

日本語学習者用電子化辞書の開発に向けて

白井太三, 乾健太郎, 太田武昭, 徳永健伸, 田中穂積

東京工業大学工学部, 大学院情報理工学研究所

{taizo, inui, ota, take, tanaka}@cs.titech.ac.jp

1 はじめに

電子化辞書が普及し始めてからすでに久しい。電子化辞書は、従来の書籍辞書にくらべ、はるかに大規模な記述のなかから高速に情報を検索することができるという利点や、ハイパーテキスト化により多様な情報を多様な関係で組織化することができるという利点を潜在的にもっている。このような特性をもつ電子化辞書は効果的な言語学習を実現する上で重要な役割をはたすことが期待できる。ところが、現在の日本語電子化辞書の多くは、書籍辞書の記述を電子化したものに検索機能や画像・音声データを付加した構造をもっており、データベースの情報量や情報の組織化という点では従来の辞書と大きく違わないのが現状である [10]。電子化の特性をさらに活かした辞書を開発すれば、言語学習の効率がさらに向上すると考えられる。とくに、外国人用日本語辞書については、かな・ローマ字表記や他言語訳を載せている辞書 [12] や用例中心の辞書 [11] などがいくつかの書籍辞書が編纂されているが、電子化は立ち後れている。また、外国人用日本語辞書のデザインについては玉村が論じているが、玉村の議論は書籍辞書を前提としており電子化辞書については触れていない [9]。このような背景にもとづき、本稿では、外国人日本語学習者を対象とする学習用辞書に焦点をあて、電子化の特性を効果的に利用する辞書のデザインについて論じ、現在試作中のプロトタイプについて報告する。

2 電子化辞書の機能とその実現方法

電子化辞書に求められる機能を以下の4つの観点から整理し、それぞれ関する実現方法と問題点について述べる。

- 日本語理解を支援する機能
- 日本語産出を支援する機能
- 語句の定着を支援する機能
- カスタマイズ機能

2.1 日本語理解を支援する機能

日本語理解の支援には、少なくとも節や句を形態素(+構文)解析する機能と節や句の構成要素の意味を表示する機能が必要だと考えられる。

前者の機能があれば、単語単位の辞書引きだけでなく、電子化文書からの自動的な辞書引きが可能になる。日本語のように分かち書きされていない言語では形態素解析機能の有用性はとくに高い。形態素解析については現在の技術でもある程度実用的なシステムの開発が可能であろう。ただし、学習者がキーボード入力する場合には、入力誤りを想定した頑健な辞書引きを実現しなければならない。とくに、促音や長音などは学習者の母語と干渉しやすく、学習者が誤って聞きとることが少なくない [2]。われわれは現在、学習者の認識誤りによる入力誤りと母語の音韻的特徴との関係を考慮することによって頑健な辞書引きを実現する方法についても研究を進めている。

後者の機能、すなわち語句の意味の表示は、従来の対訳辞典で意図されてきた中心的な機能である。ここで注意が必要なのは多義語のあつかいである。多義語の語義を表示する場合、少なくとも3種類の方法がある。

- 出現頻度の高い順に語義を一列に並べる方法 (e.g. [12, 13])。この方法では、学習者が頻度の高い語義を優先的に調べることができるので、相対的に早く目的の語義に到達することができる。頻度の高い語義の方が頻繁に学習者の目にとまると考えられるので、学習の効果も高くなると期待できる。
- 語義間の派生関係を木構造的に明示する方法。たとえば、岩波国語辞典 [14] はこの方法に近い。多義語がもつ語義の間には比喩的派生関係(「動物のメスが産む」たまご)と「[医者などの]たまご」)や換喩的派生関係などが認められる [4, 5]。学習者にとってはこれらの語義を個別に覚えるより、派生関係にもとづいて体系的に覚える方が語の持つ意味の広がり理解が容易になると考えられる。

- 共起語のような語の出現文脈がシステムに与えられる場合、システムは共起情報を利用した多義性解消技術 (e.g. [18]) を利用してある程度の精度で語義を推定することができるので、尤度の高い語義をハイライトしたり、尤度の高い順に語義を並べかえたりする表示方法も考えられる。

これらの表示方法は一般には両立しないので、学習者が自由に切り換えられるようにするなど何らかの工夫が必要である。

2.2 日本語産出を支援する機能

学習者が発話したり、作文したりする場合、既知の定型句とオープンクラスの語を組み合わせることによって言語表現を産出していると考えられる。このうち辞書の利用がより有効なのはオープンクラスの語の選択である。ここでは、名詞・動詞・形容詞のようなオープンクラスの語の選択を支援する機能に焦点をあてる。

学習者が語選択に利用できる資源には少なくとも次の4種類がある。

- (1) 他言語からの対訳情報
- (2) シソーラス
- (3) 例文、共起例
- (4) 選択基準の解説

(1) はもっとも一般的な資源であるが、日本人用に編纂された英和辞書や中和辞書が与える情報が語選択には十分でないことは、そもそも日本人には訳語 (日単語) の選択基準を提示する必要がないことから明らかである。たとえば、研究者新英和辞典 [19] は “go” の訳語として「(ある場所・人 (の所)・方向へ) 行く、向かう」, 「(活動などに従事するために) [場所へ] 行く、通う」などをあげているが、「行く」と「向かう」の選択、あるいは「行く」と「通う」の選択のための基準は記載していない。このように対訳情報は主として語選択の候補を調べるのに利用される。また、(2) のシソーラスも同様の目的で利用される (e.g. [6, 8, 17])。これに対し、候補間の選択基準については、(3) 日本語側の例文や共起例からボトムアップに獲得する場合と (4) 選択基準の解説からトップダウンに獲得する場合が考えられる。従来の書籍辞書では主として、(3) の情報は国語辞典やのような日単語を五十音順にソートした辞書に記載されており (e.g. [13]), (4) の情報はシソーラス的な構造をもつ辞書に記載されている (e.g. [17])。

ここで注意したいのは、上の (1) の情報は他言語の見出し語、(2) は概念階層、(3) は日本語の見出し語、(4) は概念階層というように、いずれも異なるタイプのイン

デクスによって構造化されている点である。日本人が外国語文章を書く場合に英和・和英辞書や類語情報を利用するのと同様、外国人学習者も日本語産出の際にはこれら異なるタイプの情報、すなわち異なるタイプの辞書をすべて利用する必要があると考えられる。したがって、複数のタイプのインデクスを共用できる電子化辞書の特性を活かしてこれらの情報を統合することは語選択の支援に効果的であると言える。

この統合には少なくとも2つの相補的な方法が考えられる。一つは、個々の記述の間に相互参照のリンクをはることによってハイパーテキスト化する方法である。ハイパーテキスト化の利点は学習者の単語検索の負担を抑えられるという点だけではない。たとえば、異なる用例「金をとる」「飛行機をとる」からそれぞれ「とる」の異なる語義記述のページにリンクをはるというように、ハイパーテキストではリンクの文脈依存性を利用することができる [7]。

もう一つの方法は上の (1)~(3) の情報を利用して用例を自動検索するインタフェースを提供するというアプローチである。上で述べたように、用例は主として、複数の候補から最適な語を選択する際の基準として利用される。すなわち、語選択は表現したい内容にできるだけ類似した用例を検索する過程と見なすことができる。そこで、図1のような用例検索インタフェースを提供することによって語選択を支援する方法が考えられる。学習者は、表現したい意味に近い英語の語句あるいは日本語の既知の語句を入力し、日本語の用例を検索する。検索した用例からリンクをたどって辞書の語彙項目に行けるようにしておけば、日本語見出し語、多言語見出し語、概念階層とは異なる第4のインデクスとして用例を利用することができる。用例の検索には、事例ベースの機械翻訳 (e.g. [1]) や多義性解消 (e.g. [18]) で提案されている照合技術が応用できるだろう。ただし、言語学習支援が目的のシステムでは必ずしも最尤解を特定することが目標ではない。学習者がシステムと対話しながら検索結果に応じて類義語や共起語を入力に付加したり、意味的照合の厳しさを調節したりすることができる環境を提供することが重要である。既存の電子化辞書システムにも文字列照合による例文検索機能が見られるが、ここで述べる用例検索は照合に対訳情報やシソーラス、ジャンル情報などを利用する点で、より一般的な照合機能を想定している。言語学習という文脈を前提にする場合、今後様々な試行を通じて入力インタフェースや類似度計算法の洗練をはかる必要がある。また、どのような用例をどれくらい集めれ

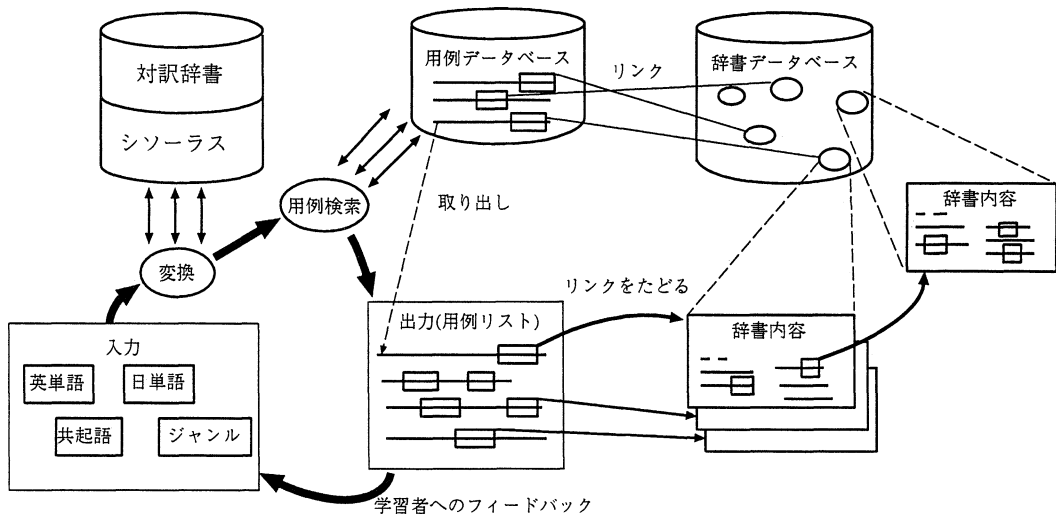


図 1: 用例検索インターフェース

ばよいのかに関して何らかの指針が必要である。

2.3 語彙の定着を支援する機能

語彙の定着を支援する機能としては次のようなものが考えられる。

- 電子単語集の生成・ブラウジング
- 練習問題の作成
- 付箋の編集
- しおりの検索

学習者の中には単語のリストを自作して暗記しようとする者が少なくない。そこで電子化辞書は、学習者が辞書検索する過程で語彙項目を選択的に登録することができる単語集編集機能をそなえていることが望ましい。この単語集をハイパーテキスト化して電子化辞書とリンクさせておけば、学習者はこれを自作のインデクスとして利用することもできる。

2.4 カスタマイズ機能

日本語学習の進度には個人差があるので、情報の表示の仕方に関して以下のようなカスタマイズができることが望ましい。

- かな・ローマ字表記の有無の選択
- 英語などの他言語への対訳の有無の選択
- 例文、説明文の難易度の選択
- 表示させる情報項目の選択
- 表示する用例の量に関する選択
- 多義語の語義の表示数に関する選択

3 プロトタイプを試作

我々は現在2節の議論にもとづきプロトタイプの試作を進めている。プロトタイプ試作の目的は、辞書の記述、用例の選択、ユーザインタフェースなどの適切性に関し学習者や専門家から広く意見を求め、設計の洗練をはかることである。そのため、今回のプロトタイプは World Wide Web ブラウザを利用する cgi プログラムとして実装している。

日本語理解を支援する機能 各語彙項目の記述は IPAL 日本語基本 { 動詞, 形容詞, 名詞 } 辞書 [15, 16] をもとにしている。また、シソーラスには分類語彙表を用いている [3]。例文・共起例などから辞書項目へのリンク情報は IPAL のハイパーテキスト化を行っている梁らの研究成果 [7] を利用している。IPAL に記載されていない読みがな、英語訳、中国語訳、語義間の派生関係などについては一部の語に限って手作業で追加した。図 2 に語彙項目の表示例を示す。見出し語検索の場合、まず見出し語のすべての語義について英語訳・語釈文・例文 (1 文のみ) を表示する。各語義はそれぞれ詳細情報を表示する別のページにリンクしている。詳細情報のページでは、語釈文・類語・反義語・文型・用例・慣用・活用例 (動詞のみ)・分類語彙表へのリンク・共起語・複合語 (名詞のみ) をデフォルトとして表示する。形態素解析には JUMAN[20] を利用した。

日本語産出を支援する機能 分類語彙表を利用し、シソーラスから語彙項目へたどることができるようにした。図 1 に示した用例検索インターフェースは現在実装中である。

語彙の定着を支援する機能 単語集を生成・ブラウジングする機能を現在実装中である。

カスタマイズ機能 読みがな・ローマ字表記, 英語訳, 中国語訳を一部の見出し語に追加し, 表示の要不要を学習者が選択できるようにした。

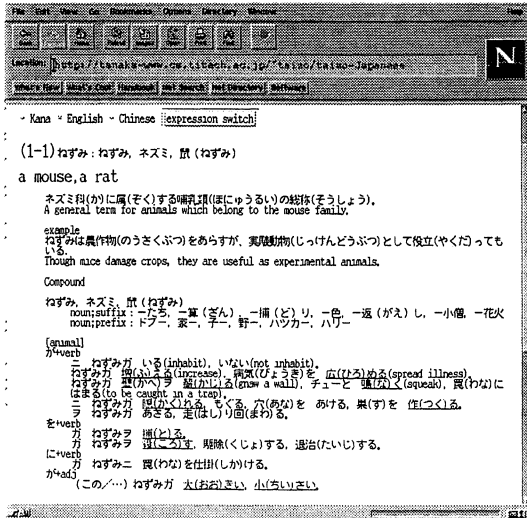


図 2: 名詞「ねずみ」の詳細画面

4 おわりに

本稿では, 電子化の特性を活かした外国人日本語学習者用辞書のデザインについて論じ, 現在試作中のプロトタイプについて報告した。2 節で論じた機能の中には実装には至らずに, 今後の課題として残されたものも少なくない。実装にいたらなかった理由としては, 辞書の基礎とした IPAL の情報不足の他に, プラットフォームとして用いた cgi の自由度の制限があげられる。前述のように, プロトタイプの試作は学習者や専門家からの意見を集めることを目的としている。現在も少しずつシステムの拡張・改良を進めているが, <http://tanaka-www.cs.titech.ac.jp/~inui/JLD.html> ですでに部分的に公開している。

謝辞

プロトタイプの試作にあたり情報を提供して下さった北陸先端大の梁慶昇氏, 奥村学氏に感謝します。

参考文献

- [1] L. Cranias, H. Papageorgiou, and S. Piperidis. A matching technique in example-based machine translation. In *Proceedings of the 14th International Conference on Computational Linguistics*, Vol. 1, pp. 100-104. COLING '94, 1994.

- [2] フォード順子. 聴解ディクテーションの「誤聴」分析—中・上級の文法の困難点を探る—. 筑波大学留学生センター日本語教育論文集, Vol. 7, , 1992.
- [3] 林大. 分類語彙表. 秀英出版, 1966.
- [4] 国広哲弥. 語彙論と辞書学. 言語, Vol. 24, No. 6, pp. 38-45, 1995.
- [5] 桑畑和佳子. I P A L 名詞辞書における下位区分間の関係記述の試み. IPA 技術発表会予稿集, 1995.
- [6] S. I. Landow and Bogus R. J., editors. *The Doubleday Roget's Thesaurus in Dictionary Form*. Doubleday, 1987.
- [7] 梁慶昇, 奥村学. I P A L 辞書の自動的ハイパーテキスト化. 言語処理学会年次大会予稿集, 1996.
- [8] D. Summers. ロングマン英語アクティベータ. 丸善, 1993.
- [9] 玉村文朗. 外国人のための日本語辞書構想. 言語, Vol. 24, No. 6, pp. 54-61, 1995.
- [10] 矢沢真人. 国語電子辞書引き比べ. 言語, Vol. 22, No. 5, pp. 22-27, 1993.
- [11] 高橋一夫ほか(編). 外国人のための基本語用例辞典. 文化庁, 第 2 版, 1987.
- [12] 望月孝逸ほか(編). 基礎日本語学習辞典. 凡人社, 第 1 版, 1986.
- [13] 竹林滋(編). 日英辞典. 研究社, 第 1 版, 1992.
- [14] 西尾実, 岩淵悦太郎, 水谷静夫(編). 岩波国語辞典. 岩波書店, 第 5 版, 1994.
- [15] 情報処理振興事業協会技術センター. 計算機用日本語基本動詞辞書 IPAL, 1987.
- [16] 情報処理振興事業協会技術センター. 計算機用日本語基本形容詞辞書 IPAL, 1990.
- [17] 遠藤織枝ほか(編). 類語例解辞典. 小学館, 第 1 版, 1995.
- [18] 藤井敦, 乾健太郎, 徳永健伸, 田中穂積. 動詞多義性解消における格要素の貢献度について. 情報処理学会 自然言語処理研究会, Vol. 111, No. 9, pp. 55-62, 1996.
- [19] 岩崎民平ほか(編). 新英和中辞典. 研究社, 第 4 版, 1983.
- [20] 松本裕治, 黒橋禎夫, 宇津呂武仁, 妙木裕, 長尾真. 日本語形態素解析システム JUMAN 使用説明書 Version 2.0, 1994. 2.