2ZH-5

自然言語研究，学習，教育支援を目的としたコーパス解析システムの考察と提案

三浦 宏太† 坂本 泰伸‡ 藤野 玄大†

東北学院大学大学院 人間情報学研究科† 東北学院大学 教養学部 情報科学科‡

1. 研究背景

大量の英文書が収録されたコーパスが，主に自然言語研究で利用されている．従来，コーパスに収録される英文書は，紙媒体が中心であり，自然言語研究者は，語の用法やコロケーションの研究[1]を手作業で行っていた．しかし，手作業による語彙の計測は，研究者への作業負担が大きく，その結果を正確な数値として議論に用いることができなかった．

情報技術の発展に伴い，コーパスの電子化が進んだ．これにより，コーパスには，英文書に加え，単語の品詞情報などの付加的な情報である属性情報が保持されるようになった．電子化されたコーパスは，コンピュータによる正確な解析が可能となったことで，定量的な議論が可能となり，自然言語研究が発展した．これにより，これまで主に自然言語研究で用いられていたコーパスが，外国語学習や外国語教育といった分野へ応用されるようになった[2]．

本稿では，コーパスを解析するシステム（コーパス解析システム）に対して考察を行う．さらに，その考察を踏まえ，リレーショナルデータベースである PostgreSQL(RDBMS)を用いた新しいコーパス解析システムとして，自然言語研究や学習，及び教育の支援を目的としたシステム(支援システム)の提案を行う．

2. 研究目的

コーパスに対する解析は，主に英文書や属性情報に対して行われる．研究者は，研究目的に応じて独自のフォーマットでコーパスを構築し，コーパス解析システムもそのフォーマットに従って開発される．そのため，コーパス解析システムは，フォーマットが異なる属性情報に対して解析を行うことが難しい．コーパスを用いた研究では，他者のコーパスは貴重なデータソースとなり，研究者間での共有が求められる．また，解析対象となる属性情報は，単一の属性情報が主に単語に対して付与される．しかし，属性情報を単語だけではなく，文や熟語といった英文書の構成要素に対して付与でき，さらに研究者による任意の複数の属性情報を扱うことができれば，コーパスを学習支援や教育支援といった分野に応用することが可能となる．

3. 現状のコーパス解析システムの考察

コーパス解析システムは，解析対象となるコーパスで用いられている単一の属性情報のみ解析を行うことができる．そのため，異なる属性情報に対して解析を行うことができず，コーパスの共有が難しい．しかし，狭い分野を対象とした研究では，解析対象となる英文書や属性情報といったデータソースの収集が難しく，他の研究者が用いているコーパスの共有が望まれる．

また，学習支援や教育支援を考えると，英文書の構成要素に対する属性情報の付与が求められる．しかし，現状のコーパス解析システムでは，主に単語に対して属性情報が付与され，英文書の構成要素である文や熟語といった対象に付与することが難しい．さらに，付与する属性情報は，単一の属性情報しか扱えず，研究目的に応じて，複数の属性情報を使い分けて付与することができるコーパス解析システムは少ない．このような問題を解決するには，DataBase Management System(DBMS)による英文書と属性情報の柔軟な管理が必要となる．

4. 新しいコーパス解析システムの提案

新しいコーパス解析システムとして，RDBMS を用いた支援システム(図 1)を提案する．支援システムでは，コーパスに収録された英文書と属性情報を蓄積するフォーマットを統一することで，研究者間でのコーパスの共有化を行う．また，

Consideration and suggestion of the corpus analysis system for the purpose of the NLP, learning, education support.

†Graduate school of Human Informatics, Tohoku Gakuin University

‡Department of Faculty of Liberal Arts, Tohoku Gakuin University

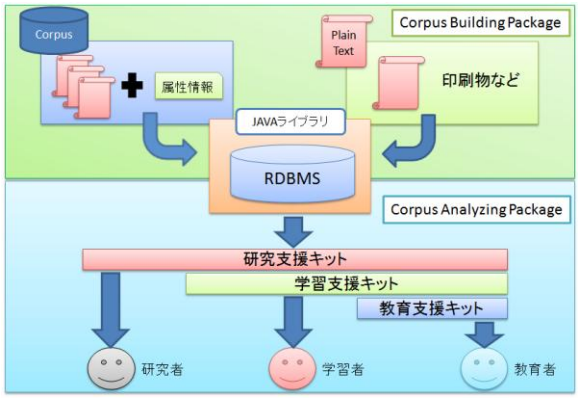

図 1　支援システムの全体像

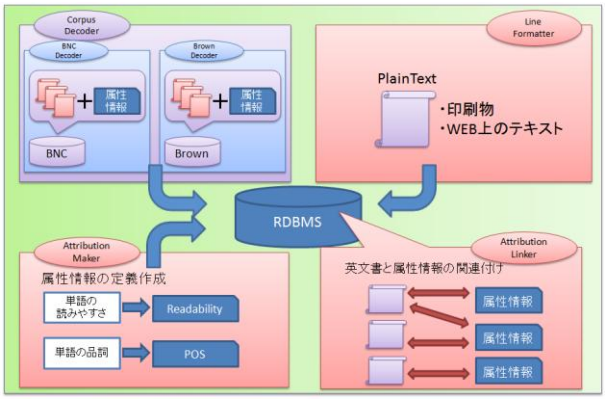

図 2 Corpus Building Package

本来，階層構造を持つ英文書の管理には，階層型データベースが適している．しかし，学習支援や教育支援といったコーパスの応用を考えると，英文書の構成要素に対して，任意の複数の属性情報の付与が求められる．そこで，英文書の構成要素や複数の属性情報の管理には，リレーションを用いることで柔軟に管理することができる RDBMS を採用する．また，RDBMS を用いることで，英文書に対して複数の属性情報を関連付けることが可能となる．支援システムは，この RDBMS を中心に，データソースを蓄積する部分である Corpus Building Package(CBP)と，RDBMS に対して解析を行う部分である Corpus Analyzing Package (CAP) によって構成される．

4.1 Corpus Building Package

CBP(図 2)は，支援システムの中心である RDBMS に英文書と属性情報の蓄積を行う．コーパス解析では，他者のコーパスに対して解析を行えることが求められる．そのため，支援システムでは，既存のコーパスを蓄積するアプリケーションとして Corpus Decoder を提供する．また，既存するコーパスに収録された英文書以外にも，利用者は自由に解析を行う英文書のコンテンツを選択できることが求められる．そこで，紙媒体の書籍や WEB 上の英文テキストといった PlainText を RDBMS へ蓄積するアプリケーションとして Line Formatter(LF)を提供する．

Corpus Decoder や LF によって蓄積された英文書に着目すると，既存コーパスには他者によって属性情報が既に付与されており，PlainText には属性情報が付与されていない．利用者の要求を考慮すると，他者が定義した属性情報ではなく，自身が必要とする属性情報を新たに定義し，英文書に対して再付与できることが望まれる．また，PlainText のような属性情報を持たない英文書には，属性情報を付与する枠組みが必要となる．そこで，支援システムでは，利用者による自由な属性情報の定義を可能とする Attribution Maker と，RDBMS に蓄積された英文書と属性情報の対応付けを行う Attribution Linker を提供する．

4.2 Corpus Analyzing Package

CAP は，RDBMS に収録された英文書や属性情報に対して，語の共起関係や品詞の連鎖関係などの検索を行う．このような検索結果に対して，二次的な加工を行うアプリケーションを用意することで，中学英語で用いる熟語を含む英文抽出などの学習支援を行うことができる．さらに，テスト問題の配布や収集といった枠組みを提供することで，教育支援への利用も可能となる．

5. まとめ

本稿では，現状のコーパス解析システムへの考察を行い，問題点を上げた．さらに，考察から得た問題点を踏まえ，新しいコーパス解析システムとして，RDBMS を用いた支援システムについて論じた．支援システムにより，コーパスの共有が可能となり，利用者による任意の複数の属性情報を英文書の構成要素に対して付与できると考える．

謝辞

本研究は，平成 22 年度科学研究費補助金基盤研究(C)(22500891)による助成を受け進められている．ここに深く謝意を表する．

参考文献

[1]: 後藤一章　“英語コーパス解析に基づく類似名詞と共通共起同士の検出：コロケーション時点への活用”

[2]: 佐野洋，中村隆宏　“BNC を利用した英語教材作成とその提供 Web サイトの開発”　2004-CE-77 (8)　2004/11/20　社団法人　情報処理学会　研究報告