E-010

コーパスを利用した研究・学習・教育支援システムの開発に向けた RDBMS の評価
Evaluation of an RDBMS for a corpus-based language research, teaching and learning supportive system

藤野 玄大† 三浦 宏太† 坂本 泰伸‡
Genta Fujino Kouta Miura Yasunobu Sakamoto

1. 研究背景

近年，コーパスの解析結果を利用し，自然言語研究や外国語学習等に応用する活動が活発に進められている．自然言語研究者は，コーパスに蓄積された英単語の並びや付与された属性情報に着目して解析を行い，その結果を自身の研究に利用している[1,2]．これらの解析に使用される属性情報は，コーパスを利用する目的や利用者ごとに異なることが一般的である．このように，利用者が必要とする任意の属性情報をコーパスに対し付与することや，既に存在するコーパスの属性情報を利用して新たな属性情報を再生産することは，更なる応用に利用できると考えられている．

本稿では，リレーショナルデータベースである PostgreSQL（RDBMS）を用いた英文書の管理と，英文や単語に付随する複数の属性情報を柔軟に管理するテーブル設計について述べる．

2. 研究目的

これまでに，我々はコーパスを利用した研究や学習を支援するシステムの開発を進めてきた[3]．この支援システムでは，大量の英文書と属性情報を RDBMS で管理する．コーパス中の英文や単語，属性情報に対して SQL による検索を行い，その結果に対して二次加工を施すアプリケーションを用いることで，研究だけではなく，学習や教育活動の支援をすることを目的としている．このような支援を行うには，大量の英文書の管理や柔軟な属性情報の管理が必要となる．

解析対象となる英文書は，文の集まりで構成されており，文は単語の連鎖によって構成される．すなわち，階層構造を持っている．一般的にこのような階層構造を管理する場合には，XML データベースや階層型データベースが利用される．しかしながら，支援システムでは英文書が持つ階層構造を管理するだけではなく，英文書を構成する個々

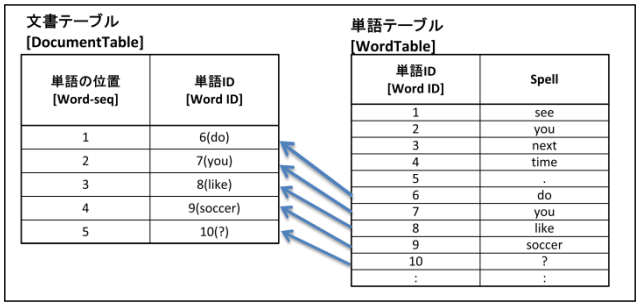

図 1 英文書を構成する単語の管理

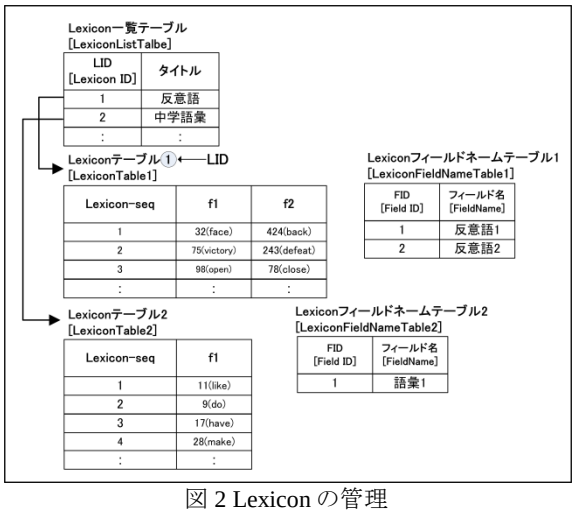

図 2 Lexicon の管理

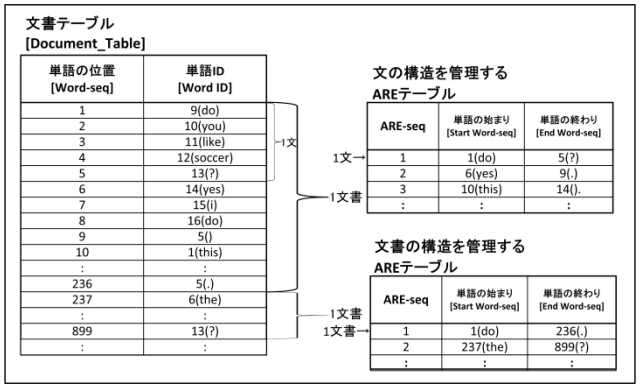

図 3 ARE の管理

の要素に対して複数の属性情報を関連付ける必要がある．そこで，英文書の構成要素と複数の属性情報をリレーションで容易に管理することができる RDBMS を採用した．また，熟語や語彙，レマといった単語の集合を内部で管理することで，学習や教育活動における更なる支援も視野にいれている．

3. RDBMS で管理する英文書の構成要素

3.1 単語と Lexicon の管理

解析の対象となる英文書は，最小構成要素である単語の連鎖によって構成される．支援システムでは，個々の単語に対し，重複しない一意の値（単語 ID）を割り振ることで，各単語を識別する（図 1）．また，語彙やレマ，反意語や同義語といった何らかの明示的な関係にある単語を，Lexicon として個々の集合毎にテーブル（Lexicon Table）で管理する（図 2）．これらの Lexicon Table には，重複しな

† 東北学院大学大学院人間情報学研究科 Graduate school of Human Informatics, Tohoku Gakuin University
‡ 東北学院大学教養学部 Department of Faculty of Liberal Arts, Tohoku Gakuin University

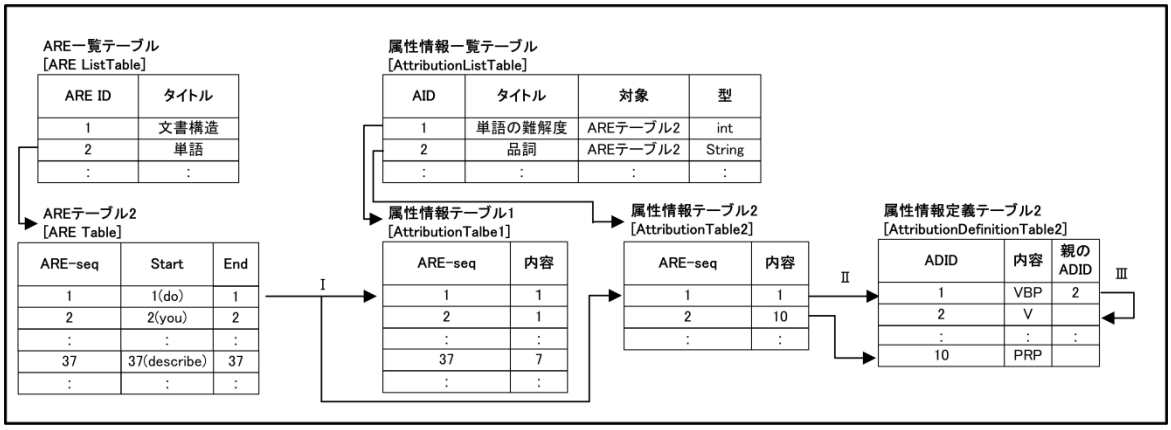

図 4 ARE に対する属性情報の付与

い値（LID: Lexicon ID）がテーブル名に用いられ，この値が Lexicon 全体を管理するテーブルに登録されている．また，個々の Lexicon テーブルの構成情報（Lexicon テーブルの名前やフィールド数）を保持するテーブルも存在し，同じく LID で関連付けている．

これらのテーブルを用いて Lexicon を管理することにより，ある語彙を含む英文の抽出や，反意語や同義語，レマを対象とした全文検索も行うことが可能となっている．

3.2 英文書の構造管理

英文書に対する検索では，文書中の文や単語，単語の連鎖（共起）といったような，様々な英文書中の構成要素やそれに付随する属性情報に着目した処理が必要となる．支援システムでは，まず，複数の英文書を唯一のテーブルに追記していく形式で英文書を管理する[3]．英文書の内容は，英単語の連鎖によって構成され，英文書中の行や段落といった構成要素は，我々が ARE（an Allocation of Remarked Elements）と呼ぶ概念に基づいて管理する（図 3）．ARE は，各構造の単語の始まり位置と終わりの位置を記録している．また，これらの記録される位置情報に対して，重複しない値（ARE-seq）を割り振ることで個々の位置情報を識別する．ARE の集合は，利用目的毎に一つのテーブルに収められ，重複しない一意の値（ARE ID）によって管理される．

支援システムでは，この ARE を共通の外部キーとして利用しながら，内部で利用されている構成要素（文書，文）と属性情報を関連付けている．利用者の必要性に応じて，段落や章といったその他の構造表現や，文書中に出現した熟語や連語などの位置なども記録することが可能である．また，検索で得られた結果も ARE 形式で記録する．この ARE の考えにより，利用者の定義した幅の広い属性情報に対する検索や，検索結果に対する再検索も実現可能となっている．

4. RDBMS を用いた属性情報付与

4.1 付与可能な属性情報の型

英文書を構成する単語に対して付与される属性情報には，品詞情報等の文字列型や，単語の難解度のスコア等の数値情報を付与する場合がある．我々の支援システムは，英文書を構成する要素に対し，数値型と文字列型の 2 種類の属性を付与する機能を提供する．

4.2 属性情報付与の方法

解析に利用される属性情報は，利用者毎に定義することができ，ARE が表す英文書の構成要素に対して，リレーションを用いることで付与される．個々の属性情報は，重複しない値(AID: Attribution ID)を割り振ることで識別する．

付与する属性情報が数値型の場合には，ARE-seq に対して数値情報を直接関連付けることで管理する（図 4 - Ⅰ）．また，属性情報が文字列型の場合には，属性情報を表す文字列に対し，重複しない値（ADID: Attribution Definition ID）を割り振ることで個々の文字列を管理し，その値と ARE-seq を関連付けることで文字列型を付与する（図 4-Ⅱ）．また，文字列の場合には，親子関係のような階層構造も管理している（図 4-Ⅲ）．

5. まとめ

解析対象となる英文書の構造の単語の始まりの位置と終わりの位置を管理する ARE を用いることで，英文書の持つ文や単語等の構造を管理することが可能となった．また，この ARE に対し，複数の属性情報を文字列と数値で付与することが可能となるテーブル設計を提案した．このようなテーブル設計を用いて，属性情報を付与する対象を従来の単語だけではなく，ARE によって管理される文書や文に対し付与することで，様々な支援が実現可能になると考えられる．これらのテーブルの性能評価では，The British National Corpus の蓄積を行い，評価を行った．

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．また，この研究に関する多くの議論を，東北大学の岡田毅氏と立命館大学の田中省作氏を始めとする多くの方と行い，有益なコメントをいただいた．ここに深く謝意を表する．

参考文献

[1] 朱京波, 片上大輔, 新田克己, “ウェブベース英単語学習支援システムの提案”, 電子情報通信学会 ET2005-76, (2006-01)

[2] 内堀朝子, 中條清美, “コーパスを用いた文法・語彙指導―基本的な名詞句構造に関する暗示的および明示的指導の組み合わせ―”, 日本大学生産工学部研究報告 B,(2010)

[3] 藤野玄大,” コーパスに基づく研究・教育・学習支援システムで利用する英文書管理の提案と評価”,平成 22 年度第 5 回情報処理学会東北支部研究会(2011)