日本語学習者コーパス検索ツールの開発

淺尾仁彦（京都大学）・李在鎬（情報通信研究機構）

Development of a search tool for Japanese-learner corpora

Yoshihiko Asao (Kyoto University), Lee Jae-Ho (NICT)

本発表では，発表者の開発したコーパス検索支援ツール E-KWIC を事例を交えて紹介する。E-KWIC には，学習者コーパスの検索に特化した「KY コーパス用」と，研究者の手持ちのデータに自由に適用できる「茶まめ用」との 2 種類が用意されている。これらは Microsoft Excel 上のマクロとして動作するため，多くの研究者にとって導入が容易となっている。

従来，「KY コーパス」(鎌田・山内 1999) をはじめとする様々な学習者コーパスの構築が試みられ，こうしたコーパスを利用することで，豊富な研究成果が挙げられてきた。しかしながらこの種の研究の多くは，単純な文字列ベースの用例検索に依存している場合が多いという難点があった。発表者らのグループは，形態素解析ツール「茶筌」を用いて，KY コーパスに形態素情報を付与したうえ，人手によって誤用や言い直しなど習得研究に有用な追加情報を付与したデータを構築した (李他 2008)。「KY コーパス用」 E-KWIC は，このデータを検索するためのツールである。

本ツールでは語の基本形による検索に加え，品詞や意味分類による検索，さらに正用または誤用のみの検索，学習者の母語やレベルを限定しての検索が可能となっている。また，ユーザは検索結果を Excel に出力し，編集・加工することができる。

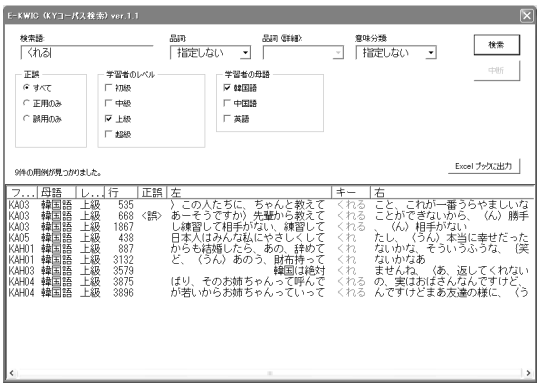

図 1　E-KWIC (KY コーパス用) による検索例

李他 (2008) のフォーマットに合わせて加工することで，ユーザの手持ちの学習者データの分析に役立てることもできる。

また，「KY コーパス用」E-KWIC から学習者コーパス向けの機能を除いた「茶まめ用」E-KWIC を作成した。「茶まめ」(小木曽他 2007) は形態素解析を手軽に行うためのツールであるが，その結果を利用して検索を行うツールは存在していなかった。本ツールを用いると，ユーザの手持ちのデータに対し，形態素解析を用いた検索を簡単に実行でき，習得研究のみならず日本語研究の幅広い分野で役立てることができる。

本ツールは発表者のウェブサイト http://www30.atwiki.jp/corpus-ling/pages/55.html で公開されており，利用者からフィードバックを得ながら継続的に改善していく。

参考文献

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.

鎌田修・山内博之 (1999). 「KY コーパス」. Ver 1.1.

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.