日本語学習者コーパス検索ツールの開発

Development of a search tool for Japanese-learner corpora

浅尾 仁彦 (京都大学大学院)　　李 在鎬 (情報通信研究機構)

asaokitan@ling.bun.kyoto-u.ac.jp

背　景

- 学習者コーパスの需要の高まり
  - 文字列ベースの検索では不十分
  - 形態素解析のコスト

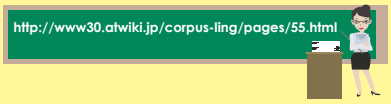
▼

- KY コーパス (鎌田・山内 1999) に基づくタグ付き学習者コーパスの作成 (李ほか 2008)
  - 品詞情報(茶筌)
  - 意味分類(分類語彙表)
  - 誤用や言い直し(手作業)
- 検索ツールの必要性

開　発

「E-KWIC　KY コーパス用」

- Excel 上で実行できる
  - 広い層の研究者が利用できる
- 検索機能
  - 基本形による単語の検索
  - 品詞／意味分類による検索
  - 正用／誤用別の検索
  - 学習者母語の絞り込み
  - 学習者レベルの絞り込み
- 検索結果を Excel のワークシートとして保存できる
- 発表者ウェブサイトで公開，フィードバックを得て継続的に改善



KY コーパス (オリジナル)

T: 町の感じとかどうですか, 人の感じとか
S: はい, ここ, 大阪の, 人はたいへん親切だし, 物価はちょっと高いんですけど, 便利, 交通も便利だし, はい, そんなに寒くないだし, 私ここが好きです, 〈あーそうですか〉はい
…

▼

KY コーパス (加工済み)

寒く	サムク	寒い	助動詞
ない	ナイ	ない	助動詞
だ	ダ	だ	助動詞-誤用
し	シ	し	助詞-接続助詞
、	、	、	記号
私	ワタシ	私	名詞-代名詞

▼

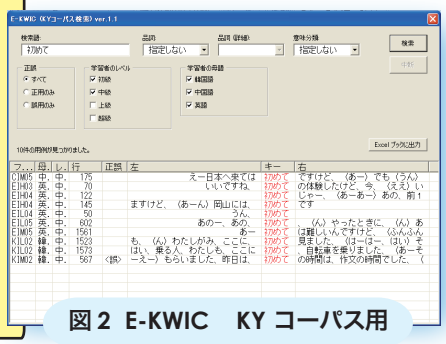


図 2 E-KWIC　KY コーパス用

茶まめ用

- 「茶まめ」(小木曽ほか 2007)
  - 簡単な形態素解析ツール
  - 辞書「UniDic」を利用
  - 解析結果を Excel に出力できる
- 解析結果を利用した検索ツールがない

▼

「E-KWIC 茶まめ用」

- 「E-KWIC　KY コーパス用」から学習者コーパス向け機能を除いたもの
- 手持ちのデータに自由に適用できる

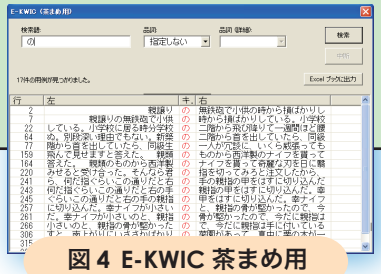


図 4 E-KWIC 茶まめ用

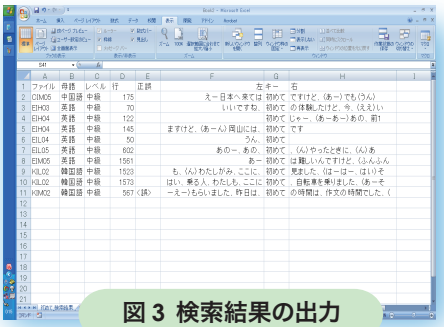


図 3 検索結果の出力

利 用 例

- 「の」の誤用
  - 「古いの切手」「小さいの弟」

母　語	正　用	誤　用
韓国語	1,602	58
中国語	2,018	117
英　語	1,634	35

- 中国人学習者に有意に多い ($p<.001$)
- 学習者レベルが上がると改善
  - 初級…誤用 10 件 (15.6%)
  - 中級…誤用 102 件 (13.8%)
  - 上級…誤用 77 件 (6.5%)

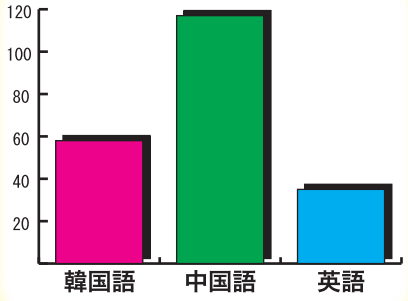


図 1 誤用の「の」の数 (母語別)

文　献

[1] 李在鎬・淺尾仁彦・濱野寛子・佐野香織・井佐原均 (2008). 「タグ付き日本語学習者コーパスの開発」. 『言語処理学会第14回年次大会発表論文集』 pp. 658-661.

[2] 鎌田修・山内博之 (1999). 「KY コーパス」. Ver 1.1. (http://opi.jp/shiryo/ky_corp.html)

[3] 小木曽智信・小椋秀樹・伝康晴 (2007). 「日本語研究に適した形態素解析ソフトウェア―「UniDic」と「茶まめ」―」. 『日本語学会 2007 年度秋季大会予稿集』 pp.255-262. (http://www.tokuteicorpus.jp/dist/)

JSLS2008. 2008年7月12-13日. 於静岡県立大学