

## シソーラス辞書におけるデータの管理

6 E-5

内田真由美\* 安高みわ\* 内海正樹\* 森谷精徳\*\*

\*株式会社 東芝 \*\*東芝アドバンストシステム株式会社

## 1.はじめに

全文検索においてはユーザが思い思いの言葉で検索を行なうことができるため、同じ意味を持つ単語や関連語などが文書内に記述されていてもその文書が得られない場合がある。

このような検索漏れをなくすためにシソーラス辞書を用いるシステムが多くなっており、シソーラス辞書を用いた検索は、全文検索の機能の一部として重要な役割を果たしている。

シソーラス辞書が重要な位置を占めるにつれ、シソーラス辞書の扱うデータ量も増え、シソーラス辞書そのもののデータ構造に対する改良も必要になってきた。

今回、シソーラス辞書のデータに着目し、データ構造に対する検討を行なった。本稿では、作成したシソーラス辞書のデータ管理方法について述べる。

## 2.これまでのシソーラス辞書

## (1)従来のシソーラス辞書構造

従来のシソーラス辞書形式を図1に示す。

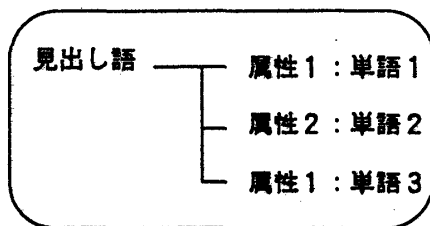


図1 これまでのシソーラス辞書

シソーラス辞書の一般的な形式は、"見出し語"とそれに対する"単語"から構成され、その"見出し語"と"単語"を"属性"によって定義づけしてい

る。"属性"とは、"同義語"、"上位語"、"下位語"など見出し語と単語との関係のことを言う。

## (2)問題点

従来のシソーラス辞書の具体例を図2に示す。

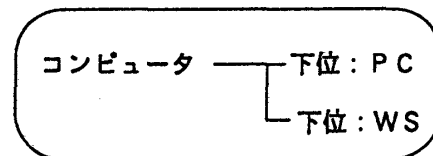


図2 これまでのシソーラス辞書の具体例

図2ではコンピュータの関連語として"PC"と"WS"という語が定義されている。さらに"電子計算機"という語を"コンピュータ"の同義語として追加した例を図3に示す。

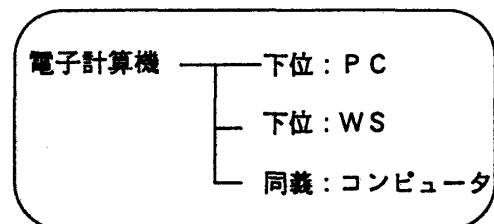
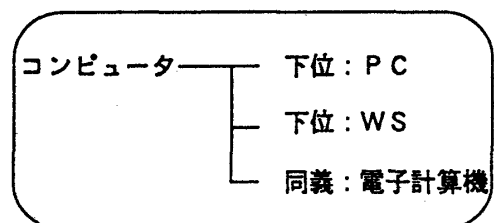


図3 辞書への同義語の追加

図3の見出し語"コンピュータ"と"電子計算機"の辞書の内容はほぼ同じである。にもかかわらず、別々の見出し語であるため、辞書内にはダブった内容が登録される。

また、"コンピュータ"や"電子計算機"の内容に"オフコン"を追加しようとする、整合性をとるためにも"電子計算機"という見出し語に対しても"オフコン"という語を追加しなければならない。

Data Management in a Thesaurus

Mayumi UCHIDA\*, Miwa ATAKA\*, Masaki UTUMI\*, Kiyonori MORIYA\*\*

\*TOSHIBA Corporation

\*\*Toshiba Advanced System Corporation

この方式の問題点は、

- ・辞書内容がダブっており、無駄が多い
- ・同義語が別々の見出し語として登録されているため、辞書の整合性のチェックが困難

というものが挙げられる。

### 3. シソーラス辞書の改良

#### (1) 同義語の統合

2で述べた問題点の一つである内容のダブリは、同義語ごとに同じ内容のものを定義してしまうためであると考えた。そこで、同義語による定義をまとめるために、同義語を一つの塊にまとめることにした。ここでは、この塊をシノプス（概念）と呼ぶ。

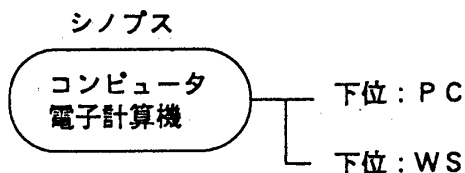


図4 シノプスを用いたシソーラス辞書

これにより、同義語による同じ辞書内容の登録がなくなり、ダブリが少なくなった。また、同義語がひとかたまりで登録されているので、同義語ごとに辞書の整合性がとれるようになった。

#### (2) 属性の定義

さらに、シノプス単語の関連づけに着目した。図4の状態だと、「PC」に下位語などがあった場合、「PC」を含むシノプスを作成しなければならない。それでは、同じ単語が辞書内に何回も現われるため、PCなどのシノプスに関連する語もまたシノプスに入れることにした。これにより、全ての語はシノプス内に格納されることになった。

同義語以外の関連づけは、シノプスとシノプスの関連づけによって定義づけすることにした。

#### (3) 複数のシノプスへの関連づけ

図5においては、一つのシノプスからのみ関連づけが行なわれている。しかし、シノプスには上下関係はないので、どのシノプスから他のシノプスへ関連づけをしてもかまわない。

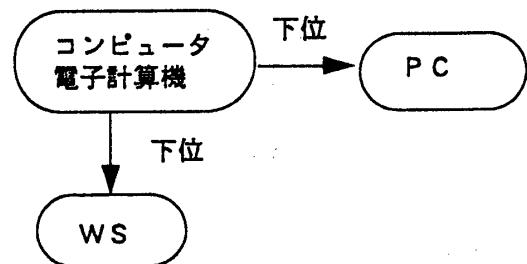


図5 改良したシソーラス辞書形式

これにより、見出し語と単語のダブリもなくなった。また、図5の方式において複数のシノプスに属性を定義できるようにすると図6のようになる。

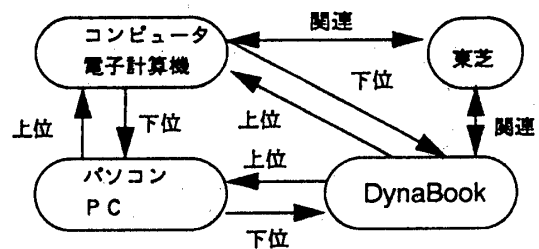


図6 複数の関連づけをしたシソーラス辞書

これにより、シノプス間を結ぶだけで関連を指定できる範囲が広がり、類似語として得られる語の範囲が広がった。

### 4. おわりに

全文検索システムで使用するシソーラス辞書の形式について述べた。今回のシソーラス辞書の形態にしたことにより、

- ・同じ内容の登録がなくなり、無駄が減った
- ・同義語をまとめることにより、辞書の整合性がとれるようになった

という結果になった。またこの結果により、

- ・語のダブリが少なくなり、辞書のサイズが小さくなった
- ・語の登録回数が減り、ユーザカスタマイズしやすくなった

などの利点があった。

シソーラス辞書は全文検索のみならず、多くの分野で有効に使用することができると思われる。