Seeds

キーワード:コーパス、頻度、機械可読、正規表現、教材開発

テキスト処理による計量的な言語分析

Masatsugu Ono

ひと文化系領域
言語科学・国際交流ユニット

小野(おの) 真嗣(まさつぐ) 准教授

Phone:0143-46-5882 Fax:0143-46-5889
E-mail:onomasa@mmm.muroran-it.ac.jp
URL http://www.mmm.muroran-it.ac.jp/~16999408/

大量の電子テキストから言語を計量的に分析

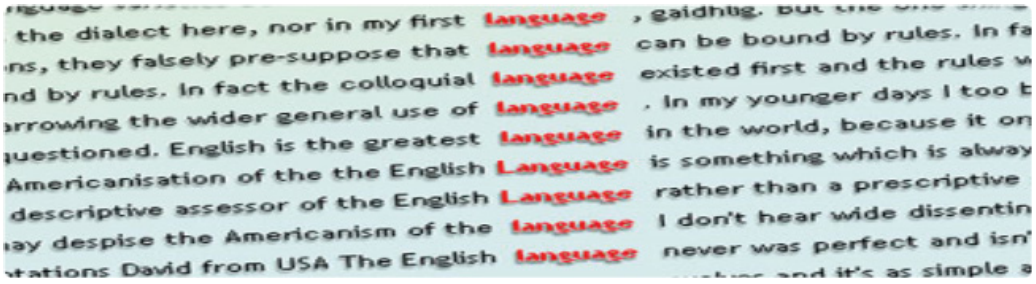
研究の目的

新聞や雑誌など毎日何気なく私たちは文字を目にしながら言語に触れている。それらを言語情報として大量に集め、コンピュータ処理で様々に加工することによって、語の使われる頻度や使われる状況など、辞書には記述されていない内容が近年になって徐々に解明されてきている。これらのデータを分析し、外国語教育における言語材料や教材開発への応用可能性を探究する。

研究の概要

目的に合った語彙習得のために…

コンピュータ上のコーパスを用いた言語研究は歴史的にはまだ浅いものの、これまでに様々な分析手法やツールが開発されており、今後も新たに開発されていくことは間違いない。しかしながら、分析のための基本的な考え方はほぼ同じである。大量の言語資料から目的の表現を抽出し、その表現の共起関係や頻度及び使用域といった言語特性を観察し統計的に処理していくことにある。正規表現を用いたプログラミングはコーパス分析に大変役立ち、アイディア次第で既存有償ツールに縛られない、柔軟な分析が可能となる。



Seeds テキスト処理による計量的な言語分析

研究(開発)のアピールポイント

◆研究の新規性、独自性

これまでの人間の感性や直観に基づく言語記述から、各種言語資料の集積によるデータ分析に基づく記述に変化しており、語法に関してエビデンスの提示といった言語の可視化が実現できるようになる。

◆研究に関連した特許の出願、登録状況

なし

◆従来研究（技術）と比べての優位性

コンピュータの性能向上により、言語資料の大規模化、処理速度の高速化が進んでおり、PerlやPython等のプログラミング言語による正規表現を駆使した分析手法が求められている。

研究(開発)のビジョン、ステージ

◆適応分野

語学教育(英語・日本語)

◆研究のステージ

(基礎研究) 応用段階

◆製品化、事業化のイメージ

目的別語彙集(例:受験英語単語帳、語法辞典等)

企業等へのご提案、メッセージ

◆研究（開発）に関連して、あるいはそれ以外に関われる業務

教科書、付属問題集の他、辞書、文法書といった書籍編纂の相談など。

◆利用可能な設備、装置など

英語・日本語の各種コーパスが利用できます。

◆教員からのメッセージ

今後もコーパスを基にした言語研究の分野は益々発展を続け、コンピュータで処理できる言語資料も日々刻々と量的にまた質的にも精錬されていくことでしょう。同時に新たな分析手法も求められ、学際的に英知を結集した共同研究が望まれています。

小野 真嗣