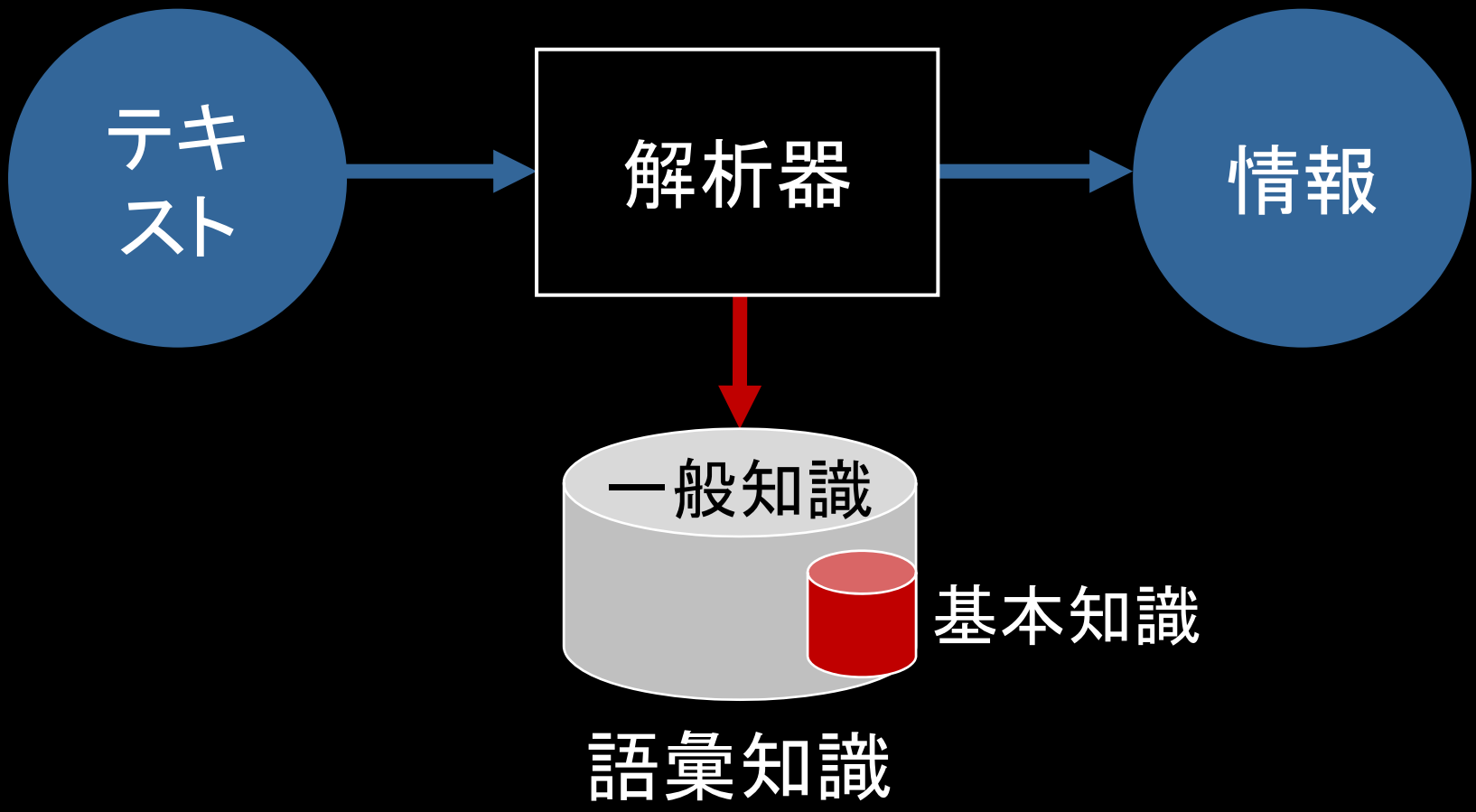
オンライン未知語獲得

京都大学 黒橋研究室
村脇 有吾

2009年12月27日
第2回入力メソッドワークショップ

言語の理解に語彙知識が必要

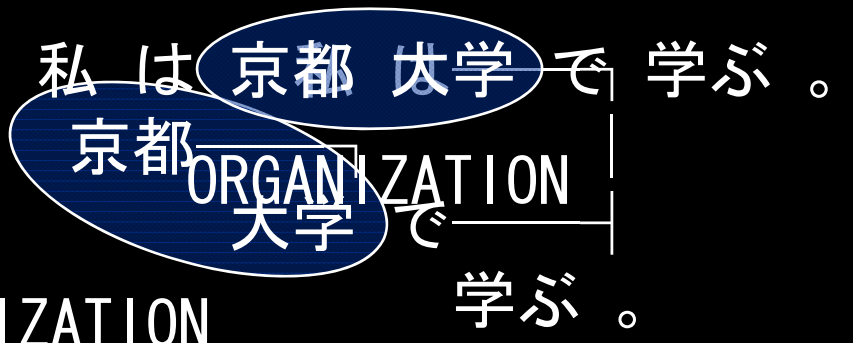


段階的な言語解析

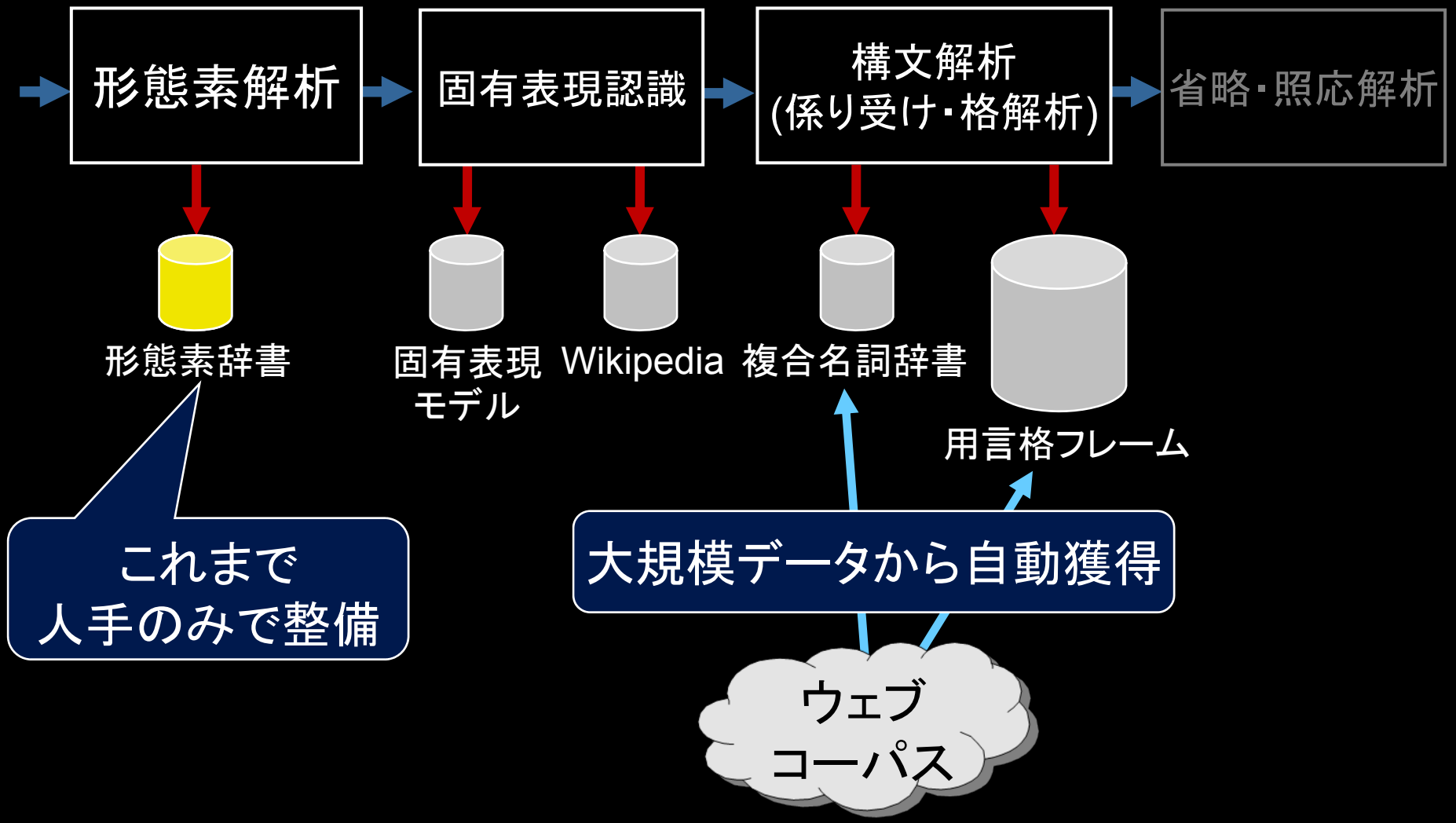
形態素解析 → 固有表現認識 → 構文解析 (係り受け・格解析) → 省略・照応解析

応用:
- 情報検索
- 機械翻訳, etc

私は京都大学で学ぶ。



解析に必要な語彙知識


形態素解析
固有表現認識
構文解析
(係り受け・格解析)
省略・照応解析
形態素辞書
固有表現
モデル
Wikipedia
複合名詞辞書
用言格フレーム
これまで
人手のみで整備
大規模データから自動獲得
ウェブ
コーパス

未知語による解析誤り

- ついったー → つ + いった + ー
- 倖田來未 → 倖 + 田 + 來 + 未
- ググ-る → ググ + る
- ムカつ-く → ムカ + つく
- 喰ら-う → 喰 + ら + う
- うざ-い → う (鵜/雨) + ざい (剤/材/…)
- けいはんな → けい (刑/型/形/…) + はん (判/半/反/…) + な (名/菜)

未知語獲得のタスク設定

- 一般的な語彙は人手で整備済み
  - 約12万形態素 (うち基本語彙3万)
- 足りない語彙 (未知語) を自動獲得
  - テキストから
- (人手による獲得語彙の修正は原則なし)
- (オンライン処理)

IME開発との共通点と相違点

- 共通点
  - 英語にない問題
  - 未知語が大問題
- 相違点
  - 深い言語解析がやりたい
  - 読みなんて飾りです
  - 形態素の単位はそれなりに重要
    - 「ただしイケメンにかぎる」は4形態素

人手による形態素の記述例

品詞

(名詞 (普通名詞

語幹(見出し語)

((読み まつり)

(見出し語 祭 (まつり 1.6))

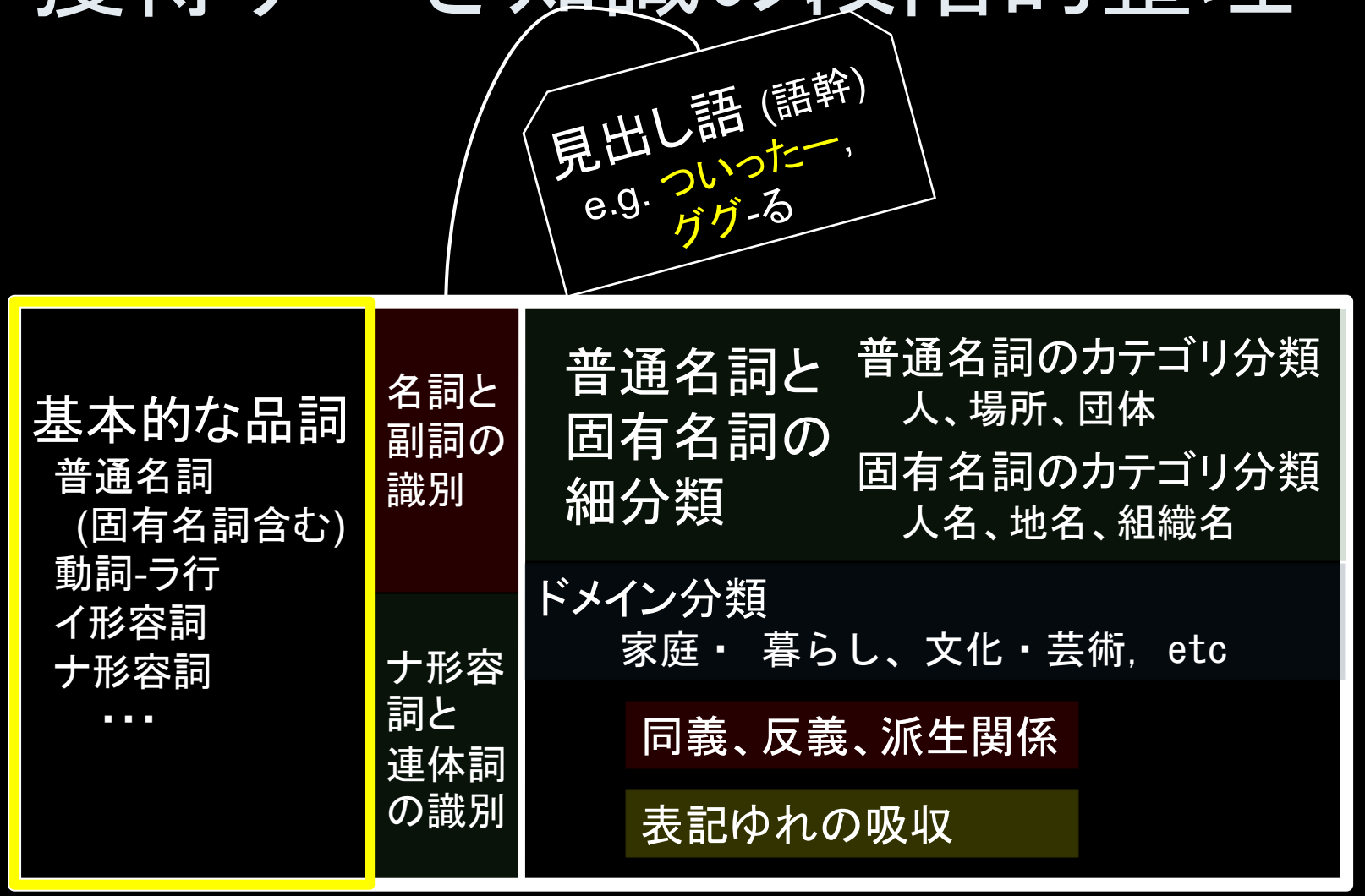
表記ゆれの吸収

(意味情報 "代表表記:祭/まつり
漢字読み:訓
カテゴリ:抽象物
ドメイン:家庭・ 暮らし;文化・芸術
動詞派生:祭る/まつる")

)))

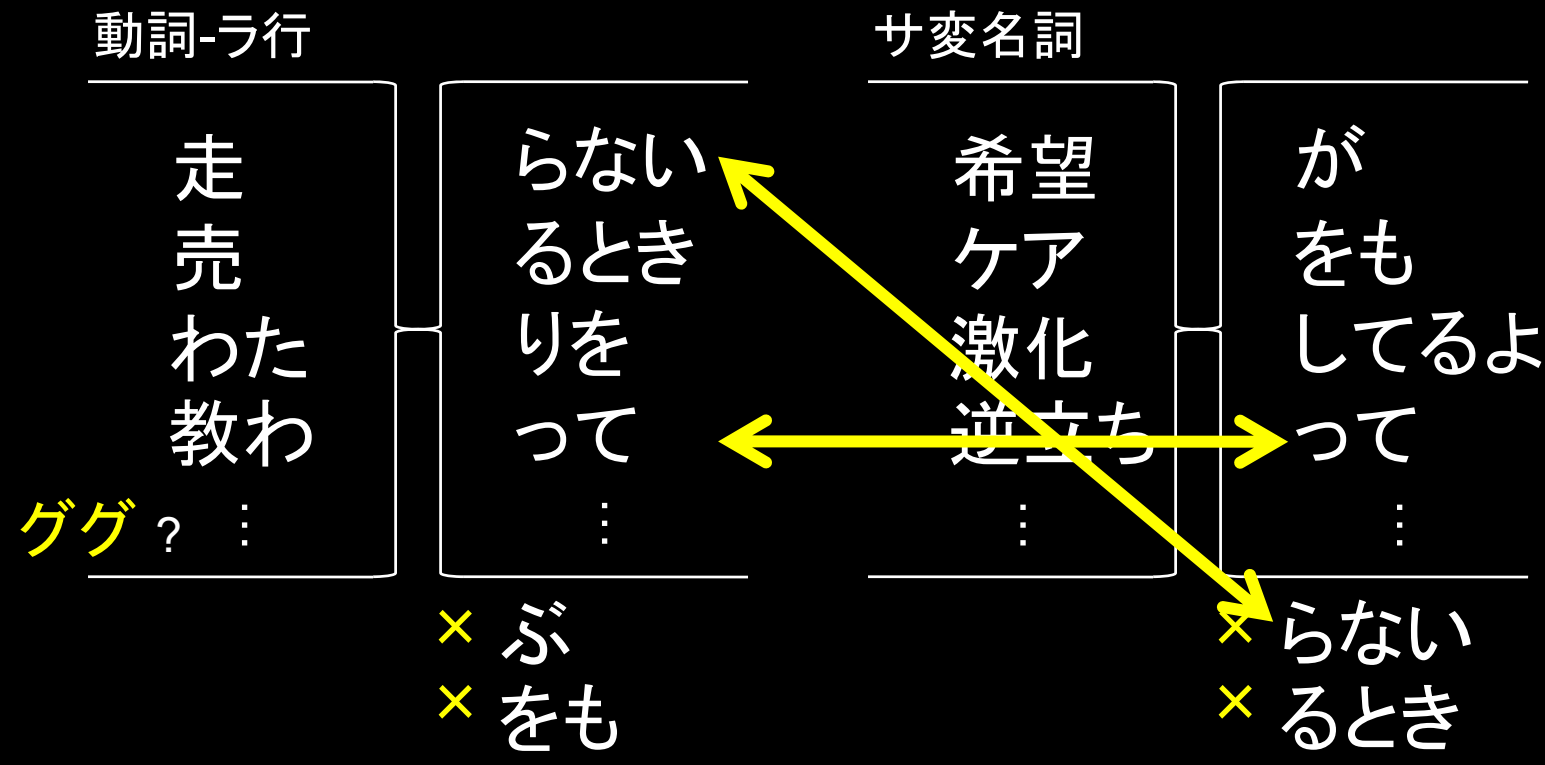
応用処理に必要な意味情報

獲得すべき知識の段階的整理



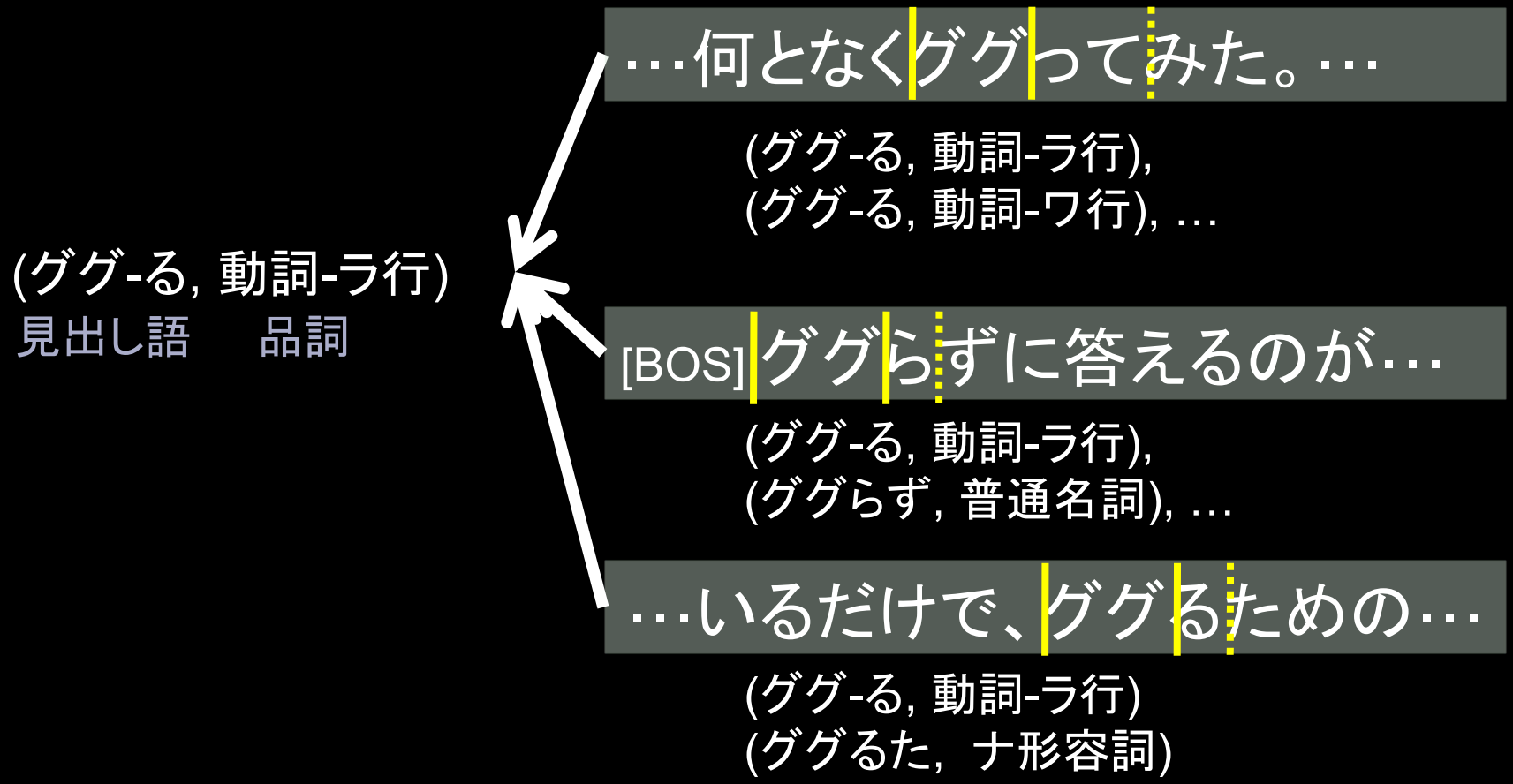
形態論的制約を用いた品詞識別

- ある形態素の語幹に後続し得る付属語列は、品詞ごとに形態論的制約に従う
- 制約充足のチェックにより未知語候補の品詞が絞り込める

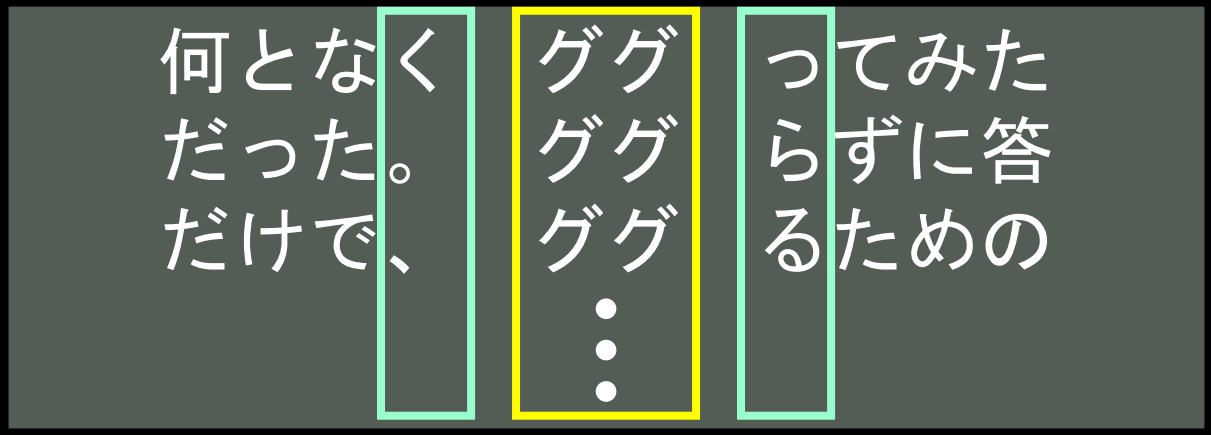


複数用例の比較による曖昧性解消

テキスト中の用例



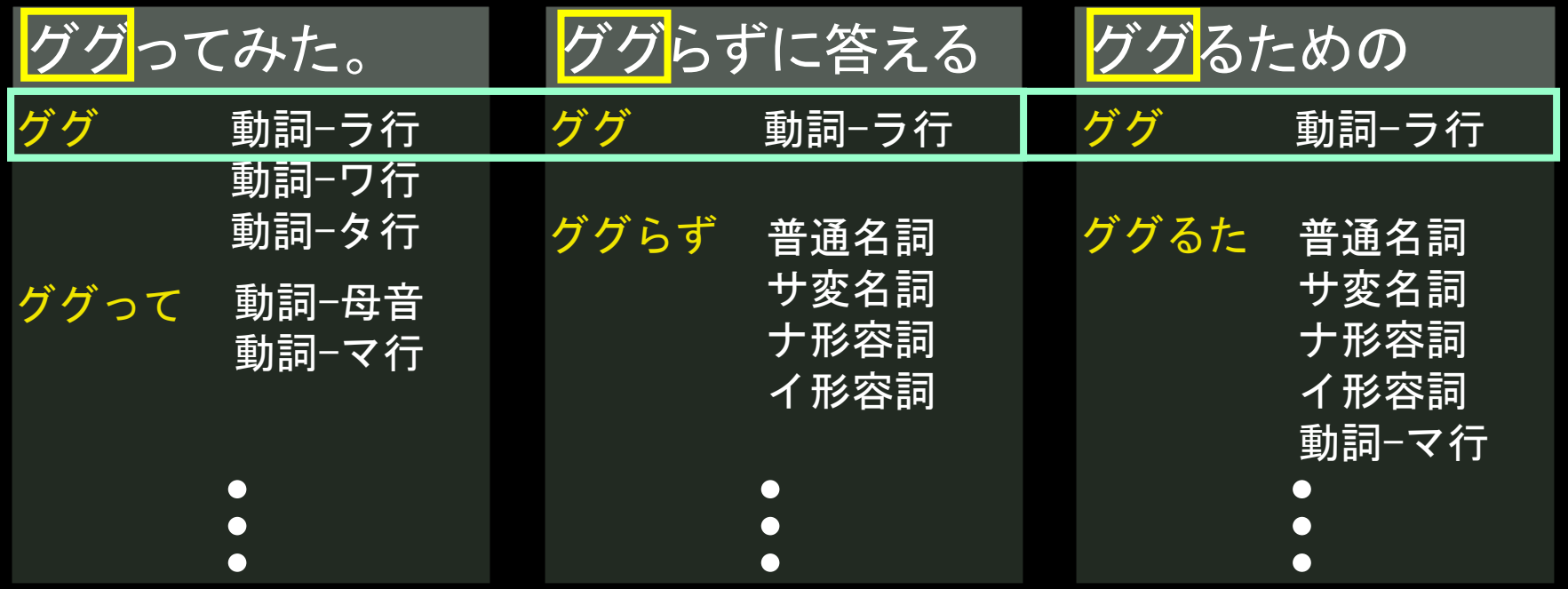
従来手法: バッチ処理



1. コーパス全体をソート
2. コーパスから任意の文字列 (語幹候補) を抽出
3. 前後の環境 (文脈) ベクトルにより品詞を決定

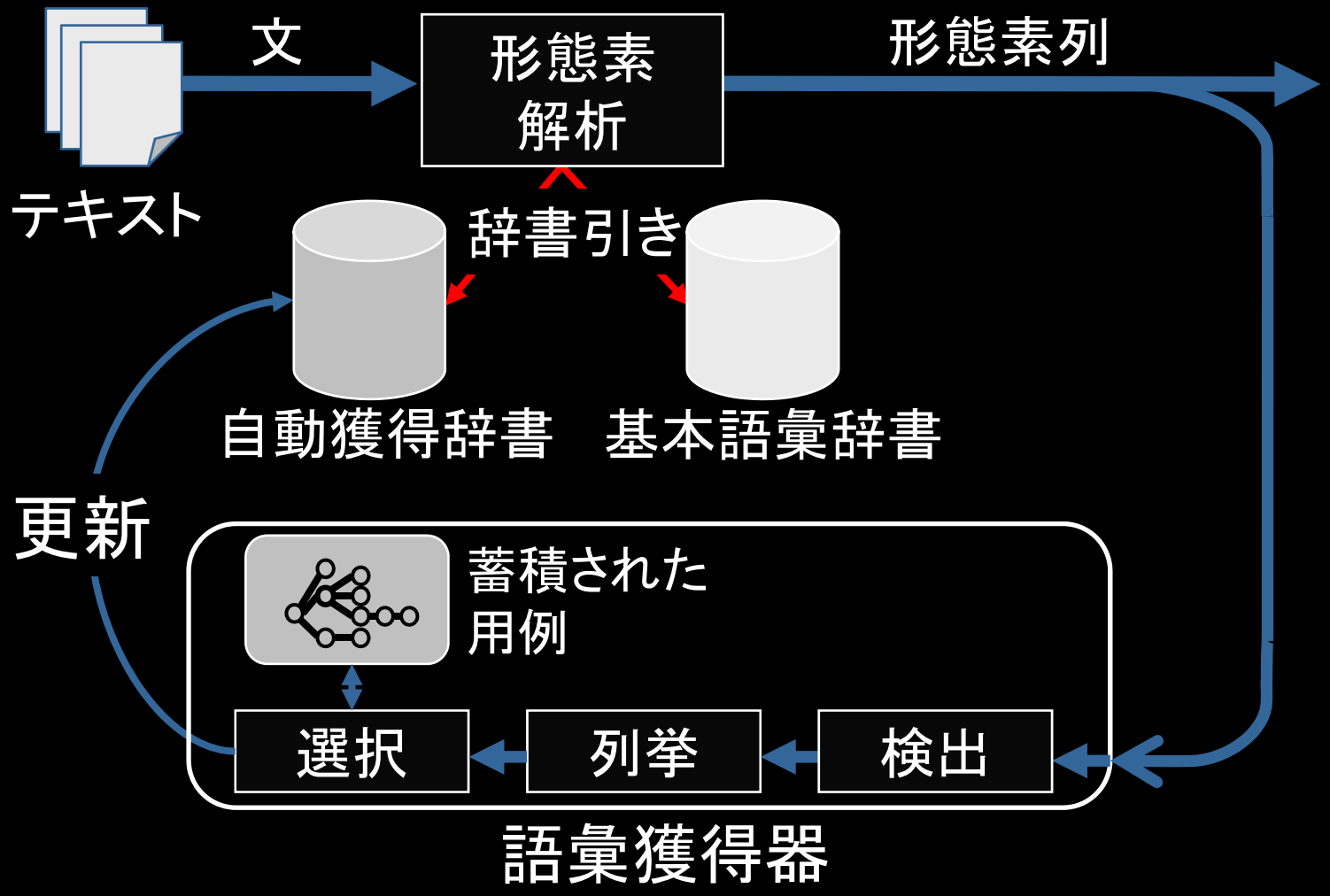
[Mori+ 1996]

提案手法: オンライン処理



- 用例を蓄積し、適宜取り出して比較
- 十分に曖昧性が解消されたら獲得

オンライン未知語獲得


文
形態素
解析
形態素列
テキスト
辞書引き
自動獲得辞書
基本語彙辞書
更新
蓄積された
用例
選択
列挙
検出
語彙獲得器

実験結果

- ウェブページ1,000件からの獲得数: 74-460
- 獲得語の精度: 97.4-98.5%
  - 名詞: 94.1-100%
  - カタカナ語: 67.9-79.4%
- 獲得時点で利用した用例数: 4-7 (中央値)
- 獲得により形態素解析の精度が改善

未知語の獲得例

- ついったー:普通名詞
- 倖田來未:普通名詞
- ググ-る:動詞-ラ行
- ムカつ-く:動詞-カ行
- 喰ら-う:動詞-ワ行
- うざ-い:イ形容詞
- けいはんな:普通名詞

オンライン化の動機

- 生のテキストはいくらでも手に入る
  - 一気に処理するには大きすぎるし、その必要はないはず
- 対象テキストをあらかじめ決めたくない
  - 新しいテキストが日々生まれている
  - どんな大規模テキストにも低頻度語は存在
    - cf. Zipfの法則
- コーパス本位ではなく語彙本位の獲得
  - 入力テキストを徐々に増やす
  - 曖昧性が解消されたら獲得

言葉を覚えるボット

- せっかくオンライン化したのになかなか活かせない
- Twitterに注目
  - 流行っている (@murawaki)
  - ジャーゴンが使われる
    - ついったー, ふぁぼったー, ばずる, ふぁぼる
  - バッチ処理は不向き
    - リアルタイムにつぶやかれる

Twitter / Home - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://twitter.com/

zeregtsegch 「ばずる:子音動詞ラ行」を覚えた。
1:33 AM Dec 26th from web

hazuma @yuzuruu 笹本優子さんって、じつはぼくの妻のの詩の朗読に参加していただいたことがあり、そのときご挨拶しているのです。という点でも懐かしかったりしてw
1:33 AM Dec 26th from web in reply to yuzuruu

yuzuruu @hazuma あらまぁ^^;
1:33 AM Dec 26th from web in reply to hazuma

yukatan 小田和正がばずってる。メジャー音楽の祭典を見ながら、きょう載せた記事の「来年はレコード会社が3社ぐらいに減るのでは」という@tsudaさんのお話を思い出す。レコード会社がなくなっても音楽はもちろん、なくならないとはいえ・・・ http://ow.ly/Pziq
1:32 AM Dec 26th from HootSuite Reply Retweet

jienotsu 貴闘力
1:32 AM Dec 26th from movatwitter

Done