

結論のおよぶ範囲…144/母集団と標本…144/コーパス日本語学と検定…145	
11. 有意差の考え方…147	147
帰無仮説とは…147/検定が分かりにくい理由…147	
12. カイ自乗検定：新旧字体間の頻度差…148	148
なぜカイ自乗検定を取り上げるのか?…148/なぜウェブ統計ソフト STAR なのか?…149/ウェブ統計ソフト STAR の利用…150/クロス表の検定…153	
13. 相関係数の検定：使用頻度の時代変化…158	158
検定の目的…158/分布の形の確認…158/相関係数の検定…159	
14. まとめ…160	160

第5章 コーパスを利用した研究例 間淵洋子 165

1. コーパスと日本語学…165	165
コーパス日本語学とは…165/コーパスの研究利用とコーパス日本語学の現状…166	
2. コーパス日本語学の分野分類…168	168
3. 音声・音韻…169	169
『日本語話し言葉コーパス』を用いた音声・音韻研究…169/その他のコーパスを用いた研究…174	
4. 文字・表記…177	177
新聞記事データベースを用いた研究…177/新聞以外のコーパスを用いた研究…180/文字・表記分野でコーパスを利用する際の留意点…182	
5. 語彙・用語・文法…183	183
コーパス利用の多様性…183/新聞記事データベースを用いた研究…184/書籍データを用いた研究…191/サンプルコーパスを用いた研究…195/複数の資料を用いた位相研究…198/専門文章のコーパスを用いた研究…203/現代語以外のコーパスを用いた研究…205/話し言葉のコーパスを用いた研究…209	
6. 文章・文体…213	213
文章の特徴を明らかにする研究…213/語の文体を解明する研究…216/比喩表現に関する研究…218	
7. コミュニケーション…219	219
談話研究と対話コーパス…219/その他の研究…224	
8. これからのコーパス日本語学…225	225

**コーパスデータの作成
—OCR ソフトを利用して—**

はじめに

本稿では、OCR ソフトを利用したコーパスの作成法、及び、その注意点について紹介する。OCR ソフトを使ってコーパスを作る方法にはいくつかのものがあるが、ここでは、最も一般的な、書籍をイメージスキャナ（以下、単に「スキャナ」とよぶ）で読み取って電子化する方法についてみていく。

以下、1節でOCRとはどのようなものか簡単に紹介し、2節でOCRソフトを使うことの利点について述べたうえで、3節で現在入手可能なハードウェアとソフトウェアについてみる。そして、4節でOCRソフトを使った作業の流れを紹介し、5節で各種文書をOCRソフトで認識した結果を報告する。最後に、6節でOCRソフトでコーパスを作成するうえで重要な著作権について簡単にふれる。

1. OCRソフトとは

OCRソフトのOCRとは、Optical Character Readerの略称である。ウェブ上の辞典、IT用語辞典 e-Words (<http://e-words.jp/>) には次のようにある（2010年6月22日更新）^(注1)。

光学式文字読取装置。手書き文字や印字された文字を光学的に読み取り、前もって記憶されたパターンとの照合により文字を特定し、文字データを入力する装置。

この「装置」には例えば、郵便物を自動的に振り分けるために郵便番号を読み取ったり、振込用紙の口座番号を読み取ったりするものなどがある。この機能を持った、パソコン用のソフトが「OCRソフト」である。「IT用語辞典 e-Words」には先に引用した部分に続いて次のようにある。

スキャナで読み取った画像から文字を識別して文書に変換する OCR ソフトもある。

ここではスキャナにしか言及していないが、ソフトによっては、この他、PDF 文書やデジタルカメラで撮影した画像などからも文書（テキストファイル、PDF ファイル、各種ワープロファイル）への変換が可能である。この OCR ソフトを用いて言語研究用のコーパスを作成する方法を紹介するのが本稿の目的である。

2. OCRソフトを用いてコーパスを作成することの利点

ここ15年ほどのパソコンの急速な普及^(注2)と歩調を合わせるように、パソコン上で利用できる書籍等もかなり増え、それらは多くの研究者からコーパスとして利用されている。その代表的なものがかつて販売されていた『CD-ROM 版 新潮文庫の100冊』（1995年）であり、また、1997年に「開館」した「青空文庫」^(注3)に収録された作品であろう。さらに、研究向けのコーパスも充実しつつある。朝日、毎日、読売、日経の各新聞社からは、学術研究のための記事データ集が発行されており、国立国語研究所による『現代日本語書き言葉均衡コーパス』の開発が進められてもいる。

このように、日本語研究に利用できるコーパスはすでになんかの分量があり、実際、これらを用いた研究成果も多数発表されている。このことからすると、あらためて自らコーパスを作成する必要性を感じないという向きもあるだろう。

しかしそれでもなお、OCR を用いてコーパスを作成する利点を挙げると

すれば、《好きなものを、好きなだけ》という点に集約されよう。このうち、「好きなだけ」という点は作業時間を確保できる限りにおいて、という制約が加わるが、「好きなものを」という点に関しては大きな利点となろう。

例えば、周知の通り青空文庫は無償で入手できることもあり便利であるが、収録されている作品は著作権の関係上、没後50年以上たった著者のものである^(注4)。したがって、どうしても作品が限られてしまい、研究目的によっては必ずしもメリットがあるとはいえなくなってしまう。

この点を補おうと他のコーパスを求めようとしても、『CD-ROM 版 新潮文庫の100冊』は現在入手できず、中古市場でもかなりの高額となっている。各新聞社の記事データ集も、おおよそ10万円前後と高額なものもあり、個人では（あるいは大学教員の個人研究費でも）入手しにくい。もし入手できたとしても、それとて調査できる対象のバリエーションは限られてしまう。

研究の必要性という観点からはもちろん、好みや思い入れという場合も含め、自らの調査対象として必要なものを入手しようとするれば、OCRソフトを用いて機械可読な形式にしておき、手元に置いておくことは、現在でも必要性があると考えられる^(注5)。

3. 必要なハードウェアとソフトウェア

3.1 用意するもの

OCRソフトを用いてコーパスを作ろうとした場合、OCRソフトの他にパソコンとスキャナが必要となる。

このうち、パソコンについては、現在入手可能な機種であればもちろん、数世代前のパソコンでもおそらく十分であり、特別高性能なもの（あるいはパーツ）は必要ないと思われる。そこで、以下ではスキャナとOCRソフトについてみていく^(注6)。

なお、ここでは、パソコンとスキャナの接続、スキャナのセットアップ、そしてOCRソフトのインストールについては言及しない。適宜、それぞれのマニュアル等を参照されたい。また、パソコンとスキャナはUSBによっ