計算機シソーラスシステムの概念スキーマ設計

7F－7

藤代 一成・池辺 八洲彦・竹川弘志+
筑波大学電子・情報工学系 +筑波大学第三学群情報学類

1. 背景と目的

自然言語辞書(以下,単に「辞書」と呼ぶ)は,語学を含むあらゆる言語活動に必要不可欠な道具であるにもかかわらず,現状の大半が「紙の上の印刷」によって,リニアテキスト化されている.利用者に非効率な逐次検索を強いたり,①語彙の意味の相互関係を表現しない,②構造上の改善きす欠陥をもつ.これらの欠陥を加えて,①辞書は本来,間断経こい言語活動の影響を受け,緩慢な年変化を続ける自然言語の姿を正確に映出すべきものであるから[1],特徴的なデータ更新操作を含む履歴情報管理が必要になり,②さらに用途に応じ群たかけるなら,辞書はまさにデータベースの興味深い対象となる.

本稿の目的は,比較的構造の明解な辞書の例としてシソーラス(thesaurus)と呼ばれる類語辞書をとり上げ,その計算機化のための概念スキーマ設計を行うことである.そのためまず次節で,シソーラスのデータ構造の実際を検討した後,3節で,一種の拡張E－Rモデルを用いて同シソーラスの概念スキーマを設計する.4節では今後の拡張方針をまとめることにする.

2. シソーラスの実際

シソーラスとしては,紙上でのシノニム(synonym:同義語)の引き易さがポイントのROJET'sII[2]を選んだ.同シソーラスでは,シノニムを規定する単語の意味(meaning)を,

指示的意味(denotation)＋
含意(connotation)＋
適用範囲(range of applicability)

から構成されると定義した上で,ほぼ同様の指示的意味をもつ単語までをシノニムとして認めている.したがって,同シソーラスのシノニムは通常,引用文中での完全な互換性をもつわけではない.

図1は,同シソーラスの構造を端的に示すための紙面の抜粋である(記述を短くするために日本語を用いた).

登録語(entry)は,見出し語(lemma)に品詞を組み合わせている.これ以外にホモグラフ(homograph;同形意義語)のための個別番号を伴うこともある.登録語は,主(main)登録語と副(sub)登録語に分かれる.図1では,enoughとsufficientの名詞が各々の形容詞の副登録語(インデントされている)となっている.副登録語は,歴史的な派生関係を表現する他,2語動詞等を基本動詞に関連づける(例:fade in→fade).各登録語は,複数の意味要素(meaning element)をもつことができ(順番は不特定),それぞれにシノニム相互参照リストが付く.重要なイディオム(idiom)は,未登録ながらシノニムリストに含める場合がある.図2に図1のシノニムの相互参照関係が無向グラフで表現できることを示す.

adequate 形
1.中庸な. 1. ACCEPTABLE.
2.過不足ない. 2. SUFFICIENT 形.

enough 形
過不足ない. SUFFICIENT.
enough 名
適量: Give me enough to eat.
Syns: adequancy, sufficiency, sufficient.

sufficient 形
1. 過不足ない: We had just sufficient fuel to get us across the desert.
1. **Syns:** adequate, comfortable (Informal), competent, decent, enough, satisfactory, sufficing.
2. 中庸な. 2. ACCEPTABLE.
sufficient 名
適量. ENOUGH 名.

図1 シソーラス紙面の例

グラフ理論的に見た同シソーラスの大きな特徴は,各連結成分がシノニムの集合を表し,同一連結成分中,自分以外のすべての頂点と隣接するような頂点が唯一つ存在することである.この頂点は,1次(main)登録語,それ以外は2次(secondary)登録語と呼ばれている(図2のsufficient:形とenough:名が1次登録語である).2次登録語どうしは,1次登録語を介してのみ相互参照可能である(enough:形は1次登録語sufficient:形を介してのみadequate:形を参照できる).実際の紙面では,1次登録語の意味要素だけに例文が含まれ,Syns:の後にフルシノニムリストがくる.また2次登録語は,1次登録語だけをシノニムとしてもち,それは大文字で示されている.

さらに同シソーラスは,利用者の便宜を図るために,登録語のさまざまな

The Conceptual Schema Design for a Computer-Based Thesaurus System
Issei FUJISHIRO, Yasuhiko IKEBE, and Hiroshi TAKEGAWA
University of Tsukuba

特性をラベル(label)で表現している.ラベルには,歴史的,用法,方言,語源,分野,頻度の6種類が用意されている.図1では,enoughのシノニムリスト内のcomfortable(informal)が,「過不足ない」の意味で使用される場合に,口語的な用法であることを表している.この限定は局所的であり,見出し語としてのcomfortableには付加されていないことに注意されたい.

sufficient:形

adequate comfortable … enough:形

enough:名

adequancy sufficiency sufficient:名.

図2 シノニム相互参照グラフ

3. シソーラスの概念スキーマ

2項で述べたシソーラスのデータ構造は,紙面のままでは以下のような欠点をもっている:

①(データの重複)意味要素が,関連するシノニムの個数分だけ重複して紙面に現れる.逐次検索時の便宜は図れるにしてきても,紙面の増大は免れない.

②(非効率な検索)利用者が2次登録語を想起した場合には,シノニム登録語を参照する必要があるため,1次登録語を検索する必要があるため,再度,シソーラスの方式から検索するこれらは,シソーラスのデータ構造をにの基づいたデータ構造全体を離れ,検索対象を1次登録語に限定することで,検索の操作を簡単化する.

以下では,その目的のために,2項で述べたシソーラスの概念スキーマの結果に基づいて,同じ構造分析の概念スキーマを設計する.シソーラスの一人が提案したSGDM(Semantic Graph Data Model)[3]と呼ばれる,一種の拡張ER概念モデルを用いることにする.このモデルは,汎化を始めとする選択抽象化機構をもち,関連結合度やシソーラス性も陽に表現でき,後のリレーショナル・モデル等を用いた実装容易に変換できるといった特長をもっている.

図3に,シソーラス概念スキーマを表現するSGDMダイアグラムを示す.ただし,属性等は説明に必要なものだけを示してある.各実体型には,独自の識別子(surrogate)が対応しているので,属性はすべて記述子の役割をしているものとする.

①ホモグラフ番号の有無により,単語/連語を区別したが,それは汎化によって登録語にまとめられる.単語には1次/2次登録語の区別を付加した.

②従属関係は,主/従登録語の関係を記述する.

③連関は,多対多の関係を表現する.2次登録語は,1次登録語の属性をもとに,「同義」「多対1」「1次/2次」に区別した.

④見出し語は,連関番号を用いて,主/従関係を表現する.

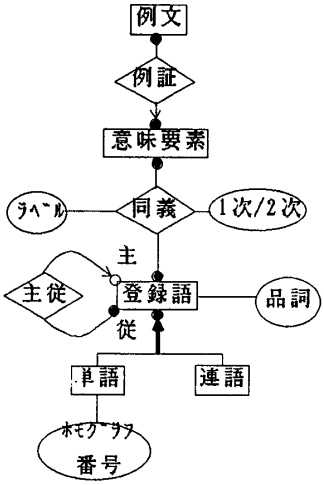

図3 シソーラス概念スキーマ

4. シソーラスの拡張性

3項で設計したシソーラスのキーマ概念スキーマを,任意の登録語に対する意味要素の拡張操作により,例文や連語を登録語の階層構造に定義することが可能である.

さらに,次のような拡張方法を考えられる:

①例文や連語を登録語の再帰的集約(recursive aggregation)として表現すれば,例文-連語-単語の言語要素の階層が現れ,例文レベルのきめ細かい検索が可能になる.

②単語の難易度を何らかの方法で定義すれば,利用者の習熟度に応じたサブシソーラスを適応的に構成することができる.

現在,これらを含めた概念スキーマの詳述化と,対応するパイロットデータベースの構築を検討中である.

参考文献

[1] 池辺 八洲彦, 藤代 一成: 英語学習/教育のための「英語トピックス辞書」構築について, 本大会講演論文集(1990).

[2] ROGET's II: The New Thesaurus, Houghton Mifflin (1980).

[3] Fujishiro, I.: A Logical Design Methodology for Link-Oriented Databases Using Data Model Mapping, Doctoral Dissertation to Faculty of Science, the University of Tokyo (1988).