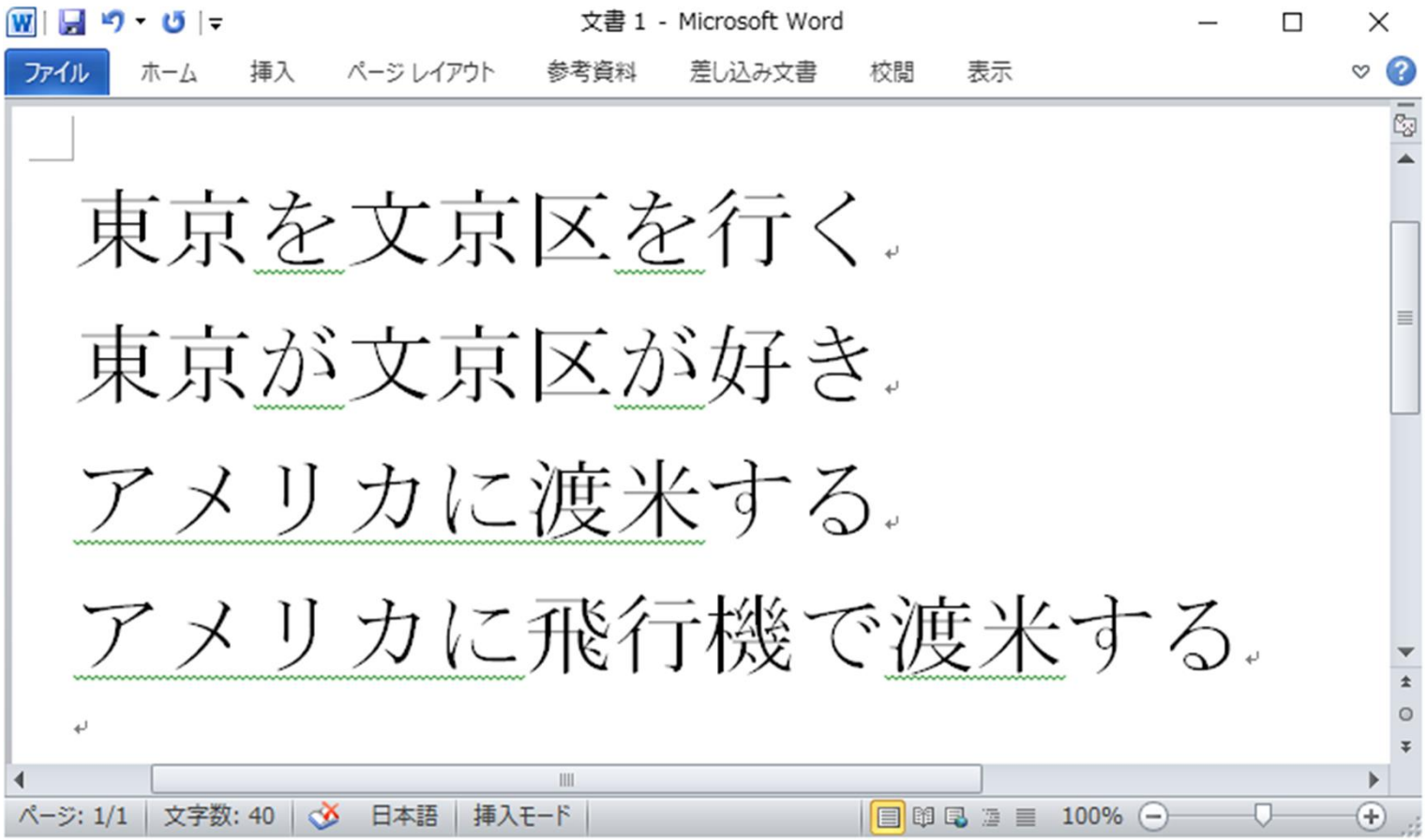
筑波大学
University of Tsukuba

文書校正におけるReviewと活用するための分類およびルール化

筑波大学大学院
津田　和彦

2019/11/01　13:30-14:20

文書校正はご存じですか？？



ソフトウェアレビューの一課題

制作物の多様性

・ソフトウェアの多様性
・各工程の成果物の多様性
・書式や記載方法の多様性

関係者の多様性

・背景知識・スキルの多様性
・興味の多様性
・コミュニケーション問題

規則性が乏しい＝ルール化が困難

暗黙知 → 形式知

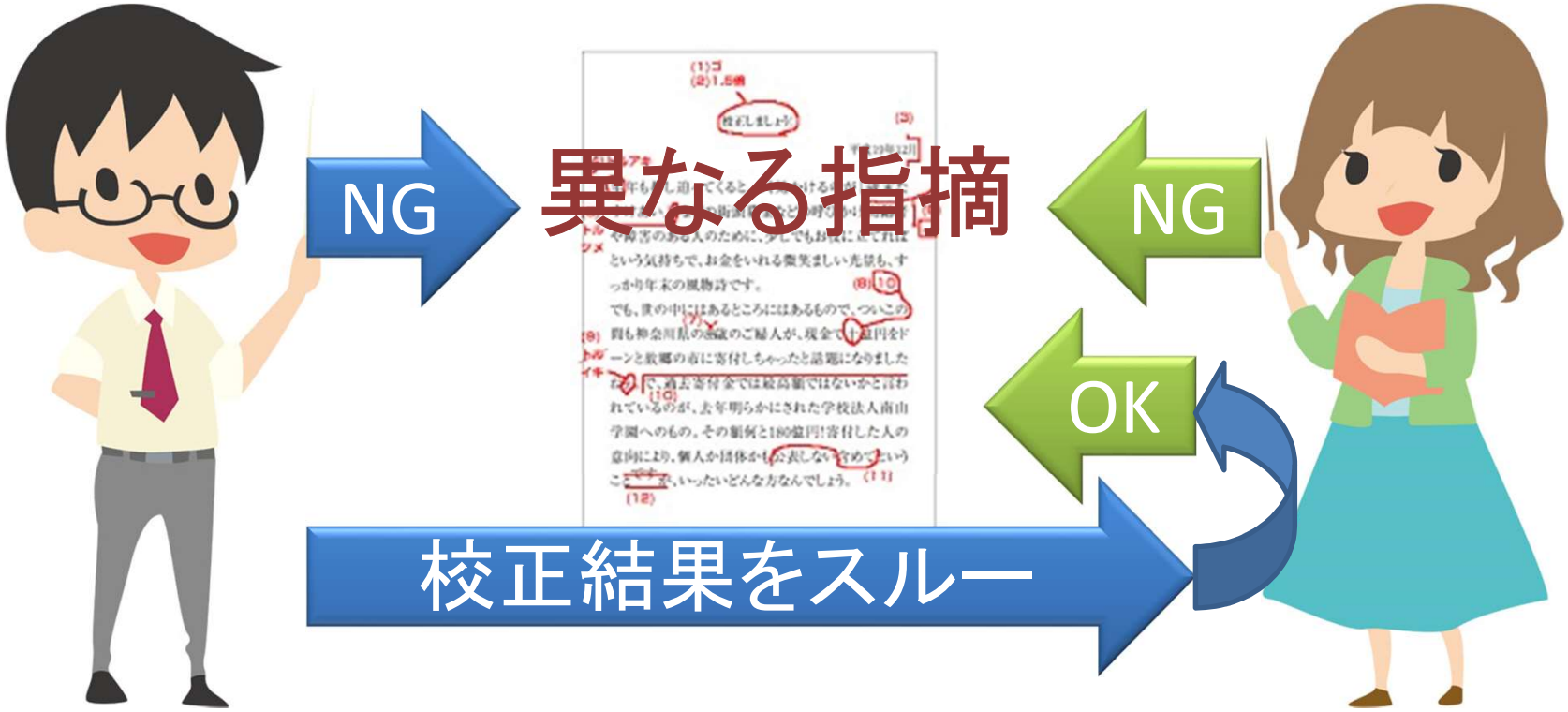
ソフトウェアレビュー課題と文書校正課題

ソフトウェアレビュー		文書校正
ソフトウェアの多様性	＜	文書話題の多様性
各工程の成果物の多様性	＜	作成文書の多様性
書式や記載方法の多様性	＜	書式や記載方法の多様性
背景知識・スキルの多様性	＜	執筆者の背景知識・スキルの多様性
興味の多様性	＜	執筆者の興味の多様性
コミュニケーション問題	＞	執筆者は1人だが…

文書校正が形式知できるのなら
ソフトウェアレビューも形式知に？

文書校正支援機能開発の動機

- 英語では，SpellCheckは当然の機能
- 国語が苦手　＝　コンピュータで指摘して欲しい
  →誰が訂正しても同一の指摘？？



文書校正には合格ラインがある？？

日本語文書校正支援システムの歴史

- 1990年頃　新聞社が誤字チェックを要望
  - 昔，皇室の方の誤植を見つけ朝刊を止めた
  - 誤字チェック，使用不可語チェック，文法チェック
- 1990～1992年頃　雑誌社が興味を示す
  - 原稿文字数を短縮するシステム
- 1992～1994　MicrosoftがSpell Checkに興味を示す
  - 世界で共通の機能・サービスを提供
  - Microsoft Office構想
  - 日本語対応は日本法人での開発が始まる

プロ用（新聞社・雑誌社）向けの校正支援1

- 文長，文節長のチェック
  - 文長は25文字，40文字が基本
- 旧字のチェック
  - 渡辺
  - 渡邊
  - 渡邉

パーソナルメディア株式会社HPより引用

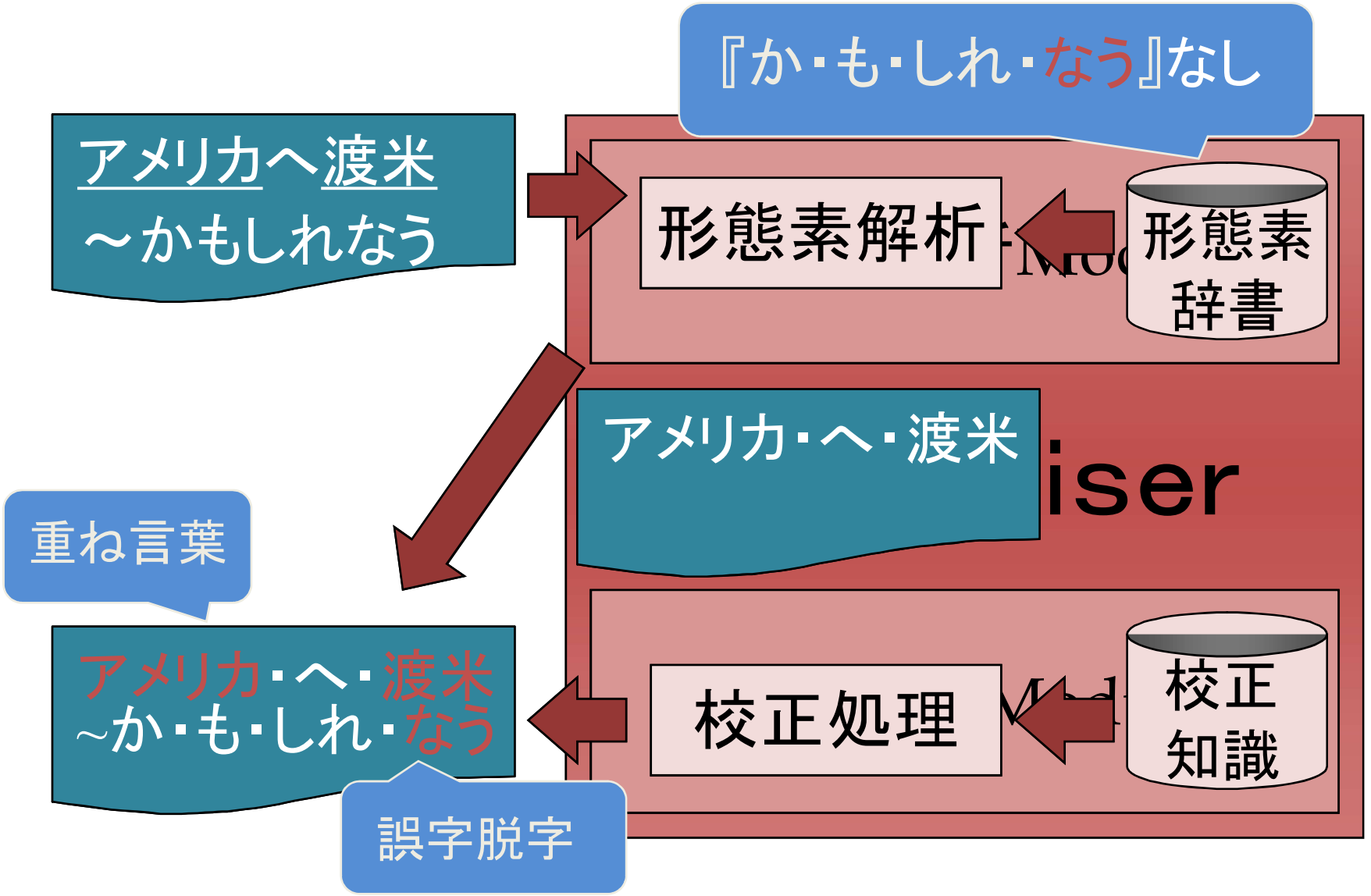
プロ用（新聞社・雑誌社）向けの校正支援2

- 使用禁止用語のチェック＆言い換え表現への置換
  - “土方” ⇒ “建設作業員”
  - “支那そば” ⇒ “中華そば”
  - “めかけ”, “妾” ⇒ “愛人”
  - “満州” ⇒ “中国東北部”
- 課題： 単純な文字列照合では実現できない
  - 「土方さんは・・・」 人名なので○
  - 「支那そばや」自体が店名なので○
  - 「満州事変」⇒「中国東北部事変」は×
  - 「法律上パチンコ店の・・・」の部分文字列内

プロの校正（赤ペン付き）原稿の確認

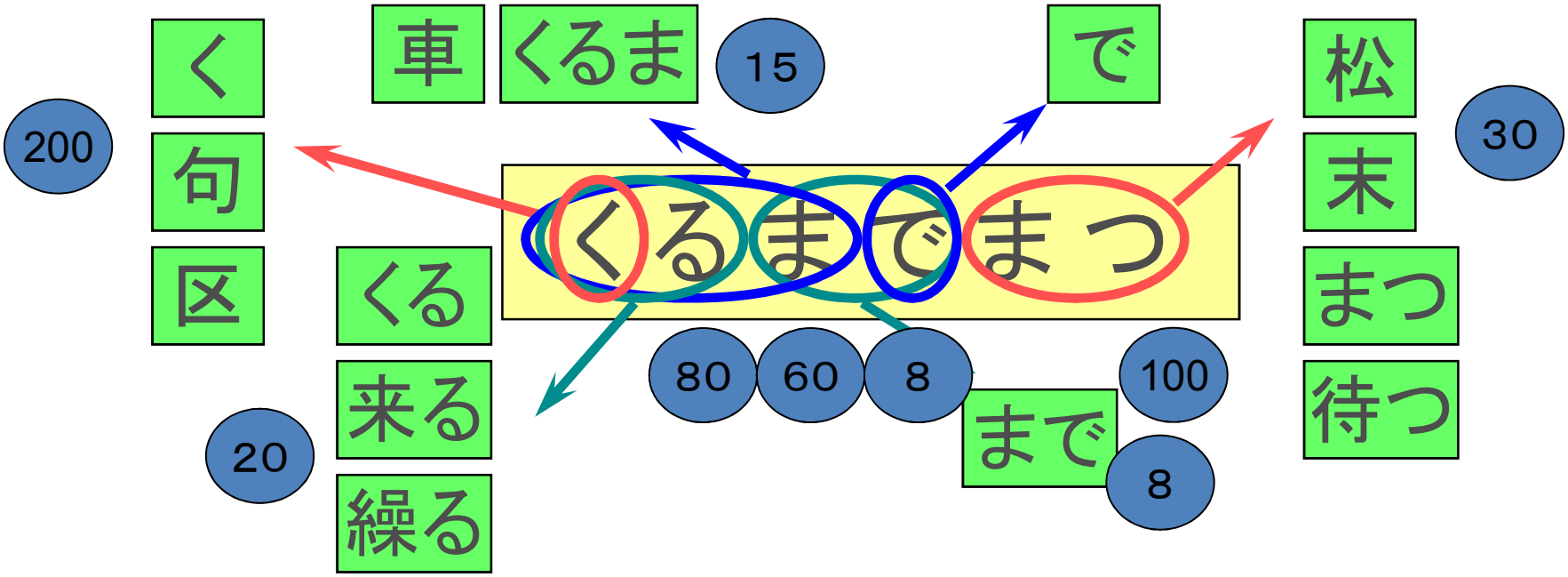
- 移動時間を利用して原稿の確認：
  - 月火:会社(大阪),水木:客先(東京),金:大学(徳島)
  - PC-9801NV(2ndバッテリ込み)約4Kgと校正原稿を（キングファイル2冊）を持って移動
- 文法の重要性を確認
  - 1Sentence 1meaningと格文法

日本語文書校正支援のフロー



形態素解析　最長一致法

- 前方から長い単語を優先的に切り出す手法
  「車／で／待つ」？　「来る／まで／待つ」？



形態素解析　コスト最小法

- 下記，資料参照

MeCab 汎用日本語形態素解析エンジン（工藤 拓）
https://www.jtpa.org/wp-content/uploads/2014/06/MeCab.pdf

予算不足

- 形態素解析を「最長一致法」から「コスト最小法」へ変更したことによるコストの増大
- チェック機能を拡張したことによるコストの増大
  - 商標
  - 当て字：滅茶苦茶（メチャクチャ），珈琲（コーヒー）
  - 同音語：カガク（科学/化学），シリツ（私立/市立）

- 新聞社＆雑誌社が販売に同意
- Microsoft社
  - 日本語部分は日本法人で開発
  - 世界共通の機能サービスの提供

販売対象
Microsoft

素人の校正（赤ペン付き）原稿の確認

- 移動時間を利用して原稿の確認：
  - 月火:会社(大阪),水木:仲間(東京),金:大学(徳島)
  - PC-9801NV(2ndバッテリ込み)約4Kgと校正原稿を（キングファイル1冊）を持って移動

【課題】

赤ペン付き原稿がない

↓

ないなら作る

自分で

単語単位の間違い

- カタカナ文字
  - （バ＝ヴァ）　　（バ≠ヴァ）
    バイオリン　○　　バイク　○
    ヴァイオリン○　　ヴァイク×
  - 文字単位の優先順位付け
    （バ＞ヴァ）　ヴァイオリン（∋バイオリン）
    　　バイク
- 漢字
  - 対象物により異なる漢字を利用
    酒を作る×　酒を造る○

慣用句の間違い

間違い	正解
頭をかしげる	首をかしげる
汚名挽回	汚名返上
舌の先の乾かぬうちに	舌の根の乾かぬうちに
取り付く暇がない	取り付く島がない

副詞の呼応の間違い

	間違い	正解
たぶん、おそらく	する	するだろう
まるで	でした	ようだ
もしかすると	です	かもしれません
決して	すべきです	ない

ミスの種類と分類

- 文法ミス
  - 助詞の用法，２重否定，受動態
- 言葉づかい
  - 重ね言葉，あいまい表現，副詞の呼応，慣用句
- タイプミス
  - ‘U’と‘I’，　‘I’と‘O’
- かな漢字変換ミス
  - “ふかけつ”　⇒“不可欠”が○なのに“不可決”

ミスの形式化　⇒　文書短縮のルールにて説明

文書短縮のルール化（同音異義語への置換）

- 優秀にもかかわらずA大学に落ちた
- 折鶴を作りはじめようとしている

［x_1：接続詞，助述…］ → ［y_1：x_1と同品詞・概念］

品詞・概念	形態素
接続詞・背反	しかし，けれども，が，ところが，なのに，でも，にもかかわらず，とはいえ，さりとて，されど，しかるに
助述・起動直前	ところだ，つつある，はじめようとしている，はじめつつある，かけている，

文書短縮のルール化

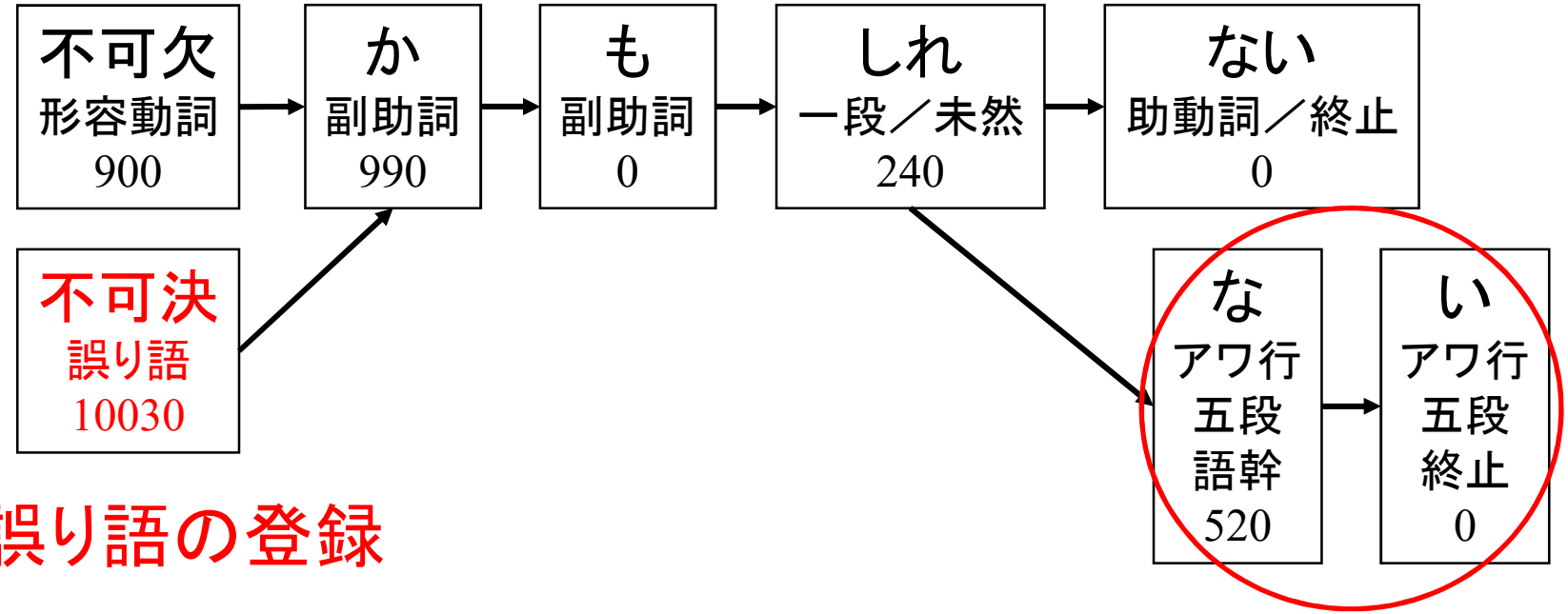
- Intelプロセッサーと AMGプロセッサーでは・・・
- レベル1とレベル2の差は・・・

[x_1：接頭辞or名詞]
[x_2：名詞]
[x_3：接続詞orカンマ]
[x_4：x_4=x_1]
[x_5：名詞]

[x_1：接頭辞or名詞]
[x_2：名詞]
[x_3：接続詞orカンマ]
[x_5：名詞]

誤字脱字チェック

- 形態素解析のコスト増大によりチェックする

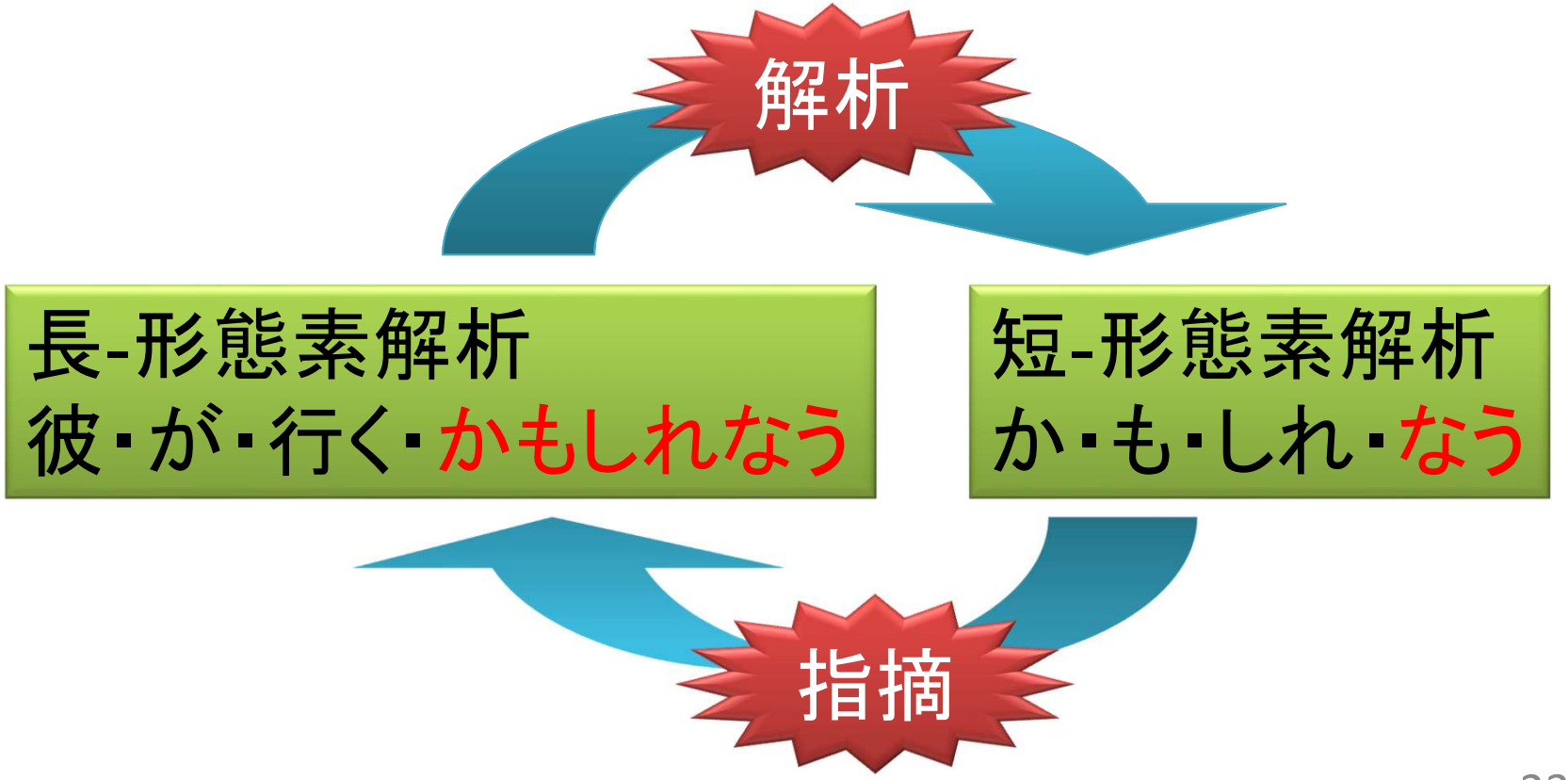


誤り語の登録

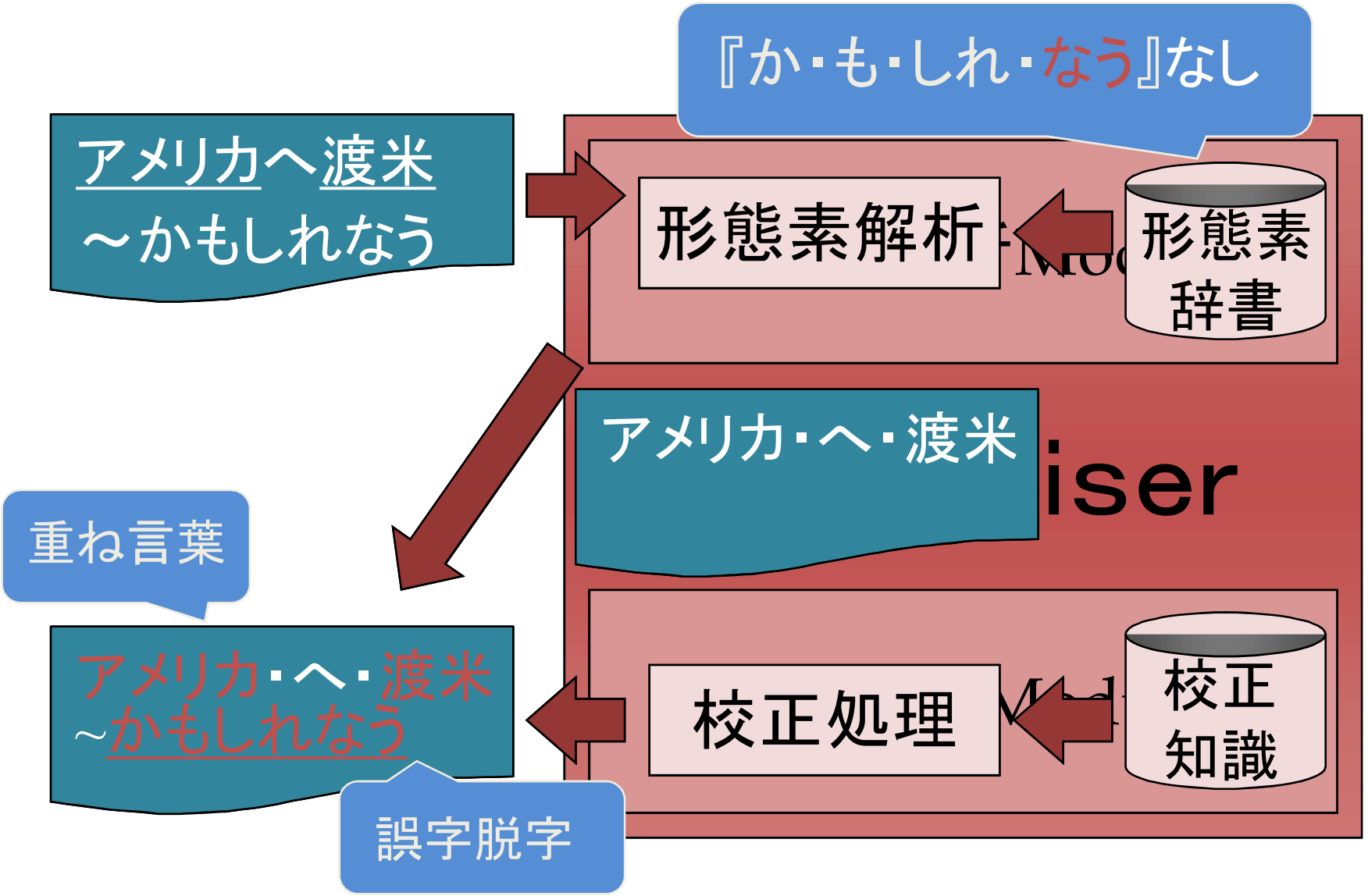
単文字の連続

誤字・脱字

- 長短-形態素を登録
  - Example: かもしれない（長-形態素）
  - か・も・しれ・なう（短-形態素）



日本語文書校正支援のフロー



Add OnからAdd Inへ

- WORD　58,000円，SpellViser　12,800円
- １万本売れればAdd OnからAdd Inへ

実現した文書校正支援の機能

誤字脱字	表現誤り
当て字	かなづかい
まぎらわしい同音語	商標，商品名
新字体，旧字体	送りがな
かな書き推奨	正式名称と略称
単位	くだけた表現
文語調	重ね言葉
副詞の呼応	あいまいな表現
二重否定	助詞の用法確認
繰り返し使用語句	指示語
受動態	文体

30年前の開発過程を思い返して

- ノートパソコンとキングファイルは重かった・・・

- 校正者Aと校正者Bは，異なる校正結果となるが，互いの校正結果への指摘はない

校正には合格点／OKの閾値がある　？

- 100点は望まず，80点を目指した開発
- 地道な確認作業からボトムアップ式で分類は構築

大半は地道な原稿の確認作業

まとめ

- 執筆原稿もソフトウェアドキュメントも正解はない

- 正解がないなら，間違いを取り除く
- 見るのは全体，作業は局所
- 地道な作業，成果は1歩づつ