

語の関連性に着目した辞書リーディングシステム

A Dictionary Reading System using Relationships between Words

松村敦*, 末次美央, 宇陀則彦

Atsushi MATSUMURA, Mio SUETSUGU, Norihiko UDA

* 筑波大学大学院 図書館情報メディア研究科

Graduate School of Library, Information and Media Studies, University of Tsukuba

〒305-8550 茨城県つくば市春日1-2

E-mail: {matsumur, uda}@slis.tsukuba.ac.jp

辞書を読み物として活用することを目的として、語の関連性に着目した辞書リーディングシステムを構築した。本システムは、1つの語を引くと、同音異義語、反義語、有名フレーズに共起する語など7種類の関連語を提示する。これによって、利用者は興味を持つ語をたどることが可能となり、辞書を読み進めることができる。実際にシステムを利用してもらい、アンケートを実施したところ、辞書を面白く読み進めるためには、未知の語・未知の関連の発見が重要であることが示された。

We constructed a dictionary reading system using relationships between words. Whenever any words are looked up in our system, related words of seven classes are provided, such as synonyms, antonyms, words co-occured in well-known phrases, and so on. Consequently, users can read the dictionary through the network of interesting words. In the preliminary experiment, we found that unknown words and unknown relationships between words were important to make dictionaries attractive reading.

キーワード: 辞書リーディングシステム, 国語辞書, 語の関連性, シソーラス, フレーズ,
Dictionary Reading System, Japanese Dictionary, Relationships between words, Thesaurus, phrase

1 はじめに

通常、辞書は語の意味を調べるものであるが、それだけではなく、読み物のように使われることもある。例えば、『新明解国語辞典』は内容のユニークさから多くの読者を得ており、『新解さんの謎』[1]のように、その内容を楽しむための本が出版されたりしている。また、阿川は、漢和辞書を読むことに言及しており、『大漢和辞典』の「木部」を読んだときの方法や、そのときの木偏の漢字に対する所感を述べている[2]。

このように辞書を読み物として捉えた場合、その読みやすさを実現するには語の意

味を次々と調べていくための動機付けが必要である。始めから順に読んでいくという単純な方法も可能ではあるが、文脈の存在しない語の並びを読み続けることは必ずしも楽しいことではない。自然に次の語を調べるようにするためには、次の語へのつながりが必要となる。

そこで本研究では、語の関連性に着目し、語の並びに文脈を与えることを試みる。語の関連性を利用して辞書を読み進めることによって読み物としての辞書を提供するシステムの実現を目指した。関連に沿って語を続けて引くことにより、語と語を比較し

表1 語の関連性と分類

大分類	小分類	例
音	同音異義語	愛-藍, 愛-i/I
	同じ音を含む語	愛-知り合い, 無垢-椋鳥
表記	同じ漢字を含む語	海-海女, 糖分-血糖値
	複合語・熟語	象形-文字, 消費-消費者
意味	同義語・類義語	愛-恋愛, 火刑-焚刑
	反義語	会う-別れる, 愛-憎
	上位・下位	ネズミ-モルモット, 公害病-病気
	同分類の語	文庫-新書, 脂肪-たんぱく質
フレーズ	故事ことわざ	恋-盲目
	歌詞	春-小川, 小川-さらさら
	作品タイトル	春-爛漫, 赤-黒
	有名フレーズ	春-別れ
エピソード	社会的エピソード	レッド・パージ-GHQ, 花見-酔っ払い
	個人的エピソード	愛-重い, 白衣-穴
その他	-	食べる-箸, 花見-桜

差異や類似性を発見することが可能となる。その結果、辞書を読む面白さを感じることができるとはならないかと考える。本稿では、国語辞典を対象として構築した辞書リーディングシステムについて紹介し、その効果を議論する。

2 語の関連性の分析

2.1 語の関連の収集

システムに実装する語の関連を検討するために予備調査を行い、人が連想する語の関連を収集した。調査対象は15名で、次の手順で語のつながりのネットワークの作成を依頼した。まず、連想の始まりとなる語を決め、ホワイトボードなどの中央に書いてもらった。次に、この語を出発点として、連想される語をすべて書き出し、語と語を線で結んでもらった。

この調査で合計170の関連を収集した。収集した関連を、谷口ら[3]及び、小野[4]の研究で利用されている分類を参考にして整理したところ、音、表記、意味、フレーズ、エピソード、その他の6つの大分類に分けられた。この6分類は表1に示すようにさらに細かく15の小分類に分けられる。

ここで、「同分類の語」とは、同一の上位語を持つ語を指す。また、エピソードとは、何らかのエピソードを背景とする関連であり、社会的な話題と個人的なものに分けられた。なお、その他には、「花見」-「桜」のように文化や育った環境を背景とするものなどがあつた。

語の関連には、言語資源を使うことにより機械処理が可能なもの（音、表記、意味、フレーズ）と、個人の文脈に依存し揺れが大きいもの（エピソード、その他）があることが分かった。

表 2 語の関連付けの結果

小分類	件数	(%)	平均	最大値	最小値
同音異義語	33117	(0.3 %)	0.80	34	0
同じ音を含む語	5459275	(47.9 %)	131.21	3315	0
同じ漢字を含む語	4933003	(43.3 %)	118.56	1858	0
反義語	1479	(0.01 %)	0.04	5	0
同分類の語	839121	(7.4 %)	20.17	293	0
故事ことわざ	4389	(0.04 %)	0.11	197	0
有名フレーズ	130714	(1.1 %)	3.14	9646	0
全体	11401098		274.01	10026	0

2.2 関連データの作成

本システムで扱う語の関連データは前処理で作成した。対象とする関連の種類は、前節の15分類のうち言語資源を利用して機械処理可能なもののうち、同音異義語、同じ音を含む語、同じ漢字を含む語、反義語、同分類の語、故事ことわざ、有名フレーズの7分類である。また、関連付けの対象とした語は、『分類語彙表-増補改訂版-』[5]データベースに収録される101,070件のレコードから重複する語を除いた83,218語とした。

同音異義語、同じ音を含む語、同じ漢字を含む語については、見出し語の表記と読みを利用して関連付けを行った。ただし、同じ音を含む語は、読みが2文字以上の語を対象に、他の語にその読みが含まれるかどうかを判定し関連付けた。反義語は日本辞典 (<http://www.nihonjiten.com/>) の反対語一覧と照合し関連付けた。同分類の語の判定には分類語彙表における意味的な分類を利用した。具体的には、分類番号と段落番号が両方とも同じものを同分類と定義した。故事ことわざに共起する語は同じく日本語辞典のことわざ一覧を利用し、有名フレーズに共起する語は世界傑作格言集 (<http://kuroneko22.cool.ne.jp/>) を利用して、関連を付与した。なお、ことわざ一

覧、世界傑作格言集からの単語の切り出しには、日本語形態素解析システム 茶筌 (ChaSen) ver. 2.1 for Windows (<http://chasen-legacy.sourceforge.jp/>) を利用した。

関連付けを行った結果を表2に示す。関連語の件数、1語あたりの関連語数の平均、最大値、最小値を分類毎に求めた。これを見ると分類によって関連付けの件数に大きな差があることが分かる。特に、同じ音を含む語と同じ漢字を含む語が圧倒的に多く、非常に偏ったデータとなった。また、全体の最小値は0であり、関連語が全くない語もあることが分かる。

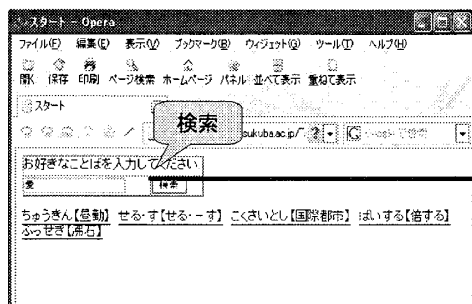
3 辞書リーディングシステム

3.1 概要

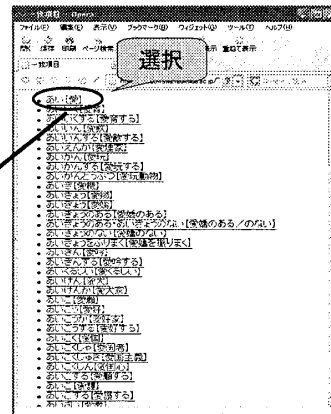
本システムは、前節で作成した語の関連を利用して、次々に語を調べ、辞書を読み進めることができるシステムである。本システムの利用の流れを図1に示す。

まず、スタート画面で最初に調べる単語を決定する。決定方法は、語を検索して検索結果リストから選択するか、ランダムに表示されている5つの語の中から選択するか2通りが可能である。本システムは辞

スタート画面



検索結果リスト



辞書記述画面

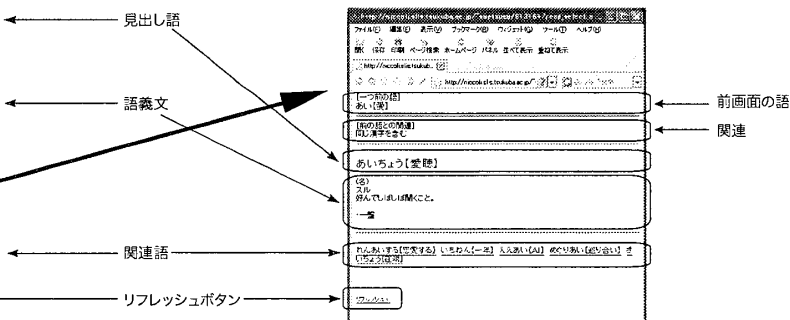
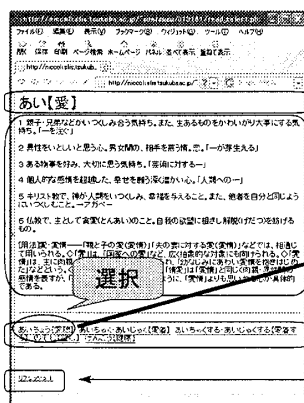


図1 システムの利用の流れ

書を読むことを想定しているため、特定の語が思い浮かばなくても読み始められるようにスタート画面ではランダムに単語を表示するようにした。

最初を選択した語の辞書記述画面には、選択した語と語義文、関連語が表示される。利用者は語義文を読むなどした後、5つの関連語の中から語を選択し、選択した語の辞書記述画面に移動する。2語目以降の辞書記述画面には、選択した語と語義文、関連語に加えて、前画面で表示した語および、その語との関連が表示される。これによって、語と語の関連を意識しながら辞書を調べることが可能となる。また、左下にある「リフ

レッシュ」というリンクをクリックするとスタート画面にいつでも戻ることができる。したがって、提示された語以外に調べたい語があれば、いつでも新しい語を調べることができるようになっている。なお、本システムは利用者を区別するために、利用に先立ってユーザ登録を行い、利用の際にはユーザ認証を行うようにした。

3.2 システム構成

本システムの構成を図2に示した。システムは、単語情報とユーザ情報を扱うデータベースとそれらを管理する語彙管理部と

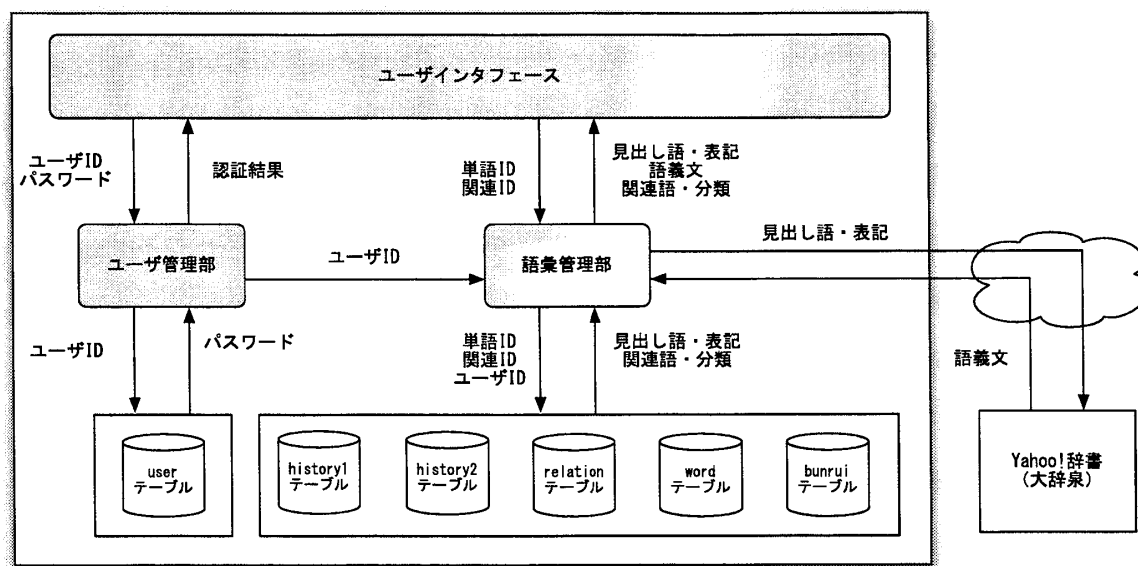


図2 システム構成図

ユーザ管理部で構成されている。

データベースの構成 本システムのデータベースは以下の6個のテーブルで構成されている。

word テーブル 関連付けの対象である単語リストを登録する。単語ID, 見出し語, 表記の3つのフィールドが設定されている。

bunrui テーブル 7つの関連についての分類IDと分類の名称を登録する。

relation テーブル 語と語を関連付けたデータを登録する。関連ID, 2つの単語ID, 分類ID, フレーズの5つのフィールドが設定されている。関連IDはあらかじめ作成している関連の通し番号である。故事ことわざと有名フレーズの2つの分類については, フレーズフィールドに元のフレーズを登録する。

history1 テーブル 利用者の履歴として, ユーザIDと7分類の各分類に対する選

択回数を記録する。

history2 テーブル 利用者のたどった経路をユーザID, 関連ID, 単語IDによって記録する。

user テーブル ユーザIDとパスワードを登録する。

語彙管理部 最初の語を選択すると, 選択した語の単語IDがシステムに送られる。送られてきた単語IDを使って, 語彙管理部は, データベースから見出し語と表記を取得する。それと同時に, 送られてきた単語IDを含む関連を検索し, その中からランダムに5つを選択する。得られた関連語の単語IDからそれぞれの見出し語と表記を取得し, 関連語の単語IDと関連IDを含む関連語のリンクを作成する。また, 得られた見出し語と表記を使い Yahoo!辞書の大辞泉を調べ, 語義文も取得する。

得られた情報は, 見出し語・表記, 語義文, 関連語(リンク付き)の順に構成し, ユーザに提示する。なお, Yahoo!辞書で見出し語と

表記に一致する辞書項目が見つからなかった場合は、「語義がありません」と表示するとともに、Yahoo!辞書の検索結果画面へのリンクを参照として表示する。

辞書記述画面で関連語を選択すると、選択した語の単語IDに加えて、表示中の語と選択した語の組み合わせが持つ関連IDがシステムに送られる。まず、単語IDを使って表示中の語と関連語の見出し語と表記を取得する。次に、関連IDを使って、関連の分類を取得する。関連語に対する語義文の取得と関連語の取得は前述と同様の処理を行う。

得られた情報は、前画面で表示した語、2つの語の関連、選択した語の表記、語義文、関連語（リンク付き）の順でユーザに提示する。

なお、ユーザが選択した単語IDと関連IDはユーザ管理部から送られるユーザIDとあわせて履歴情報として格納される。

ユーザ管理部 ユーザ管理部は、ユーザIDとパスワードを受け取り、ユーザ登録と認証を行う。また、履歴蓄積のために語彙管理部へユーザIDを渡す。

初回利用時にはユーザ登録画面でユーザIDとパスワードを入力しユーザ登録を行う。ユーザ管理部では、ユーザIDとパスワードを受け取り、ユーザIDの重複チェックを行いユーザを登録する。既に登録済みのユーザの場合は、ログイン画面からログインを行う。ログイン画面から送られるユーザIDとパスワードを利用して認証を行い、その結果を返す。

4 評価

4.1 概要

本システムの効果を明らかにするための予備的な調査として、実際にシステムを利用してもらい簡単なアンケートを実施した。

1. 普段の国語辞典との関わりについて教えてください。
 - (a) 国語辞典をどのようなときに使いますか。
 - (b) 国語辞典の利用頻度はだいたいどのくらいですか。i) ほぼ毎日 ii) 週に数回 iii) 月に数回 iv) 滅多に／全く使わない
 - (c) 国語辞典を面白いと思ったことはありますか。ある場合、それはどのようなときですか。
2. 今回のシステムを使ってみての感想をお聞かせください。
 - (a) 関連語を選択するときの決め手は何でしたか。
 - (b) 今後、このシステムで国語辞典を読みたいと思いましたか。
 - (c) 利用してみた感想を自由に記述してください。

図3 アンケート

対象は学部学生10名、大学院生3名、社会人2名の合計15名、期間は3日間とした。協力者はその期間内に自由にシステムを利用し、利用後にメールで図3のアンケートに答えてもらった。

4.2 結果と考察

4.2.1 システムの利用履歴

利用履歴から各分類が選択された回数を取得し、その平均、最大値、最小値を表3にまとめた。ただし、15人中2人に対しては、ログの取得を失敗したため、13人分の結果

表3 各小分類の選択回数(13人分)

小分類	計	平均	最大値	最小値
同音異義語	2	0.15	1	0
同じ音を含む語	129	9.93	39	0
同じ漢字を含む語	153	11.77	35	3
反義語	0	0	0	0
同分類の語	132	10.15	19	3
故事ことわざ	0	0	0	0
有名フレーズ	3	0.23	3	0
計	419	32.23	90	11

となっている。まず全体の合計を見ると、1人当たり平均約32回、最も多い人では90回語を引いていることが分かる。3日間という短い期間にもかかわらず、これだけ辞書を引いているのは、本システムの手軽さによる効果と考えられる。アンケートでも、「単純にことばを引き続けているだけで、結構楽しい」といったように、ことばを引き続ける楽しさを指摘する回答が見られた。

一方、分類毎の集計を見ると、反義語と故事ことわざは、一度も選択されなかった。これに対して、同じ漢字を含む語、同じ音を含む語、同分類の語が多く選択されていた。このような偏りが生じた主な原因は、システムが関連を選択する際に、表2に示される関連データの偏りの影響を受ける実装となっていた点である。利用者の分類毎の選択の嗜好については、システムの実装方法を検討した上で再度検証する必要がある。

4.2.2 関連語提示の効果

語の選択基準についてのアンケート結果を見ると、15人中7人が、関連語の選択の決め手として「聞いたことのない語」、「知らない語」のように未知の語をあげていた。一方、自由記述では「思わぬところで聞き

なれた語が出てきて面白い」、「予想外の単語や知らない単語が出てきて面白い」などの感想も見られた。このように、利用者が辞書を引く面白さを感じるのは、知らない語が出てきた時と語自体は既に知っているが関連が未知であった時であることが分かる。本システムは、多様な関連を考慮することによって、辞書を読むことを面白くする未知の語・未知の関連の提示に成功している。

しかしながら、関連語の選択の決め手として分類に着目した回答をしたのは2人であった。今回の実験では、関連性の分類は面白さの大きな要因ではなかった。実際に関連の種類まで考慮することが辞書を読むことにどの程度貢献するかということは、関連の選択頻度の偏りの詳細な分析を行うことで、今後より深く追求していくべき課題である。

また、関連語の選択回数は普段の国語辞典の使用頻度とは関係がなかった。したがって、語の意味を知りたいという明確な目的がなくても、手軽に面白さを感じながら辞書を読んでいることが分かった。また、「楽しみながら語彙を増やすことを目的とした」というように、学習効果を期待する記述もアンケートには見られた。

本システムの提供する関連語に基づいて

手軽に読み進めるという仕組みは、面白く読むだけでなく、語の理解を深めるという学習の面においても意義があると考えられる。

4.2.3 言語資源統合の課題

アンケートより辞書記述画面に「語義がありません」と表示されることが利用者にストレスを与えていることが分かった。原因は分類語彙表に収録された語を対象とした点である。分類語彙表と大辞泉では収録されている語が異なる上に表記が一致しないこともある。本システムは辞書を読むことを想定するため、国語辞典の語彙をベースにシステムを構築する必要がある。それと同時に、その他の言語資源と統合する際に、表記の差異等を吸収する仕組みが必要である。

4.2.4 利用者参加型システムへの拡張

今後の機能の拡張としては、利用者の履歴の活用がある。本システムには利用者がたどった経路を記録する仕組みがある。今回は履歴分析のために実装したが、これを使って利用者が自分の履歴を参照できるようにすることが考えられる。自分のたどった経路がネットワークとして可視化され、視覚的に把握可能となれば、利用者はまた別の面白さを得ることができると考える。

さらに、人ならではの関連を扱うことも考えられる。エピソードによる関連のように、既存の資源で機械的に処理することが困難な関連にも対応可能となれば、利用者が感じる面白さの幅はさらに広がるだろう。そのためには、関連付けを前処理で行い固定化するのではなく、利用者が語と語の関連を登録できる機能を加える必要がある。これによって利用者参加型のシステムの実現を目指したい。

5 おわりに

本研究は、国語辞典の記述を語の関連性に基づいて読むシステムを構築し、読み物としての辞書の実現を目指した。利用者実際に利用してもらったところ、辞書を読み進めるために重要となる要素は、未知の語・未知の関連の発見であることが明らかになった。また、語の意味を知りたいという明確な目的がない場合にも手軽に辞書を読み進められるということも、面白さを感じさせる要因であることが分かった。

今後の課題は、関連の偏りの解消と他の関連について実装を行い、利用者による評価実験を行うことである。さらに、利用者参加型の仕組みを提供することによって、より面白く辞書を読むシステムに拡張することを目指したい。

参考文献

- [1] 赤瀬川原平:「新解さんの謎」,東京,文藝春秋,317p.,1999.
- [2] 阿川弘之:「辞書を読む」,本,Vol.24, No. 3, pp.5-7, 1999.
- [3] 谷口すみ子; 赤堀侃司; 任都栗新; 杉村和枝:「日本語学習者の語彙習得 -語彙ネットワークの形成過程-」,日本語教育, No. 84, pp.87-91, 1994.
- [4] 小野正樹:「語彙ネットワーク -日本語母語話者と日本語学習者の自然連想方の調査から-」,筑波大学留学生センター日本語教育論集, No.16, pp.11-19, 2001.
- [5] 国立国語研究所編:分類語彙表,増補改訂版,2004,(CD-ROM).