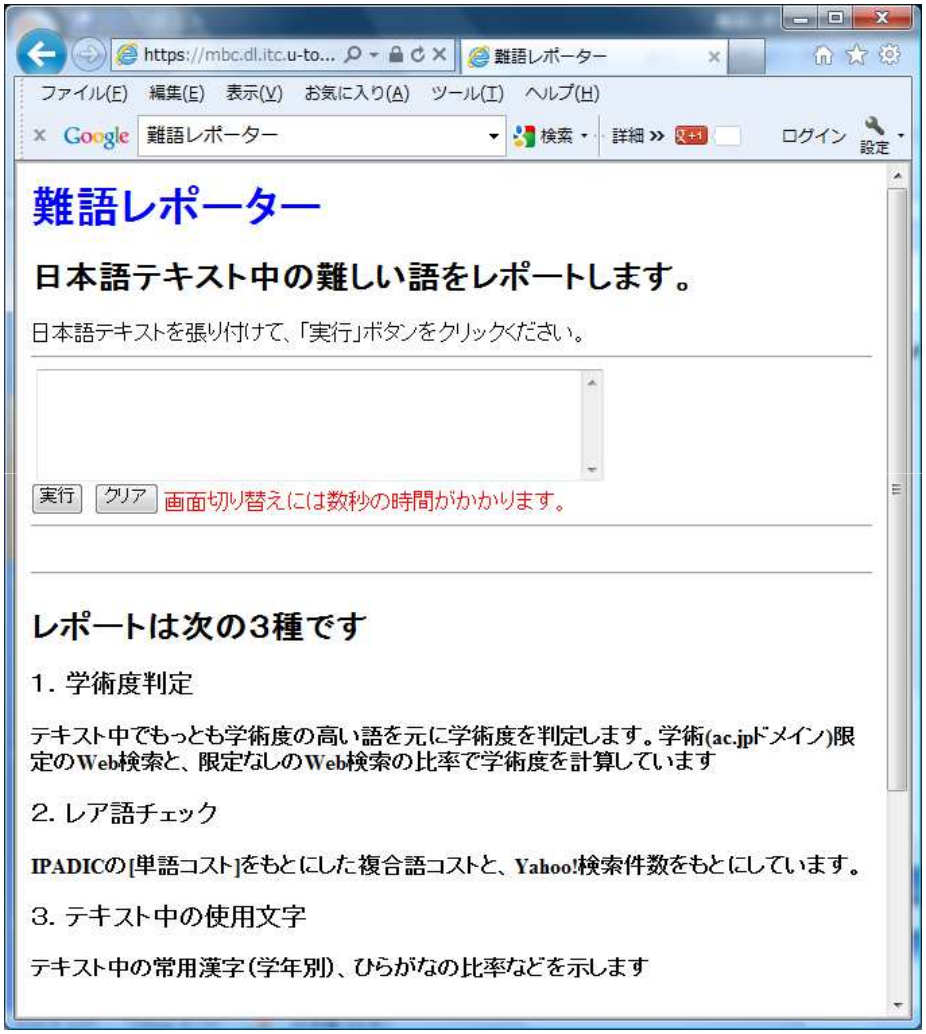
「難語レポーター」

平成24年3月26日

東京大学柏図書館　前田朗

画面イメージとサイトURL



https://mbc.dl.itc.u-tokyo.ac.jp/nango_report/

はじめに

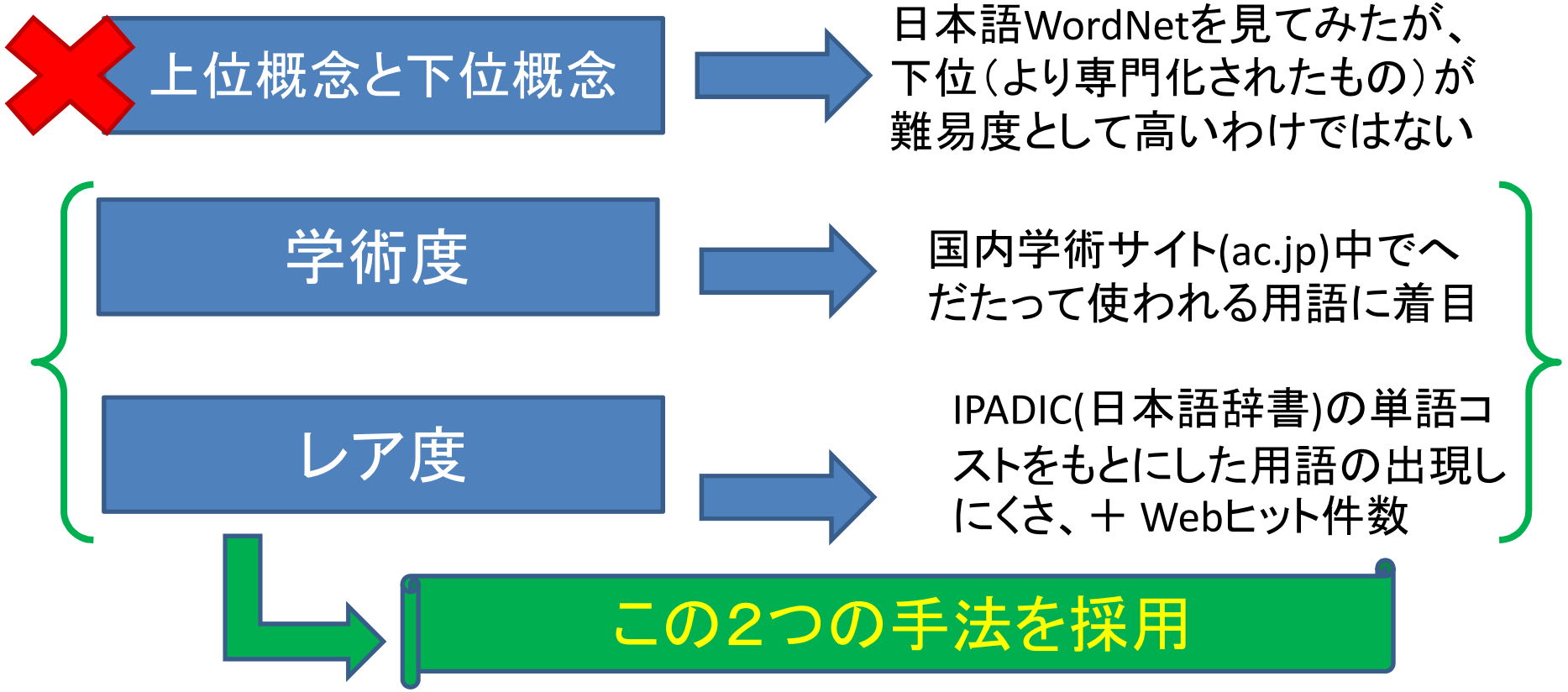
- 文の難易度を判定できると
  - 本の推薦・選定 or 蔵書の評価
  - その他、一般向けにも使えるかも
- 「難読度」の既存研究
  - 帯2 （教科書コーパスによる機械学習）
  - ひらがなや文章長さ述語数の組合せ

一定量のテキストが必要（書名のような短いフレーズでは厳しい）
図書の内容紹介の難読度と本体の難読度が必ずしも一致しない

従来の「難読度」とは
別の視点で考える

個々の用語の難易度に着目する

→　短いフレーズでも有効に働くはず



常用漢字集計

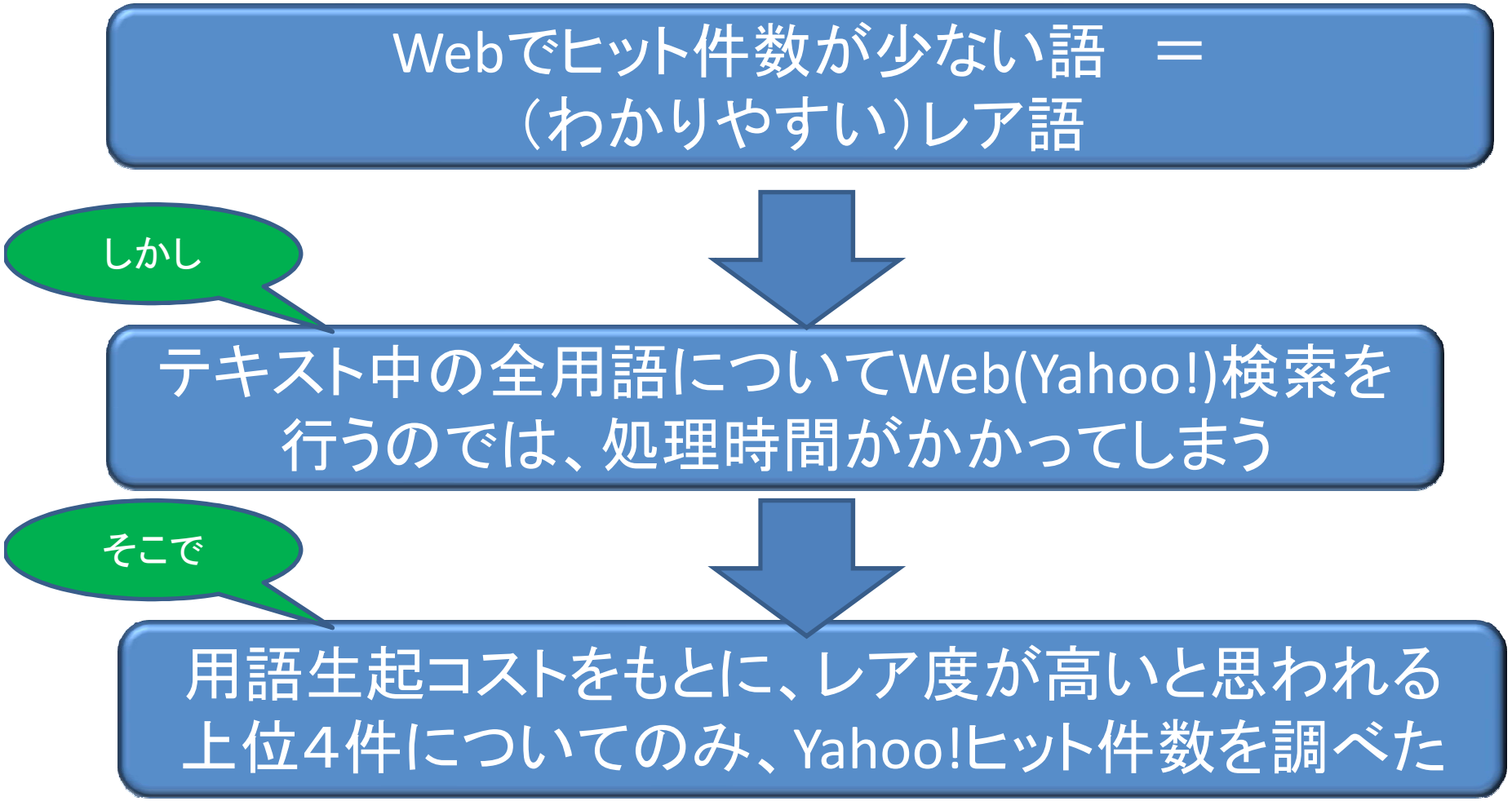
- 十分なテキストの量が多いと「難読度」の推定に使える
  - 中学生以上で学ぶ漢字や、小学n年生で学ぶ漢字、ひらがななど数量を集計
- 参考情報として「難語レポーター」に組み込み
- 情報漢字表は、Wikipediaから取得

手法の詳細

レア語抽出
1-単語コストと複合語

- 言選WebのMeCab.pm のコードを流用
- IPADICには、個々の形態素に、形態素の出現しやすさを示す「形態素生起コスト」をもつ
  - 「形態素生起コスト」が高いほど出現しずらい
- 「用語生起コスト」
  - 形態素生起コストの独自拡張
  - 複合語をなす個々の形態素の「形態素生起コスト」を掛け合わせる
  - 普通に乗算を行うと桁数が大きくなるので、log（自然対数が底）にて計算

レア語抽出
2-Yahoo！　ヒット件数



レア語抽出
3-処理フロー

テキスト

形態素解析（和布蕪）

形態素

用語生成（言選Web）

全用語

各用語の「拡張単語コスト」を計算

上位4件の用語

Yahoo! ヒット件数調査

上位4件の用語

学術度判定

- 学術度　＝　国内学術サイトのヒット件数 ÷ 日本語Webサイトのヒット件数
- レア語と同様に用語生起コストの上位4件のみ判定
- 判定基準

学術度	判定
0.5より上	学術度が非常に高い
0.1～0.5	学術度が高い
0.05～0.1	学術度あり
0.01～0.05	学術度判定不可
0～0.01	学術度が低い

[課題]テキスト全体としての評価

- テキスト全体としての評価をどのように計算で出すかは、これからの課題
- 現状では、テキスト中のもっとも学術度の高い語で学術度を判定させているが、テキスト量（対象となる用語数）が多いと、学術度を高く判定しかねない