『コーパス日本語学ガイドブック』の概要

（同書の抜粋）

コーパス日本語学ガイドブック

田野村忠温・服部匡・杉本武・石井正彦

特定領域研究「日本語コーパス」日本語学班

2007

はしがき

平成１８～２２年度文部科学省科学研究費補助金特定領域研究「代表性を有する大規模日本語書き言葉コーパスの構築：２１世紀の日本語研究の基盤整備」（領域代表者：前川喜久雄氏、領域略称：「日本語コーパス」）を１つの柱とする国立国語研究所の日本語書き言葉コーパス構築５か年計画が現在進められている。これにより、2011 年には待望の現代日本語の大規模な均衡コーパスをついに誰もが利用できるようになる。

電子媒体の言語資料の特性を生かす形でコーパスを日本語研究に利用するにはそれなりの知識や技術が必要となる。本書は、特定領域研究「日本語コーパス」の計画研究班の１つである日本語学班のメンバー４人がこれまでの経験を踏まえ、日本語研究にコーパスを役立てようとする方の参考になりそうな初歩的な情報を提供するものである。どこからでも読める形にしてあるので、興味のあるところから読んで実際にパソコン上で試していただければ幸いである。

第１部本編では、日本語研究者が自ら独自の目的・方法でコーパスを利用できるようになるための方法や手順などを紹介する。第２部各種解説編には、日本語コーパスを利用するうえで必要となる一般的な知識や既成のソフトウェアに関する情報などを収めた。コーパスを研究に生かしたいがどこから手を付けてよいか分からないという方にとって本書がコーパス処理の世界への道案内となることを願っている。

本書の各章が直接の処理対象とするのはいわゆるプレーンテキストである。一般に利用可能な付加情報付きコーパスがない現状では好むと好まざるとにかかわらず選択の余地がないわけであるが、プレーンテキストには、豊富に入手でき、扱いが容易で、特定の文法解釈に制約されることがないという利点がある。プレーンテキストという形態でのコーパスは、付加情報付きのコーパスが作られその利用が普及した後も日本語研究の精密化に重要な役割を果たし続けることであろう。

なお、本書では現在日本語研究者に最も広く使われていると思われる OS である Windows XP 日本語版の使用を想定している。もっとも、本書の内容の多くは Windows のバージョンには依存しないはずである。また、日本語テキストの文字コードは Shift-JIS であることを前提としている。

本書および添付 CD の内容に関して更新や訂正の情報があれば特定領域研究「日本語コーパス」の Web サイトにある日本語学班のページ（http://www.tokuteicorpus.jp/g_jpling/）に掲載していく予定である。不審な点などがあればまずはそちらをご覧いただければ幸いである。

田野村 忠温

2007 年 9 月

目次

第１部　本編

第２部　各種解説編

日本語KWICソフトウェアKWIC

田野村 忠温

コーパス利用の普及のためには簡単に使えるコーパス処理ソフトウェアの存在が不可欠だと思われるが、残念ながら日本語に関しては研究者が手持ちの電子テキストをそのまま使って簡便にKWIC索引を生成することのできるソフトウェアがない。

そうした状況にかんがみ、このたび日本語KWICソフトウェアを作成してみた。ここではその機能・用法と、検索結果の利用に関する簡単なヒントについて述べる。

内容

AWK プログラミング入門

田野村 忠温

テキスト処理に適したプログラミング言語として AWK と Perl、ほかには Python や Ruby が広く使われている。ここでは AWK とソートプログラム sortf を使ってテキストから文字列を検索したり KWIC 索引を生成したりする方法を紹介する。例題では英語のテキストを処理対象とするが、日本語のテキストを処理する場合にも基本的な考え方はそのまま通用する。

AWK はほかの言語に比べて機能が限られているが、手軽に学べるという利点がある。また、AWK を学んだ経験は、さらに進んで Perl その他の言語を学ぶときにも大いに役に立つであろう。

内容

Perl・sortf・秀丸エディタを用いた用法分析の一例

服部　匡

Perl スクリプトによって簡単な KWIC 索引を生成する。それを利用し、Perl・sortf・秀丸エディタを組み合わせて語の用例の仮分類や計数を行っていく方法の一例を紹介する。

内容

perl によるテキスト処理

杉本　武

コーパスを用いた日本語研究には、テキスト・データが不可欠であるわけであるが、現在では、一般的には、各紙の新聞データや青空文庫のような小説のテキスト・データが利用できる。ただし、これらのデータは、必ずしも日本語研究のために作成されたものではないため、検索などの処理をするためには扱いにくい点があることがある。

ここでは、上のような一般的に入手できるテキストデータから必要な情報を抽出、整形し、扱いやすいデータを作成する perl のスクリプト、また、そのデータを検索するための簡易 KWIC のスクリプトを紹介する。

内容

Excel で語彙表を作成する

石井 正彦

難しいプログラムや特別なソフトを使わず、手軽に語彙表を作ることはできないか。こうした要望に応えるべく、Excel（Microsoft Office Excel 2003）を使って語彙表を作成する方法を、新聞コーパスを例として、まとめてみた。全体を８回に分け、各回とも、「今日の目標」「解説」「新出の技法」（「準備」）「作業の手順」という構成にした。なお、Excel による語彙表の作成は、伊藤雅光『計量言語学入門』（大修館書店、2002）も行っているが、伊藤の方法はあらかじめ作成した語彙表のデータを Excel に読み込んでいろいろな作業を行うのに対して、ここでは、Excel 自身に語彙表を作らせるようにしている。

内　　容

特定領域研究「日本語コーパス」平成 19 年度研究成果報告書（JC-L-07-01）

コーパス日本語学ガイドブック

田野村忠温・服部匡・杉本武・石井正彦

2007(平成 19)年 9 月 7 日発行

文部科学省科学研究費補助金特定領域研究「日本語コーパス」日本語学班

〒562-8558 大阪府箕面市粟生間谷東 8-1-1 大阪外国語大学 田野村忠温研究室

電話・FAX: 072-730-5177

E-mail: tanomura(at)osaka-gaidai.ac.jp (2007年9月30日まで)

tanomura(at)let.osaka-u.ac.jp (2007年10月1日より)