２００７年５月２８日

（報道発表資料）

独立行政法人国立国語研究所

大規模書き言葉コーパスのオンライン試験公開

～KOTONOHA「現代日本語書き言葉均衡コーパス」～

独立行政法人国立国語研究所（東京都立川市、所長：杉戸清樹）は、現在構築中の『現代日本語書き言葉均衡コーパス』のデータの一部、約 1000 万語分をインターネット上で試験公開します。（http://www.kotonoha.gr.jp/demo/）

国立国語研究所は、明治から現代にいたる日本語の電子化資料をコンピュータ上で公開しようとする KOTONOHA 計画を基幹的なプロジェクトとして推進しています。『現代日本語書き言葉均衡コーパス』（コーパスとはコンピュータ上に蓄積された大規模な言語資料のこと）はその一環として昨２００６年度から構築を開始したものであり、２０１１年の完成時には１億語を超える量の現代日本語の書き言葉データとして公開する予定です。

試験公開の目的

『現代日本語書き言葉均衡コーパス』をだれもが利用できるコーパスとするためには、全サンプルに著作権処理を実施する必要があり、その総数はおよそ３万件に及ぶものと想定されます。しかし昨今では、著作権保護、個人情報保護意識の高まりを反映して、著作権者との連絡にかかるコストが著しく増大する傾向にあり、著作権処理の成否がプロジェクト全体に大きく影響する状況となっています。

今回オンライン試験公開を開始する目的のひとつは、著作権者の方々にこのデモサイトを利用

していただき、ご提供いただくサンプルが実際にどのような形で利用されるかについて理解を深めていただくことにあります。また、不特定多数の方々に対して KOTONOHA 計画で開発中のコーパスに関する情報を提供することも、もうひとつの大切な目的です。

今回公開するデータ

現時点で検索することのできるデータは、各省庁刊行の白書のデータ約 500 万語分と、ヤフー株式会社提供の「Yahoo!知恵袋」のデータ約 500 万語分の合計約 1000 万語です。それぞれについてもう少し詳しく説明します。

白書データの母集団は 2001 年から 2005 年の間に発行された白書と過去 30 年間に継続して発行されつづけた白書の全体です。この母集団から無作為に約 500 万語分を抽出しています。サンプルは白書の対象にしたがって 9 個のカテゴリー（「安全」「科学技術」「外交」「環境」「教育」「経済」「国土交通」「農林水産」「福祉」）に分類されており、カテゴリーを限定した検索も可能です。同様に白書の刊行年を限定した検索も可能です。白書は政府の刊行物ですが、やはり著作権は存在しますので、「著作権フリー」を宣言しているごく一部の白書を除くサンプルについては関係省庁から書面ないし口頭で利用許諾をいただいています。

一方「Yahoo!知恵袋」は、参加者同士で知識を教えあうことを目的とした、Q&A 形式のナレッジコミュニティーサービスです。この種のデータについては、従来から言語研究上の重要性が指摘されてきていましたが、今回、ヤフー株式会社のご厚意により、『現代日本語書き言葉均衡コーパス』のデータの一部として利用できることになりました。現在公開されている「Yahoo!知恵袋」データの総量はそれだけで 1 億語を超える膨大なものですが、今回はそこから 500 万語分の質問と回答を無作為に選択して公開対象としました。

白書のデータと「Yahoo!知恵袋」のデータとはいろいろな面で対照的です。白書が硬い書き言葉のひとつの典型であるのに対して、「Yahoo!知恵袋」はかなりくだけた話し言葉的な特徴を示します。表記や文法についても、白書の文書がよく統制されているのに対して知恵袋は自由奔放といえそうです。

今後の展開

本試験公開サイトには、今後とも著作権処理が完了したデータを追加してゆく予定です。今後追加を予定しているデータには以下のものがあります。

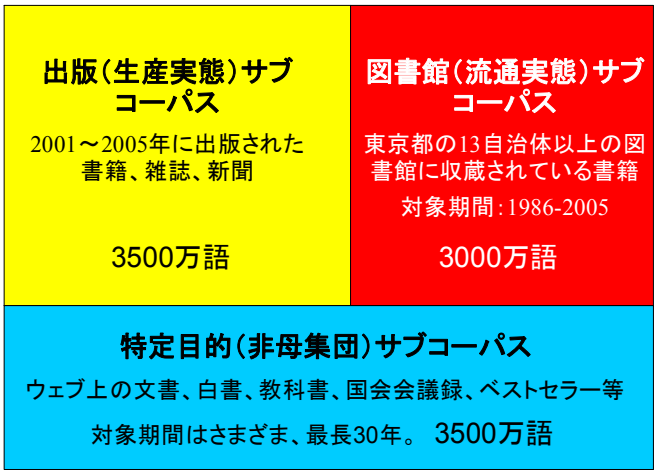
- 国会会議録　　（500 万語程度）
- 新聞記事　　（100 万語程度）
- 文芸作品　　（500 万語程度）

これらが追加されることによって、本試験公開サイトは次第に均衡コーパス（以下の「背景説明」参照）として整備されてゆくことになります。

背景説明：『現代日本語書き言葉均衡コーパス』

現在、世界的潮流として、コーパスを利用した言語研究が活発におこなわれています。このような研究には、対象とする言語の様々なジャンルのデータを偏りなく収集した「均衡コーパス」

を整備する必要がありますが日本語に関するコーパスの整備は、英語圏はもとより中国語や韓国語に比べても著しく遅れているのが現状です。『現代日本語書き言葉均衡コーパス』はこのような問題を解消して我が国の言語研究のインフラを整備するために企画されたものです。またそれと同時に、常用漢字の見直しの問題、公用文の表記法の問題、教育用基本語彙の選択の問題など、日本語の言語政策にかかわる諸々の問題を検討するためにも活用される予定です。



図に示したように、『現代日本語書き言葉均衡コーパス』は三つのサブコーパスから構成されています。左上の出版サブコーパスは 2001 年から 2005 年の間に出版されたすべての書籍、雑誌、新聞を母集団として、そこから合計で 3500 万語分のテキストを無作為に抽出するものです。

このサブコーパスは「どれだけ書き言葉が生産されたか」という観点から構築したコーパスです。書籍の販売部数には大きな差が生じますが、このサブコーパスでは 100 万部を超えるベストセラーも数百部しか売れなかった本も同じ重みで扱われている点に特徴があります。

これに対して、実際にある程度広い範囲に出回ったことが判明している書籍だけを対象としているのが、図右上の図書館サブコーパスです。このサブコーパスでは、東京都下の公立図書館に所蔵されている蔵書のうち、少なくとも 13 館以上に収蔵されている図書（34 万冊ほど）を母集団として、そこから無作為にサンプルを選んでいます。図書館サブコーパスのもうひとつの特徴は、対象期間が 20 年間におよんでいることです。

最後に、図下部の特定目的サブコーパスは、国立国語研究所が種々の研究活動で必要とするデータのうち、上述のふたつのサブコーパスでは十分にデータが集まらないものを格納するものです。例えば公共性の高い文章における外来語や難解語の言い換え提案を行うための資料として白書の本文をデータとして格納しています。また、現代の日本語を考察するためには欠かすことのできないインターネット上の文書のデータもこのサブコーパスに格納しています。

謝　辞

出版サブコーパスおよび図書館サブコーパスのサンプリングには国立国会図書館、日本図書館協会、立川市立中央図書館、都立図書館、一橋大学図書館にご協力をいただいています。知恵袋データはヤフー株式会社からご提供いただきました。文芸作品の著作権処理には、日本文藝家協会、日本推理作家協会、日本児童文学者協会、日本児童文芸家協会、日本ペンクラブにご協力いただいています。新聞データについては、読売新聞、毎日新聞、産経新聞、共同通信社にご協力いただいています。

■本件に関するお問い合わせ先
独立行政法人国立国語研究所 研究開発部門 言語資源グループ
グループ長 前川喜久雄
電話：042-540-4515
メール：kikuo@kokken.go.jp