自然言語処理　150－18
(2002. 7. 16)

単語親密度に基づく基本的語彙の選定

金杉　友子†　笠原　要††　稲子　希望††　天野　成昭††

† NTT アドバンステクノロジ株式会社
†† 日本電信電話株式会社, NTT コミュニケーション科学基礎研究所
† †† 〒619-0237 京都府相楽郡精華町光台 2-4
E-mail: †{kanasugi,kaname,inago,amano}@cslab.kecl.ntt.co.jp

あらまし　意味に関する言語処理技術の基盤となりうる概念辞書である“常識概念体系”を構築する第一歩として、人々の概念的な思考で共通して利用していると推定される基本的な語の集合(“基本的語彙”と呼ぶ)を選定した。選定の対象としては学研国語大辞典(9万5千見出し語)を用い、選定の尺度として、心理実験により評定される単語のなじみ深さを表す単語属性である単語親密度を用いた。過去の研究において12歳児の理解語彙数の推測値が2万5千と報告されており、別の語彙数調査結果から、同数の語彙を成人の94%が知っていると推測される。そこで、基本的語彙数を2万5千程度と定めた。国語辞典の見出し語について、過去の単語親密度に関するデータベースに含まれていない3万3千語の追加の評定実験を行い、9万5千語から親密度が高い2万7千語を基本的語彙として実際に選定した。
キーワード　単語親密度, 基本語, オントロジー, タクソノミー

Selection of a Basic Vocabulary Based on Word Familiarity Ratings

Tomoko KANASUGI†, Kaname KASAHARA††, Nozomu INAGO††, and Shigeaki AMANO††

† NTT Advanced Technology Corporation
†† NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation
† ††2-4, Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619–0237 Japan
E-mail: †{kanasugi,kaname,inago,amano}@cslab.kecl.ntt.co.jp

Abstract　As the first step of constructing a dictionary of word concepts, the “Commonsense Concept Database,” which will be a base for language processing technologies regarding meaning, we selected basic words which are supposed to be commonly used by Japanese adults. We selected the basic words from a Japanese dictionary in which the number of word entries is about 95,000. In a previous study, the size of the basic words which a Japanese child of twelve years knew was estimated to be 25,000. From the another recent psychological study estimating the number of the vocabulary in Japanese speakers, we were able to estimate that 25,000 of the Japanese basic words were known by 94% of Japanese adults. Therefore, we selected the number of basic words for Commonsense Concept Database to be 25,000. As a measure of selecting the basic word, we used word familiarity ratings. We did farther psychological experiments of rating familiarity of words in the Japanese dictionary which had not been listed in the word familiarity database previously published. Finally, we selected all words with a familiarity rating above five (between seven point scale) which gave us around 27,000 words out of the 95,000 entries of the dictionary.
Key words　Word Familiarity, basic words, ontology, taxonomy

1. はじめに

テキストに関わる情報処理の最近トピックの1つとして、言葉の意味情報を前提としたシステムが提案されるようになっている。例えば、Semantic Web [1] や意味情報ネットワークSIONet [2] では、サービス提供者や利用者が作成するオントロジーやタクソノミーといった言葉の意味の体系の存在を前提としている。一方、情報検索の分野では、文献 [3], [4] のような、テキストから言葉の意味の類似性判別のためのデータベースを自動作成し、それを用いて行う概念的な検索が近年注目されている。また、そこで用いられている単語の属性ベクトルのデータベース(“概念ベース”)は、テキストセグメンテーション [5]

や情報の可視化システム [6] にも利用されている。このように、言葉の意味をいかに工学的に表現し利用するかは、今後の情報処理技術の課題の 1 つと言えよう。

これを解決する現実的な方法として、適用するドメインやアプリケーションに応じたオントロジーを個々に作成し利用している。ただし、現在注目を浴びている Semantic Web のような P2P サービスでは、様々なドメインのオントロジー同士やさらにはユーザのオントロジーを比較して利用することが必要となる。これを解消する方法論の 1 つとして全てのオントロジーに共通となるであろうオントロジーの上位部分 (例えば IEEE Standard Upper Ontology, http://suo.ieee.org/) を作成し、これを参照して個々のオントロジーを作成することが試みられている。しかし、この方法でも共通ではないオントロジーの下位レベルの比較が生じてしまう。

一方、知識獲得の研究では、多くの人々が共通して保有していると予想される言葉の常識的な意味に関するデータベース (「概念辞書」と呼ぶ) を構築し利用する研究が行われており、過去に、大規模知識ベース構築が行われた [7], [8]。多数の単語について様々な概念的関連性が記述されているが、これらの大規模知識ベースは必ずしも幅広く利用されていない。その理由としては、大規模知識ベースの構造が複雑な点や単語間の概念記述に欠落などが存在する点、知識ベース自体の正しさの評価が行われていない点など様々である。

これら課題を解消して広く利用される概念辞書を作成するためのアプローチの 1 つとしては、やみくもに多数の多数の単語を対象とするのではなく、各自然言語において必要最小限の基本的な語彙を対象とした概念辞書を作成することが挙げられる。これによって、必要とされる単語間の概念的な記述を最小限におさえることができるので、記述の欠落や構造の複雑性の問題を低減することができる。

このような観点で我々は、“中規模” の概念辞書の構築を試みている [9]。国語辞典 [10] 中の基本語の語義を単位として、各種言語データベース [11]〜[13] と心理実験に基づいて概念辞書を構築することを検討している。その第一歩としては、基本語を具体的に選定することが必要である。

基本的な語彙をどのように選定するか自体は難しい問題であり、言語学や心理学など様々な分野で検討が行われている [14], [16]〜[21]。例えば国語学の研究では、教科書や雑誌などに出現する語彙の調査が行われている。その出現頻度に基づいて基本語が選定されているが、頻出しない単語の中にも人間の常識的な思考を形成するために重要な語が存在する。従って、人間の常識を構成するために必要最小限の基本的な語彙を実際に選定する場合には、テキストコーパスにおける使用頻度の調査にとどまらず、被験者を用いた心理実験によって決定することが必要である。それに対して、数百程度の語彙に対する調査とそれを元にした理解語彙数の推定 [14], [15] は行われているが、実際に基本語を信頼できる方法論で選定する試みは行われていない。

そこで本稿では、日本語の概念辞書構築のための第一歩として、単語に対する心理的な尺度の 1 つである単語親密度の評定を通して基本的な語彙の選定を行う試みについて述べる。単語親密度とは、人間が提示された単語に対して感じるなじみの程度を数値化した尺度である。日本語 7 万語に対して 32 名の被験者によって評定された単語親密度のデータベースが存在する [13]。このデータベースと新聞記事中の単語の出現頻度の比較が行われ、出現頻度と親密度の相関は高いが、出現頻度が低い単語の中に親密度が高い語があることが報告されている [22] このように単語親密度は、評定のコストはかかるがもれなく基本的な語彙を選定する尺度として有用な尺度である。

本稿ではまず、基本的語彙の選定に関して既存研究と今回の選定の試みに方針について説明する。次に、実際の選定方法およびその結果について説明し、さらに考察を加える。

2. 過去の基本語選定の研究

日常生活で多くの人間が共通して用いる語をここでは基本語と呼ぶ。基本語の研究調査は言語学、言語教育上の必要性から過去に様々に行われてきた。ここでは、基本語彙調査・選定の代表的な三種類の考え方について説明する (文献 [23], [24] を参照した。)

2.1 基礎語彙

言語教育を主なる目的として、日常生活の言語活動に必要な語彙を主として個人が主観的に決定したものが基礎語彙である。

英語では、1930 年に C.K. Ogden によって 850 語の基礎語彙 (Basic English) が提案された。Ogden は、これらのみで日常の事柄すべてが表現できると主張した。この考えを汲んで Longman Dictionary of Contemporary English(1992) では、7 万 5 千以上の見出し語の説明が 2,000 の基礎語彙 (the Longman Defining Vocabulary) のみで記述されている。日本語では、土居による 1,100 語の『基礎日本語』(1943) があげられる。

また、外国人の日本語教育のための基礎語彙のデータベースが各種存在している。文献 [20] で挙げられた 7 種類の基礎語彙のデータベースの語彙数は、約 500 から約 5000 までと多岐に渡る。注目すべきは、全データベース共通する語彙数は 278 と非常に少ない点である。

これら基礎語彙は、個人の主観に基づき基本語が選定されているので、体系的であることが特徴である。しかし一方で、語数や収録する語の選定基準は個人の主観や選定の目的に左右される点が問題である。

2.2 基本語彙

言語資料の統計分析により客観的に選定された基本語は、基本語彙と呼ばれる。これは、雑誌、新聞、教科書等の各種言語資料の調査分析に立脚するものである。資料中で使用頻度が高く、しかも対象とする言語資料中で広い分野に用いられている語が基本語彙として選定されている。例えば、文献 [16] では、Brown Corpus 中での単語の出現頻度より英語の基本語彙を選定している。日本語では、国立国語研究所が種々の言語資料についての調査を行って来た [17]〜[19]。

出現頻度が高い順に選定された基本語彙は、もととなる言語資料中で使用されている単語の大半を占めるため、その言語に

おける基本語の多くが含まれている可能性が高い。しかし、この手法ではよほど語彙量の多い資料を対象にしない限り、資料による語彙や語の使用率の偏りは避け難い。

2.3 理解語彙、使用語彙

最後に、ある個人や集団の理解できる語彙、言語活動に使用することができる語彙を被験者実験により調査して基本語を決定する手法があげられる。それぞれ理解語彙、使用語彙と呼ばれる。

日本語における理解語彙については数種類の調査がなされている。例えば文献[14]では、広辞苑中の500語を被験者に提示して語を「知っている／知っていない」かを内省させた。その結果より理解語彙数を推定している。例えば12歳の理解語彙はは2万5千語、20歳では4万8千語と見積もっている。また、文献[21]では、児童の語彙発達調査を目的として児童作文の使用語彙の調査を行っっている。

このように、ある特定の集団や特定の発達段階の人々における理解語彙数の推定ということは行われているものの、前記2種と違い、理解語彙、使用語彙調査に全面的に基づいた基本語彙表は作成されていないため、基本語の語彙リストという形ではまとめられていない。

3. 常識概念体系のための基本語 (基本的語彙) 選定

上記の通り、様々な基本語の選定方法が存在し、それぞれは示唆に富んでいるが、それらの選定方法に従って得られている基本語のリストは、これからの言語に関わる情報処理技術の基盤となる概念辞書を構成する基本語としては、十分ではない。

基礎語彙として提案された語彙は、言語教育等に利用されており有用であるが、先に述べた通り、作成した個人の主観に大きく依存するところがあるために、基盤的な利用は難しいと考えられる。一方、基本語彙は、言語資料をもとにして選定されるのでそれ自体では信頼性がある。しかし、言語資料 (コーパス) が一般的な人々が行う言語活動での基本語をよく再現しているかは明らかではない。

それに対して使用語彙や理解語彙は実際に個々の被験者の判断に基づいており、概念辞書を構成する基本語にふさわしいと考えられる。ただし、使用語彙の調査は、特定のトピックに限定された作文などで調査されているので、使用し得る語彙全てをあらわしていない点が問題である。理解語彙については、被験者による評定実験が必要なために、少数の語彙での調査結果から語彙数を推定する研究が多く、少なくとも日本語においては基本語リストという形で提案はされてはいない。また、理解語彙の調査は、被験者の「その語を知っている／知らない」という内省的な判断基準に基づいているので、判断の再現性が明らかではない点が問題である。

これらの問題を考慮して我々は、概念辞書のための基本語 (基本的語彙) を以下のような方針で定めた。

[方針 1] 基本的語彙選定の母集団に国語辞典を利用

これは、時代や性差に左右されにくい普遍的な語彙が多数収録されている言語資料から基本語を選定するためである。理解語彙の研究[14]では、20歳の理解語彙数は5万8千語と見積もられている。そこで、この2倍程度の約10万語以上の見出し語のある国語辞典を選定の対象とすれば妥当と考えた。

[方針 2] 基本的語彙選定基準に単語親密度を利用

我々は、特定の個人の主観的判断、あるいは特定の言語資料の統計分析に基づいた語数、語彙選定は行わなず、被験者による心理実験を行った結果を基本語選定のための尺度とする。しかしその尺度は、語の理解性のような曖昧なものでは再現性に乏しいと予想される。そのため、ある程度客観的な言語資料の統計調査結果との比較で有効性が確認できるものでなければならない。そこで、文献『日本語の語彙特性』第一巻 (以下 『日本語の語彙特性』) [13] で用いられている、単語親密度という尺度を利用する。

単語親密度は、刺激語として提示された単語に対して被験者が判定するなじみの度合いについての主観評価に基づく尺度で、複数の被験者において1から7までの7段階で評定された結果を平均化したものである。『日本語の語彙特性』は、新明解国語辞典の見出し語 (68,855語) を対象に、32名の被験者による評定を通して得られた単語親密度のデータベースである (一部の例: 表1)。

『日本語の語彙特性』第七巻[22]では、この結果と新聞記事 (朝日新聞14年分) の単語使用頻度との関連性が調査されている。単語親密度と使用頻度の相関係数は0.634で有意な相関があり、使用頻度の高い単語のほとんどが単語親密度も高くなった。しかし、「たまねぎ」、「からあげ」のように、使用頻度が低いにもか変わらず単語親密度が高い語が存在しており、基本語彙から洩れる基本的な単語も取得できている。従って、理解語彙を基本的語彙とする場合には、使用頻度よりも単語親密度に基づいて選定した方が好ましいと考えられる。

表1 『日本語の語彙特性』の一例

単語	単語親密度	単語	単語親密度
大きい	6.65	一簀	1.34
お母さん	6.56	穎脱	1.31
教える	6.12	衍字	1.21
おしゃべり	6.46	掩蔽	1.15
会計	6.09	杣	1.12

上記の理由により、基本的語彙の選定の尺度として単語親密度を用い、一定値以上の単語を基本的語彙とする。今回、『日本語の語彙特性』に収録されていない語については新に評定実験を行う。さらに、今回の実験の際に、過去に評定された単語を再評定し、尺度としての単語親密度の再現性についても考察を加えた。

[方針 3] 理解語彙数の推定を利用する

基本的語彙数は、過去に行われた理解語彙数の推定調査を参考にして決定する。基本的語彙の構成単語の決定に際しては、単語親密度の一定値以上の単語とする。詳しくは、次章の4.3.3項で述べる。

次章では、上記の3つの方針に従って行った、実際の選択の手続きについて説明する。

4. 実　　験

本章では、基本的語彙を選定する母集団となる国語辞典の内容とその見出し語へ単語親密度を付与する手順、そして、基本的語彙数の推定と、最終的に選定された基本的語彙について説明する。

4.1 国語辞典

基本的語彙を選出する母集団として用いる国語辞典としては、94,928 の見出し語がある学研国語大辞典 [10] を用いた。この内で『日本語の語彙特性』に収録されており単語親密度がすでに付与されている語が 62,356 存在しており、基本的語彙選定の際に、これらの語の単語親密度はそのまま利用することにした。収録されていない 32,572 語については、次に述べる追加の評定実験によって単語親密度を新たに付与した。

4.2 追加の単語親密度評定実験

18 歳以上の日本人 40 名（男女各 20 名ずつ）を被験者として募集し、32,572 語に関する単語親密度の評定実験を行った。

実験は被験者自身がパソコンを操作して進める。まず、被験者ごとに順序をランダム化した刺激語がモニタに提示される。被験者は、その語を見て同画面上の 1 から 7 のうち 1 つのボタンをクリックすることにより、該当親密度を評定する（図 1）。評定のペースは被験者個人に任せたが、一日の評定語数を 4000 程度に保つことと、1 から 7 までの値をすべて使って評定を行うことを条件づけた。

実験に先だって、各自の親密度評価基準を作るための練習を行った。『日本語の語彙特性』の語彙から今回の評定での対象とはなっていない 9,000 語を刺激語とした。また、『日本語の語彙特性』の評定結果と今回の評定結果の相関を確認するために、実験では練習時とは別の『日本語の語彙特性』中の 1 万語を 32,572 語に加えて評定させた。実験後、被験者の評定の再現性を確認するために評定した語彙から 3 千語を選出し、再度評定させた。

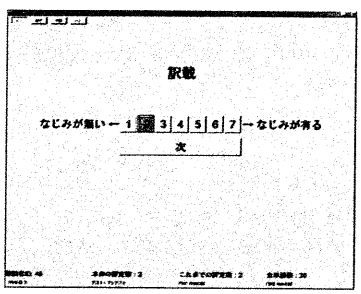

図 1　単語親密度評定実験画面

4.3 実験結果

4.3.1 被験者評定の再現性検証

実験終了後に 40 名の被験者に再評定させた 3 千の単語について、被験者ごとの評定値の相関値を調べたところ、平均は 0.71 であり評定の再現性が認められた。しかし、図 2 の相関係数に対する被験者数のヒストグラムの通り、0.8 付近をピークとする分布とともに 0.5 以下に外れ値が存在していた。そこで、相関係数 0.5 以下の被験者 (6 名) の評定データは不適切と考え、今後の評定結果から除外した。以降、相関係数 0.5 以上の評価者 34 名の評定値の平均を各対象語の単語親密度とする。

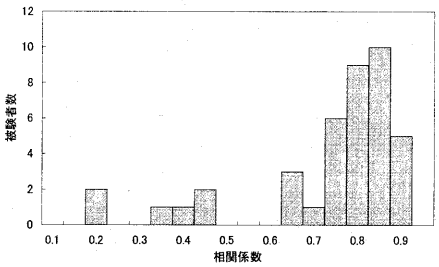

図 2　評定の再現性チェック結果

4.3.2 単語親密度の分布

国語辞典の全語彙 (94,928) の単語親密度の平均値は、3.895 であり、その分布は図 3 となった。一方、今回の実験で追加の評定実験を行った 32,572 語の親密度の平均は 3.363 であった (分布図: 図 4)。『日本語の語彙特性』と一致しない学研国語大辞典中の単語に比して一致する語の単語親密度は高い傾向にあり、基本語である可能性が高いと示唆される。ただし、図 4 の分布を見ると、学研国語大辞典のみに現われる単語でも親密度が 6 以上の語が存在する。そのため、国語辞典等の見出し語リスト複数を単純に比較して共通する語を基本語とする方法では、基本語の欠落が発生する恐れがある。従って、基本語リストを作成する場合には、今回の単語親密度のような尺度を個々の単語に関して評定することが重要であると言えよう。

表 2　基本統計量

	語彙数	単語親密度	
		平均	標準偏差
追加評定実験	32,572	3.363	1.183
国語辞典 [10] 見出し語	94,928	3.895	1.358

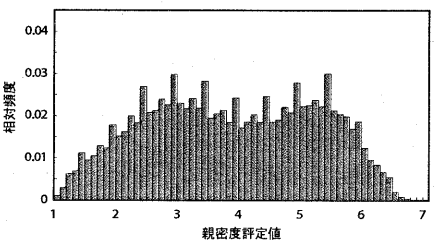

図 3　学研国語辞典の見出し語 (94,928) の単語親密度分布

4.3.3 基本語彙数の推定

過去になされた理解語彙の調査では、高校一年生を対象とした国立国語研究所の調査 [15]、阪本 (1955) によるの 6〜20 歳を対象とした調査 [14] がある。その調査結果の一部を表 3 に挙げる。小学校終了、あるいは義務教育終了時ではおおかた理解

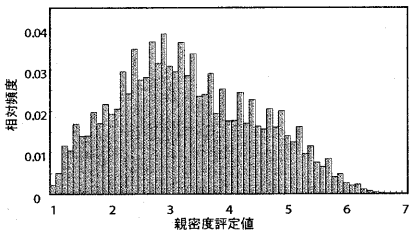

図 4　追加実験で評定した語 (32,572) の単語親密度分布

表 3　過去の理解語数調査

文献	年齢	理解語彙数 (平均)
阪本 55 [14]	12	25,668
	15	40,461
	20	48,336
森岡 51 [15]	(高校 1 年生)	30,664

語数は 2 万 5 千〜5 万語という結果に収まっている。そこで、この程度の語数を基本的語彙数と考えて選定する。

しかし、このままでは語彙数は決定できても、構成単語を決定できない。そこで、語彙数と語彙構成語を同時に特定する調査 [25] の結果から推定する。この研究では、9 千語の刺激語について、知っている／知らないを 60 名の被験者 (19 歳から 29 歳) に評定させ、その結果と刺激語の単語親密度から語彙数を推定するものである。結果として、成人の過半数が知っている可能性があると推測される日本語の語彙数を 6 万 6 千と推定している。この推定に用いられた図 5 を見ると、単語親密度の 5 以上の約 2 万 5 千語については 94%の成人が知っていると見積もられる。そこで、学研国語大辞典の見出し語に対して『日本語の語彙特性』および今回の評定実験によって得られた単語親密度の結果から同じ 5 以上の単語を選定すると、26,992 語となりほぼ同じ結果となった。これらの語もは成人ならば 94%が知っていることが期待される。そこで、この語彙を基本的語彙とした。

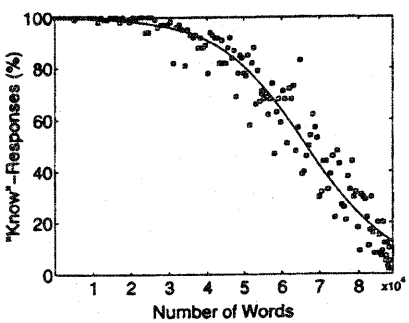

図 5　語彙数の推定 (文献 [25] より引用)

今回の選定された基本的語彙の中で親密度が非常に高い語と、親密度下限である 5 付近の値を取る語の一部を表 4 に挙げる。

表 4　基本的語彙例

単語	親密度	単語	親密度
トイレ	6.88	青臭い	5.00
TV	6.85	垢	5.00
生年月日	6.85	いたずら	5.00
京都	6.85	駆け寄る	5.00
大坂	6.85	辛うじて	5.00

5.　分析と考察

今回、概念辞書を構成する基本的語彙の選定の尺度として、単語親密度を用いた。本章は、この尺度の選定における有効性を検討する。

5.1　単語親密度の評定の再現性

図 6 は、『日本語の語彙特性』中の 1 万語について、評定された単語親密度と今回の実験で再評定した単語親密度でプロットしたものである。両者の相関係数は 0.96 と算出され、非常に高い相関があることがわかる。このことより、30 名程度の被験者による評定結果としての単語親密度は、再現性の高い安定した尺度であると言えよう。

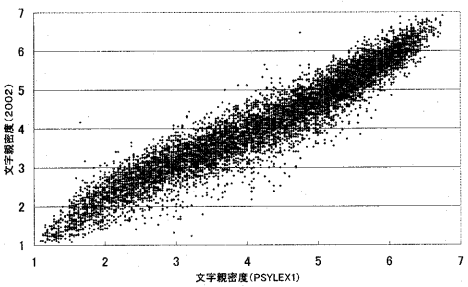

図 6　1 万語の複数評定値の相関

5.2　親密度の揺れ

上記の通り単語親密度は、再現性の高い単語尺度であることがわかったが、実験を行った時期や場所によって多少変動し、単語ごとによっても異なる。以下では、このようなパラメータの変動に伴う単語親密度の揺らぎに関して考察する。

5.2.1　個　人　差

図 7 は、今回の評定実験において、各被験者が評定した 32,572 語の親密度の平均値の分布である。これを見ると、各被験者の評定の基準は必ずしも等しくはないことが推測される。さらなる解析が必要であるが、原因の 1 つとして、被験者の理解語彙数の多少が影響していると予想される。

5.2.2　単　語　差

図 8 は、42,572 語それぞれに対して、被験者が評定した親密度を平均した時の標準偏差の分布を表す。単語ごとの評定値の平均は様々であるが、平均からの揺れを表す標準偏差は 1.2 程度であり、その分布はほぼ対称である。このことより、単語ごとの評定の揺れは存在するが、被験者によって極端に評定結果が異なる単語は少ないと予想される。

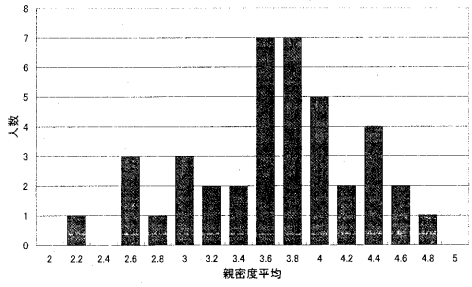

図 7　被験者ごとの評定値の平均

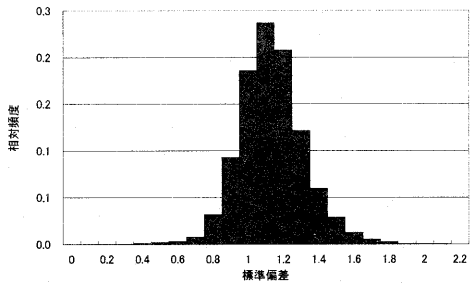

図 8　単語ごとの評定値の標準偏差

5.2.3　地域差、時代差

図 6 で見られた通り、同一の単語に対する異なる実験での単語親密度の評定結果は非常に相関が高い。ただし、個々の単語を見ると評定結果に差が生じている。『日本語語彙特性』の評定実験は、関東地区 (神奈川県厚木市) で 1995 年から 1996 年にかけて行われた。一方今回の実験は、関西地区 (京都府相楽郡) で 2002 年に行われた。この場合には、主として地域差と時代差が評定結果に多少は影響を与えると予想される。2 度の評定を行った 1 万の単語に対して、今回 (2002 年) の評定値の方が 1 以上大きかった語は 27、過去 (1995-1996 年) の評定値の方が 1 以上大きかった語は 167 であり、それぞれ全体に占める割合は小さい。これら全てを時代性、地域性で分類できるかは明らかでは無いが、一部の語については、推測が可能である。その一部を表 5 に示す。「炭疽病」や「パラリンピック」は時事的な原因で評定値に差が生じたと推測されるし、「京阪」や「銀ぶら」などは、地域差が影響していると見られる。

表 5　単語親密度評定値の差が大きな単語 (一部)

単語	評定値 (1995, 関東)	評定値 (2002, 関西)	評定値差
炭疽病	1.65	4.17	2.52
京阪	4.75	6.47	1.72
パラリンピック	4.25	5.64	1.39
銀ぶら	4.06	2.38	-1.68
しょっつる	4.06	2.17	-1.89

6.　おわりに

本稿では、人間の常識的な概念知識をコンピュータで表現する概念辞書の検討の第一歩としての基本語の選定について報告した。単語親密度を選択のための尺度として用い、理解語彙数の推定に基づいて、親密度の値が 5 以上の 2 万 7 千語を約 10 万語彙の国語辞典から選定した。

今後は、この基本的語彙の有効性をコーパス中の語彙調査と比較して検討する。また、この基本語彙と国語辞典を基にして、日常的に用いられる語義を選定して概念辞書の構築を進める予定である。

文　献

[1] T. Berners-Lee, J. Hendler, and O. Lassila, "The semantic web," Sientific America, pp. 34–43, May 2001.

[2] 星合 隆成, 小柳 恵一, ビルゲ スクバタール, 久保田 稔, 柴田 弘, 酒井 隆道, "意味情報ネットワークアーキテクチャ," 電気情報通信学会論文誌, vol. J84-B, no. 3, pp. 138 – 145, 2001.

[3] H. Schutze and J.O. Pedersen, "Information retrieval based on word senses," Fourth Annual Symp. on Document Analysis and Information Retrieval, pp. 161–175, 1995.

[4] 熊本 睦, 島田 茂夫, 加藤 垣昭, "概念ベースの情報検索への適用—概念ベースを用いた検索特性の評価—," in 情処研報, SIG-ICS 115, pp. 9 – 16, 1999.

[5] 別所 克人, "単語の概念ベクトルを用いたテキストセグメーテンション," 情報処理学会論文誌, vol. 42, no. 11, 2001.

[6] 宮原 伸二, 藤田 悦郎, 安部 伸治, 林 恭仁, "散策型映像ポータルシステム associaguide の提案," 電子情報通信学会総合大会, D-8-7, p.104, 2002.

[7] R.V.Guha and D.B.Lenat, "Cyc: A midterm report," AI Magazine, vol. 11, no. 3, pp. 32–59, 1990.

[8] Japan Electronic Dictionary Research Institute., EDR Electronic Dictionary Technical Guide, tr-042, 1993.

[9] 天野 成昭, 笠原 要, "基本語彙に対する知識データベースの構築," 人工知能学会全国大会, 2C3-09, 2002.

[10] 金田一 春彦, 池田 弥三郎 (編), "学研 国語大辞典 第二版," 学習研究社, 1988.

[11] 池原 悟, 宮崎 正弘, 白井 諭, 横尾 昭男, 中岩 浩己, 小倉 健太郎, 大山 芳史, 林 良彦 (編), "日本語語彙大系," 岩波書店, 1997.

[12] 笠原 要 , 松澤 和光, 石川 勉, "国語辞書を利用した日常語の類似性判別," 情報処理学会論文誌, Vol.38, No. 7, pp.1272-1284, 1997.

[13] 天野 成昭, 近藤 公久, "日本語の語彙特性," 第 1 巻 単語親密度, 三省堂, 2000.

[14] 阪本 一郎, "読みと作文の心理," 牧書店, 1955.

[15] 森岡 健二, "義務教育修了者に対する語彙調査の試み," 国立国語研究所年報 (2), 1951.

[16] H. Kucera and W. N. Francis, "Computational analysis of present-day American English. Province, RI," Brown University Press, 1967.

[17] 国立国語研究所, "現代雑誌九十種の用語用事 (1)," 国立国語研究報告 21. 秀英出版, 1962.

[18] 国立国語研究所, "電子計算機による新聞の語彙調査 (II)," 秀英出版, 1971.

[19] 国立国語研究所, "高校教科書の語彙調査," 秀英出版, 1983.

[20] 国立国語研究所, "日本語教育基本語彙七種 比較対照表," 日本語教育指導参考書 9. 大蔵省印刷局, 1987.

[21] 井上 一郎, "語彙力の発達とその育成 －国語科学習基本語彙選定の視座から－," 明治図書出版, 2001.

[22] 天野 成昭, 近藤 公久, "日本語の語彙特性," 第 7 巻. 三省堂, 2000.

[23] 国立国語研究所, "語彙の研究と教育 (上)," 日本語教育指導参考書 12. 大蔵省印刷局, 1984.

[24] 小池 清治, 小林 賢次, 細川 英雄, 犬飼 隆 (編), "日本語学キーワード辞典," 朝倉書店, 1997.

[25] Shigeaki Amano and Tadahisa Kondo, "Estimation of mental lexicon size with word familiarity database," in Proc. of Intl. Conf. on Spoken Language Processing, vol. 5, pp. 2119 – 2122, 1998.