3Q－3

単語の意味論的進化情報を付加した英語シソーラスデータベースシステムの概念設計 *

山崎達也† 池辺八洲彦‡ 藤代一成§
筑波大学 筑波大学 お茶の水女子大学

1 はじめに

英語シソーラスは英語句を意味によって分類配列し，それらの各語句について同義語，反意語などを記述した一種の類語辞書であり，表現したい意味内容を最も的確に表す語句を選び出すことに主に用いられる．

現在までに出版されている多数の英語シソーラスには，同一の語義を表すとしてグループ化されている各語句のニュアンスの違いを理解させるような情報を充分には扱ってないという内容上の問題点が存在する．

本研究では，この内容上の問題点を克服するための有力な補助情報の一つとして，印欧語根までさかのぼる単語の意味論的進化情報を付加した英語シソーラスデータベースシステムを提案する (印欧語根の可利用性については文献[1] を参照して頂きたい)．そのためには英語シソーラスを構成する各オブジェクトの概念スキーマを設計し，さらにその概念スキーマによるデータベースシステムが矛盾なく動作することを確認するため状態モデルを設計する必要がある．本論文ではその一例を示す．

2 英語シソーラス

英語シソーラスとは，英語句を意味によって分類配列し，それらの各語句について同義語，反意語などを記述した一種の類語辞典である．シソーラス (thesaurus) という言葉は，ギリシャ語で宝庫という意味の (*thēsaurós*) に由来する．英語シソーラスは文章の作成において表現したい意味内容を的確に表す英語句を選び出すために使われる．

なおここで言うシソーラスは Roget's 等に代表される英語句に関するシソーラスであり，情報科学の分野で言われる広い意味でのシソーラスとは違うことに注意して欲しい．

本研究では，数多くあるシソーラスの中で最も定評のある Roget's のシリーズの一つである Roget's II Internationl Thesaurus Expanded Edition(以下 Roget's II と略す) について解析を行うことから始める．

通常英語シソーラスは，各語句の意味記述をもたないため英英辞典等と併用されるが，Roget's II は，各語句の間の同義，反意などの関係を示しながら，意味記述も併せて示しているので従来のものより容易に利用可能となっている．

*Conceptual modeling of English thesaurus database system with evolving semantics of words

†Tatsuya Yamazaki(University of Tsukuba)
‡Yasuhiko Ikebe(University of Tsukuba)
§Issei Fujishiro(Ochanomizu University)

3 Roget's II の問題点

Roget's II は，使いやすさやシソーラス本来の目的の達成度を考えると多くの問題点を残している．その問題点は構造上の問題点と内容上の問題点の二つに分けることができる．構造上の問題点については既に文献[3] で論じられているので，ここでは内容上の問題点について考える．

内容上の問題点として最も顕著なものを以下に示す:

- 同義語としていくつかの語句が一つのグループにまとめられているが，各々の語句は完全に同じ語義を持つわけではない．それにもかかわらず同一グループ中の語句のニュアンスの違いをユーザに理解させるような情報を扱っていない．

この具体例を以下に挙げる:

- "ease"の一番目の語義である"Freedom from constraint,embarrassment,or awkwardness"の意味を持つとされる，"informality"，"naturalness"，"poise" などは同一グループにまとめられている．しかしそれらの語義の違いを示す情報は示されておらず，ユーザはそれらの単語を使用する局面に応じて的確に使い分けることができない．

この問題を解決するために，それぞれの語の持つ意味論的進化情報を利用することを考える．

4 単語の意味論的進化情報の利用

前節で挙げられた Roget's II の内容上の問題である，"同義語として一つのグループにまとめられた各語句のニュアンスの違いをユーザに理解させるための情報を持たない"，という点を解消するために，その語句が，どのような意味の語句に由来してきたかを示す単語の意味論的進化情報を扱う．本研究では，語源情報として文献[1] などで既にその意義が認められている印欧語根までさかのぼることにする．これが本研究の重要なポイントである．

印欧語根とは，英語の起源を，有史前までさかのぼっていくと最終的にたどりつくということが比較言語学によって確立されている推定先祖語である印欧原語の語根である．C.Watkins らの研究によって 1414 個の印欧語根が再構成されており，その意味的説明，由来する現代英単語，ならびにその発展経過などが詳しく説明されている．なお現在使われている全ての語句について，そのもととなった印欧語根が明らかにされているわけではない．しかし，ある学習用参考書に使われた単語に関する調査ではその 87%について印欧語根が明らかになっている (文献[4])．このことから印欧語根を英語シソーラスの扱う情報とすることは充分実用性のあることだと思われる．

先ほど例を挙げた"case"と"informality"の語源情報を調べてみる．"ease"の印欧語根は"ye-"である．これは投げるという意味であり，投げる→近くに横たわる→便利→快適→気軽さとその意味が変化してきた．

"informality"の印欧語根は"mer-bh-"である．これは光ることという意味であり，光ること→形→形にこだわる→形式的と意味が変化し，"formal"という語になった．"formal"に否定の"in"がついて形式的でないことを表す"informal"になり，結果的に気軽さ，気楽さ，という意味を持つようになった．

結果として"ease"は自分が快適であるという意味での気楽さ，"informality"は形式ばらないという意味での気楽さを示していると考えられる．このようにして語源情報からニュアンスの違いを知ることができる．

5 システムの概念設計

意味論的進化情報を扱うことによって同義語間のニュアンスの違いが明らかになるにもかかわらず，Roget's IIやその他のシソーラスで，それが扱われてこなかった理由の一つとして，それらが紙を媒体として構成されていたため，情報量の制限があったことが考えられる．そこで媒体を紙ではなく磁気ディスクとしデータベースとして構成する．これによって多量の情報を扱うことができ，意味論的進化情報も含めることが可能となる．

データベース化の方針として，このデータベースで扱う情報としては，Roget's IIで解説されている見出し語に関する情報と見出し語間の関係についての情報，並びに文献[1]に収録されている語句の進化情報を扱う．全ての見出し語を主見出し語として概念スキーマ上で表現する．見出し語が同一の意味記述を持つ場合，それらの見出し語は同義であるとする．各見出し語間に存在する準同義，準反意，反意の関係を概念スキーマ上で表現する．従来のシソーラスには，存在しなかった単語の進化情報を扱えるようにする．このような方針で設計したデータベースのERスキーマを図1に示す．さらにこの概念スキーマから，人間とデータベースシステムの相互関係を示す状態モデルを設計する．状態モデル作成の詳細は省略するが結果として図2に示す状態モデルが得られる．

6 まとめと今後の展望

英語シソーラスにおける同義語理解のため単語の意味論的進化情報を扱うことを提案した．単語の意味論的進化情報を扱う英語シソーラスのERモデルを設計し，それをもとに人間とデータベースシステムの相互関係を示す状態モデルを設計した．

今後は設計したシソーラスデータベースのERモデルの改良を行い，ERモデルと状態モデルをもとにデータベースの実装，英語シソーラスに適したユーザインタフェイスの設計実装を行う予定である．

参考文献

[1] C.Watkins: Indo-European and the Indo-Europeans, in *The American Heritage Dictionary, 1st college edition*, Houghton Mifflin,1969.

[2] Anne H. Soukhanov et al.(eds.): *Roget's II The New Thesaurus Expanded Edition*, Houghton Mifflin,1988.

[3] 竹川弘志: 計算機シソーラスシステムの概念設計及びその試作, 筑波大学卒業論文,1989.

[4] 佐藤和彦: 英単語学習のための辞書データベースの構築, 筑波大学大学院修士論文,1992.

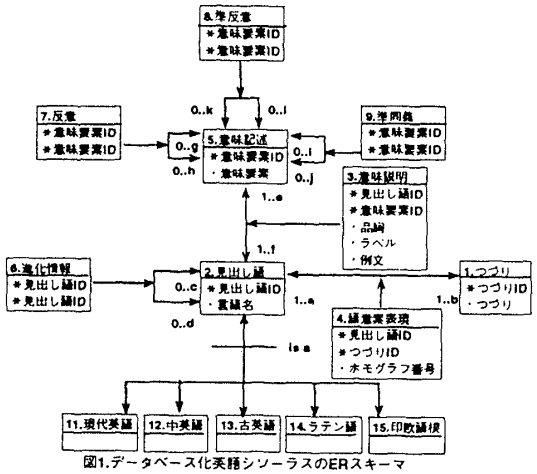

図1.データベース化英語シソーラスのERスキーマ

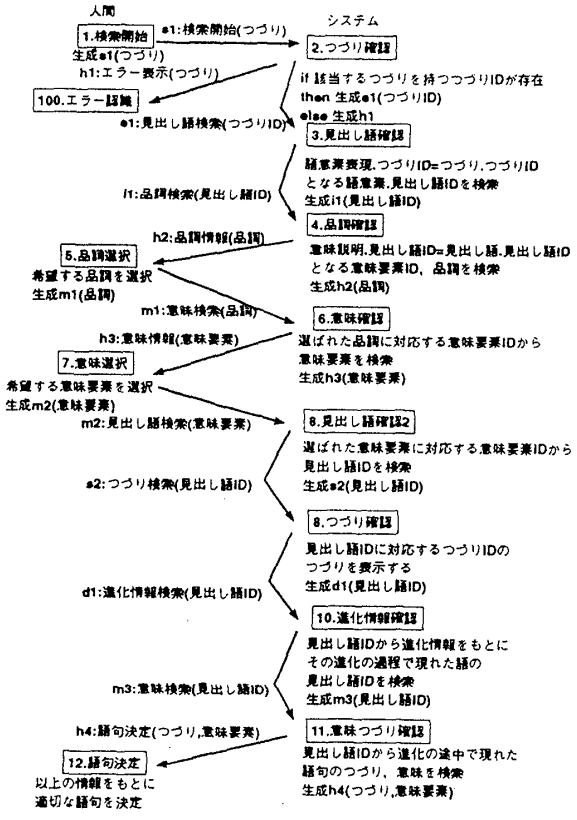

図2.人間とシステムの相互関係を示す状態モデル