「基本語意味データベース: Lexeed」の構築

笠原 要 † 佐藤 浩史 † フランシス ボンド † 田中 貴秋 † 藤田 早苗 †
金杉 友子 ‡ 天野 成昭 †
† 日本電信電話 (株) NTT コミュニケーション科学基礎研究所
‡NTT アドバンステクノロジ (株)
〒619-0237 京都府相楽郡精華町光台 2-4
{kaname, hiro, bond, takaaki, sanae, kanasugi, amano}@cslab.kecl.ntt.co.jp

概要

単語の意味を用いた情報処理技術の基盤となりうる基本語の言語知識ベースとして，「基本語彙知識ベース」の構築を進めている．本稿では，その構想と，中核となる 2.8 万の基本語の意味記述である「基本語意味データベース」の構築状況について説明する．

キーワード: 意味，統語，語用, 基本語，電子化辞書, 語義

Construction of a Japanese Semantic Lexicon: Lexeed

Kaname Kasahara† Hiroshi Sato† Francis Bond†
Takaaki Tanaka† Sanae Fujita† Tomoko Kanasugi‡ Shigeaki Amano†
†NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation ‡NTT Advanced Technology Corporation
2-4, Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237 Japan
{kaname, hiro, bond, takaaki, sanae, kanasugi, amano}@cslab.kecl.ntt.co.jp

Abstract

In this paper we describe the construction of the Japanese Semantic Lexicon: Lexeed. The lexicon contains the most familiar 28,000 thousand Japanese words, which also form the defining vocabulary for their definitions. The lexicon is designed to be a base for further information processing.

Keywords: semantics, syntactics, pragmatics, basic vocabulary, machine readable dictionary, sense

1 はじめに

テキストに関わる情報処理の最近のトピックとして，言葉の意味情報を利用したシステムの提案があげられる．例えば，Semantic Web[1] 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 [2, 3] 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注目されている．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．

これまではこの課題に対して，処理の種類や目的に応じた言葉の意味データベースを個々に作成することで対応してきた．代表的なデータベースとして，似た意味を持つ単語同士を分類したシソーラス（類語辞典）があげられる．シソーラスは，人間が類義語を選択する際に参照れるだけではなく，機械翻訳や語義曖昧性解消のための知識源 (例えば [4, 5]) として用いられている．別種類のデータベースとしては，自然言語処理用に整理した電子化辞書 (例えば Longman Dictionary of Contemporary English, LDOCE[6]) があり，単語の類似性判別や語義の曖昧性解消等に用いられている．シソーラスと電子化辞書はそれぞれ単語の意味を分類的，定義的な側面から表しているので，両者を併用して利用できれば幅広い情報処理への適用が可能となる．しかし，個別に作成されたこれらのデータベースを矛盾なく統合することは困難である．さらに，自然言語処理で利用するためには，統語情報も必要となる．

このような考えから，意味データベースを構築するために，複数の意味のデータや統語，語用情報を統合的に作成するアプローチがこれまでになされている．代表的試みとして，日本電子化辞書研究所で開発された EDR 電子化辞書 [7] があげられる．これは，日本語単語 20 万語の辞書，英語辞書， 対訳，シソーラス，コーパスが整備された大規模で統合的な意味データベースである．EDR 電子化辞書は規模の点では十分であるが，現在の日本語情報処理では十分用いられていない．実用的な意味データベースを構築するためには，多数の語彙への対応が必要となるが，データの整

合性を保つことに困難さがあり，データの質に問題が出てくる．この問題に対応するためには，必要十分な語彙に限定してデータベースを構築することが考えられるが，そのためにはどの程度の語彙を収録すれば十分であるかを定量的に示すことが必要である．

これらの課題を解決する統合的な単語の意味に関する知識ベースとして，我々は「基本語彙知識ベース」の構築を進めている [8, 9]．これは，日常生活で多くの人が共通して用いる語 (「基本語」と呼ぶ) を心理実験に基づいて選定し，これを意味，統語，語用の側面より関係付けた知識ベースである．基本語彙知識ベースをシソーラスやテキストコーパスと併用することで，基本語以外の語についても対応できることを目標としている．基本語彙知識ベース構築の際に最も重要なことは，単語の “意味の初期値” を適切に構築することであり，利用者の可読性が高く，かつ，自然言語処理での利用が容易であることと考えている．これらの条件を満たす基本語の意味記述として，国語辞典を知識源とした「基本語意味データベース」Lexeed の構築を進めており，現在は，単語表記，品詞，語義文，用例文を収集している．一例として，基本語「ドライバー」の意味記述を図 1 に示す．

本稿では，基本語彙知識ベース構築のアプローチおよび，基本語の選定，Lexeed の構築状況について説明する．

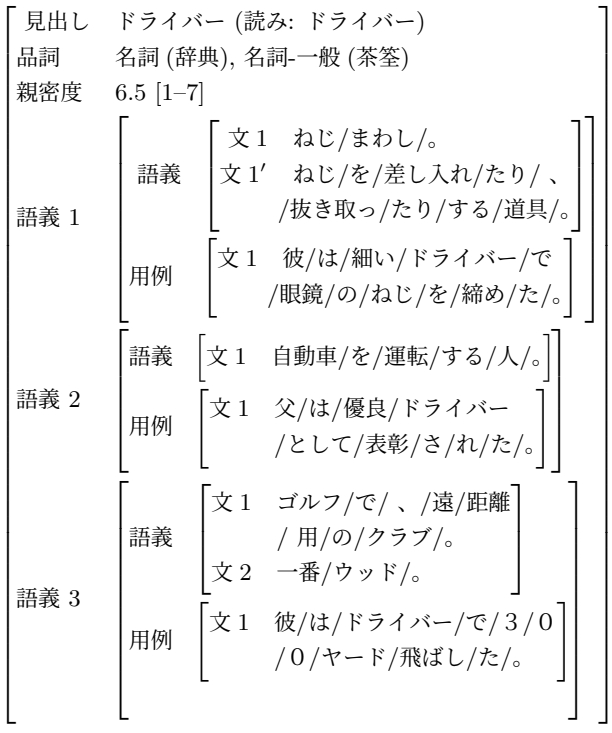

図 1　Lexeed における「ドライバー」の意味記述

2　基本語彙知識ベース

我々は，自然言語処理や知的情報処理のための日本語の基本的な言語知識ベースを目指して，以下のアプローチで基本語彙知識ベースの構築を進めている．

[1] 基本語の基本的な語義を意味の単位として設定

我々は，データベースの品質を高めることを重視して，様々な語の意味を説明するために必要十分な基本語を対象とした．基本語の意味を緻密に表現できれば，未知語の意味を基本語のいずれかの語義，あるいはその組み合わせによって表現できると考えている．基本語は，国語辞典の見出し語彙より，心理実験を通して多くの人が共通に知っていると推測される基本語を選ぶことで決定する．

このように選定される基本語であっても，現在は使われていない古い意味や，専門的な意味を語義として含んでいる．また，基本語とたまたま単語表記が同形のなじみの無い語も存在する．そこで，語義文等の意味記述を刺激として語義のなじみの度合いを評定する心理実験を行い，不要な語義や語を削除する．

[2] 高品質の語義の意味記述作成

基本語の語義の意味記述は，その語を他の語と意味において識別する基本情報であり，基本語彙知識ベースでは，これに基づいて統語や語用情報，さらなる意味情報が付与される．そのため，基本語の意味記述を適切に作成することが，基本語彙知識ベース全体の品質を決める大きな要因となる．

日本語については，LDOCE のような組織化された電子辞書が存在しないため，書籍出版用の国語辞典データの語義文から意味記述を作成する．語義文に未知語が含まれていると，その語の意味を得ることができなくなるため，語義文は基本語と機能語のみで記述する．

また，語義文が分かりにくい場合，データベース構築のために行う各種の心理実験に支障をきたす恐れがある．そのため，語義文を基本語で記述する編集作業を行う際には，わかりやすい表現に書き換えることも行う．さらに，心理実験の際，被験者の理解を助ける情報として，用例文も付与する．

[3] 基本語の意味・統語・語用情報を作成

基本語彙知識ベースを様々な情報処理における基盤として利用することを可能とするため，基本語の基本語義の意味記述として，語義文と用例文に加えて，様々な情報を付与する．まず統語情報は，語義の意味記述や用例文について構文解析を行い，その誤りを人手で修正した構文木コーパスを作成し，文法の修正も同時に行う．この過程を繰り返すことにより，基本語で記述された辞典的記述については高精度の構文解析が可能となると期待できる．

また，獲得した構文木コーパスを利用し，見出しの基本語と意味記述中の単語間の意味関係 (上位－下位，全体－部分等) を取得し，既存のシソーラスと組み合わせることにより，基本語のオントロジーが構築できる．さらに心理実験で，基本語に対して連想され

る語や語義を収集することにより，統語処理のみでは解消できない曖昧性に対処するための語用情報を収集する．これらを組織的に発展させることにより，基本語の意味記述が自己完結していて，かつ，十分な意味，統語，語用情報および，高品質な構文解析器をそなえた基本語の言語知識ベースを実現することができる．

[4] 未知語の意味を基本語の意味で表現
未知語については，コーパスにおける語の係り受け関係や共起語の傾向等による単語間の関連性の度合い，品詞等の統語情報に基づいて，どの基本語の語義と意味が類似しているかを推定し，類似した語義の意味を未知語に割り当てることを試みる．

図 2 に，基本語彙知識ベースの構成と構築の進め方を示す．まず国語辞典に含まれる基本語の語義文と用例文を収集，編集する．そして，なじみのない語義を削除し，語義間の連想関係を心理実験により収集したものが，語義文，用例文と併せて基本語の意味記述の初期値となる「基本語意味データベース」Lexeed である．

この Lexeed に基づいて構築される統語処理のための構文木コーパスが「檜」である [10, 11]．日本語の主辞駆動句構造文法 (Head-driven Phrase Structure Grammar, HPSG) にもとづいて，語義文に統語情報が付与される．「檜」では，コーパスの構築と文法の改良を併せて行うので，コーパスの構築後には，基本語で記述された辞書形式のテキストについては精度の高い構文解析が可能となる．また HPSG では，統語情報のみでだけではなく意味情報も表現することが可能となるため，見出し語に対する語義文中の各語の関係より単語間の上下関係などの意味データも取得することが期待できる．この結果と既存のシソーラスを併用して，基本語同士を関係づけることで，基本語のオントロジー構築を進める．

さらに，意味や統語だけでは曖昧性を解消できない文の解析や生成などの処理への対応としては，単語や語義のなじみの度合いのデータ等の語用情報をデーターベースとして蓄積することで対応する．このような手順で基本語彙知識ベースは，基本語の基本語義に関する意味記述「Lexeed」，統語情報「檜」，オントロジー，語用情報を持つ総合的な知識ベースとなる．

以下では，基本語彙知識ベースの基本語の選定結果と，その意味記述データベースとして構築中の「基本語意味データベース」Lexeed について説明する．

3 基本語の選定

3.1 既存研究

基本語の研究調査は言語学，言語教育上の必要性から様々に行われてきた．文献 [12, 13] によれば，基本語調査・選定には代表的な三種類の考え方がある，

まずは，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

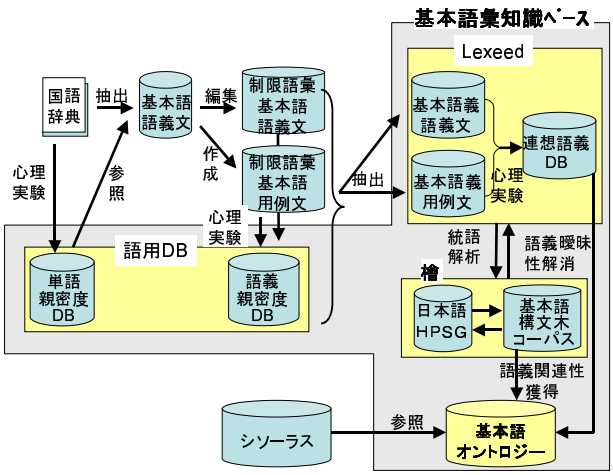

図 2　基本語彙知識ベースの構成と構築手順

した「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[6] 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．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．そのため，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はあるが，単語の意味のデータベースのための基本語としては十分ではない．

次に，言語資料の統計分析により客観的に選定された「基本語彙」と呼ばれるものがある．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．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として選定されている．例えば，文献 [14] では，Brown Corpus 中での単語の出現頻度より英語の基本語を選定している．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 [15, 16]．出現頻度が高い語は，元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が多く，偏りの無い資料を対象にしない限り，語彙の使用率の偏りは避け難く，言語資料中には頻出しないが日常的には重要な語が欠けてしまう恐れなどがある．よって，これを単語の意味データベースの基本語とするには不適と考えた．

最後に，ある個人や集団の理解できる語彙，あるいは，言語活動に使用することができる語彙としての基本語があげられる．これらはそれぞれ，「理解語彙」，「使用語彙」と呼ばれ，日本語においても数種類の調

査がなされている．例えば文献 [17] 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 2 万 5 千語，20 歳では 4 万 8 千語と見積もられている．また，文献 [18] 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，使用語彙数の推定は行われているが，前記 2 種と違い，基本語リストという形ではまとめられていないため，これらを意味データベースのための基本語情報として使うことはできない．

3.2 基本語選定のアプローチ

これらの問題を考慮して我々は，基本語彙知識ベースのための基本語を以下のようなアプローチに基づき定めた．

[1] 基本語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基本語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 [17] では，20 歳の理解語彙数は 4 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 万語以上を収録した国語辞典を選定の母集団とすれば妥当と考えた．

[2] 基本語の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基本語の選定は行わず，被験者による心理実験を行った結果を基本語選定の尺度として利用する．その尺度は，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を考えた．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』) [19] 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値で，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 [20] 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 で有意な相関があると分析されている．

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

[3] 理解語彙数の推定を利用

基本語数は，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．基本語の構成単語の決定に際しては，単語親密度が一定値以上の単語とする．

3.3 基本語の選定結果

上記のアプローチに基づいて，学研国語大辞典の電子化データ [21] の見出し語を用い，掲載されている見出し語及び子見出し語の中の名詞，代名詞，動詞，形容詞，形容動詞，副詞，連体詞，感動詞，接続詞を対象として，単語親密度 5 以上の 2.8 万語を基本語として選定した．この親密度以上の単語は，成人の 94% 以上が知っていると推測されている [22]．基本語の品詞毎の出現傾向について，表 2 に示す．

表 2 基本的語彙の品詞分類

品詞名	分類語数	割合 (%)
名詞	21806	77.1
動詞	2927	10.4
形容動詞	1611	5.7
副詞	873	3.1
形容詞	481	1.7
サ変名詞	271	1.0
感動詞	161	0.6
接続詞	66	0.2
連体詞	39	0.1
代名詞	35	0.1
合計	28,270	100.0

3.4 評価

今回選定した 2.8 万語の基本語について，一般的なテキストコーパス中の出現語をどの程度カバーしているかを調査した．テキストコーパス中の単語の出現頻度のデータベースとして，日本語の語彙特性第 7 巻 [20] を用いた．これは，1985 年から 1998 年までの 14 年間に発行された朝日新聞の記事データを日本語形態素解析システム「すもも」[23] で解析し，出現した単語や文字の傾向を分析したデータベースである．調査結果を表 3 に示す．

コーパス中の普通名詞, 動詞, 形容詞, 形容動詞について比較したところ，基本語 28,270 語中の約 97% の 27,465 語がコーパス中で出現しており，基本語のほぼ全てがコーパス中でも現れていることがわかる．一方，これらの品詞の語彙について基本語がどの程度カバーしているかを調べた結果，延べ語数で約 72%，異なり語数で約 14% であった．コーパス中には出現頻度が 1,2 回程度しか出現しない単語が多く存在するため，基本語のカバー率は異なり語数ではかなり低い．一方コーパス中で多数出現する語の多くは基本語に含まれているため，1/5 以下の語彙数であっても延べ語数では 7 割ものカバー率になったと考えられる．

この結果より，異なり語数で見た場合には，基本語

のみを対象としては新聞記事コーパスの解析などへの利用には不十分と危惧されるが，出現する延べ語数のカバー率はかなり高いため，未知語を基本語で近似するなどの方法を検討すれば，自然言語処理への基本語彙知識ベースの適用は有効と言えよう．

表 3　新聞記事中の基本語の出現傾向

	普通名詞/用言
全出現語	
異なり語数	196,547
延べ語数	129,840,929
基本的語	
異なり語数	27,465 (14.0)
延べ語数	93,777,294 (72.2)

(括弧内の数字は，全出現語彙に対する割合 (%))

4　基本語意味データベース　Lexeed

本章では，2.8 万の基本語に関する意味記述である Lexeed の構築状況について説明する．Lexeed は，図 2 に示した通り，基本語の語義より選定された基本的な語義を対象とし，形態素解析された語義文及び用例文，さらに基本語義間の連想データより構成される．基本語義の選定方法としては，語義文及び用例文を被験者に提示して，語義のなじみの度合い（語義親密度）を評定する心理実験に基づく方法を予定している．そのために，現在は，基本語の全ての語義について語義文と用例文の作成を進めている．

Lexeed では，語義文や用例文中に基本語に含まれない自立語があると，その語の意味をデータ内で参照できない．そのため，語義文と用例文は基本語と機能語のみで記述を行う．また，基本語の語義をどのように区分し，それをどのように記述するかについては，既存の電子化辞典を初期値として，それを書き換えることによって構築を進めている．以下，その構築状況について述べる．

語義区分及び語義文の初期値として，基本語の選定に用いた学研国大語辞典 [21] の電子データを用いている．2.8 万の基本語の語義数は 45,335(平均 1.6 語義/語) であり，語義文の文数は 81,100(平均 1.79 文/語義) である．4.5 万の語義それぞれについて，4 名の作業者によって語義文の基本語による書き換え及び，品詞タグ付けの作業を行った．形態素解析システムには茶筌 (`http://chasen.aist-nara.ac.jp/`) を用い，その際，形態素辞書に含まれていなかった基本語については辞書に追加した．作業は，茶筌に基づく品詞タグ付きコーパス作成支援 GUI ツール VisualMorphs[24] を用いて行った．結果の一例として，元となる学研国語大辞典に記載された語義文と，書き換えられた語義文例を表 4 に示す．基本語「朝露」（あさつゆ）の場合 “original” に対応する元の語義文について，4 名の作業者がそれぞれ異なる書き換えを行っている．また「ドライバー」については，4 名中 3 名は，単語表記を変更して同一の語義文に書き換えている．

表 4　基本語の語義文の書き換え例

ID	語義文
	「朝露」(あさつゆ)
original	朝、おりている露。朝の露。
01	朝、いろんな所に付いている水滴。
02	朝、大気が冷えて、水滴と成って、物の表面に付いたもの。
03	朝、降りている水滴。
04	朝、降りている水滴。朝、水蒸気が水滴に成ったもの。
	「ドライバー」
original	ねじまわし。
01,03,04	ねじ回し。
02	ねじを差し入れたり、抜き取ったりする道具。

次に，同じ語義について書き換えられた 4 件の語義文に対して，どれが適切であるかを別途評価した．具体的には，2 名で個々に評価し，意見が一致しない場合には第三者が評価した．その結果，最初の 2 名で判断が一致した割合は，59% であった．表 4 の例では，基本語「朝露」の場合には，04:“朝、降りている水滴。朝、水蒸気が水滴に成ったもの。”，「ドライバー」では，02:“ねじを差し入れたり、抜き取ったりする道具。” が適切な語義文として選定されている．これらの作業の結果，45,335 語義の語義中で，82% (37,248 語義) が書き換えにより元の語義文と異なるものとなった．

それぞれの語義文に含まれる自立語の品詞毎の出現傾向を表 5 に示す．“original” は，国語辞典の語義文について，形態素解析「茶筌」に添付されている形態素辞書をそのまま用いて解析を行った結果であり，“edited” は，人手で書き換えした結果である．original では，語義文中に未知語が出現しているが，edited では，それが無くなっており，書き換えによって，基本語という制限語彙での書き換えができていることが示されている．ただし，図 1 における’遠/距離/’ のような複合表現が含まれないように書き換えすべきであるかなど，検討すべき点が残されている．また，制限語彙として基本語全てを想定したが，書きえられた語義文では約 6 割 (16,914) の基本語が使われていた．

先に述べた通り，基本語の意味記述としては語義文のみではなく，用例文も対象としている．これを作成する方法としては，コーパスから基本語の現れる用例を収集し，その語義を人手で決定することも考えられる．しかし，基本語やその語義の全てがコーパス中で出現するわけではない．そこで，国語辞典編集の経験者によって用例文を作成し，語義文と同様の編集作業を現在進めている．

表 5　語義文中の品詞毎の自立語の出現語数 (異なり)

品詞名	original 語数 (%)	edited 語数 (%)
名詞	15,734 (59.2)	10,452 (61.8)
動詞	4,471 (16.8)	2,243 (13.2)
形容動詞	749 (02.8)	745 (04.4)
副詞	790 (02.9)	454 (02.6)
形容詞	591 (02.2)	357 (02.1)
サ変名詞	3,392 (12.7)	2,512 (14.8)
感動詞	56 (00.2)	49 (00.2)
接続詞	59 (00.2)	44 (00.2)
連体詞	42 (00.1)	24 (00.1)
代名詞	55 (00.2)	34 (00.2)
未知語	621 (02.3)	0 (00.0)
合計	26,560 (100)	16,914 (100)

5　基本語の語義文の分かり易さの評価

Lexeed 中の語義文と用例文は，分かり易く記述されていることが必要とされる．本章では，語義文の書き換えの結果を分かり易さの観点より評価した結果を報告する．

文や文章の分かり易さは，様々な要因に対する人の評価尺度の総体であると考えられる．例えば，計算機マニュアルの分かり易さを定量評価した研究では，分かり易さの要因として，音読できる度合いである「読み易さ」と内容を理解できる度合いである「理解のし易さ」，そして，マニュアルを検索する速度の「使い易さ」の 3 つを想定している [25].「使い易さ」は，マニュアルの利用というタスクに固有の要因と考えられるため，本章では，要因「読み易さ」と「理解のし易さ」それぞれに関連する語義文の統計量や指標に基づいて，書き換えの効果を検証する．また，小規模な主観評価の結果についても述べる．

5.1　読み易さの評価

文献 [25] において，「読み易さ」と関わりが大きいと主張されている文字種において，基本語の書き換え前後の語義文を比較した (表 6).

平仮名率，英数・片仮名率，漢字率は，基本語の語義文全てを構成する文字数における，平仮名文字，英数・片仮名文字，漢字の割合である．“original” は，学研国語大辞典における基本語の語義文 (81,100 文，1,121,982 文字) の情報である．また，“edited” は，“original” に対して基本語を制限語彙として書き換えた結果である．文数 (74,412) は，書き換え前に比べて 1 割弱減少し，文字数 (1,116,751) は，ほぼ同数となっている．書き換えによって文字数の変動がわずかであるにも関わらず，平仮名率が 53.5% から 48.2% に減少しており，反対に漢字率が 32.3% から 39.3% に増大している．[25] では，漢字率は読み易さと正の相関があり，平仮名率は負の相関があると報告されているので，語義文の文字種の比較結果より，基本語の語義文がより読み易くなったと推測できる．

しかしながら，漢字の割合が増大したことに起因して，普段見かけないような難しい漢字が多く含まれる場合，反対に語義文が読みにくくなる恐れがある．そこで，語義文を構成する文字の文字親密度の調査を併せて行った．文字親密度とは，人間が文字に対してどの程度なじみがあるかという主観的な評価値であり，文字の知覚しやすさやその文字を含む単語の認知しやすさなどの心的過程に密接に関連していると考えられている．日本語の語彙特性 第 5 巻 [26] には，JIS 漢字 6,355 文字と平仮名，片仮名，アルファベット，記号類等 6,847 文字を対象として，24 名の被験者によって 7 段階 (1:なじみが無い，7:なじみがある) の文字親密度の評定実験を行った結果が記載されている．このデータベースを用い，語義文を構成する文字の文字親密度の平均を求めた．その結果，書き換えによって文字親密度の平均値はほぼ変化がないことがわかった．また，文字種毎に文字親密度の平均値を見た場合には，片仮名・英数字で多少親密度の変動があるが，それ以外は差がない．特に漢字については，先に述べた通り，書き換えによって語義文中の文字における割合が増大しているが，その文字親密度は書き換え前の漢字の文字親密度平均と同じであることがわかった．従って，書き換えによって，読み易さに正の相関がある漢字の割合が語義文において増加し，さらにその文字のなじみの度合いは元の国語辞典の語義文と差が無いので，書き換えは，読み易さの向上に結びついていると言える．

表 6　基本語の語義文の読み易さの比較評価

	original	edited
平仮名率 (%)	53.5	48.2
英数・片仮名率 (%)	3.5	3.4
漢字率 (%)	32.3	39.3
平均文字親密度	6.19	6.20
平仮名	6.29	6.29
片仮名・英数	6.18	6.22
漢字	6.27	6.26

5.2　理解し易さの評価

語義文を理解するためにはまず，各文の単語を認識，理解し，さらに文の統語構造を解析できることが重要である．文献 [25] では，これを示す要因として文の簡潔さ，あるいは冗長さを考え，その指標として，文の平均的な長さや述語数の平均等を取り上げている．そこで，語義文の形態素解析，構文解析を行い，書き換えの効果について評価した．

形態素解析処理については，語義文の編集作業と同じく茶筌を用いた．書き換え作業前の語義文については，奈良先端大より配布されている形態素辞書をそのまま用い，自動解析した結果を分析している．そのた

め，正しく解析されていない結果も含まれている．そして統語解析は，Support Vector Machines に基づく日本語係り受け解析器 Cabocha[27] を用い，語義文の文節 (チャンク) 数を調べた．その結果を表 7 に示す．

この結果を見ると，文字数，形態素数，文節数のそれぞれにおいて，書き換えによって，語義文が多少長くなっていることがわかる．例えば単語数で言えば，1 形態素程度，書き換えによって増えている．理解のし易さとして文の記述の簡潔さで捕らえた場合には，わずかではあるが，書き換えによって理解のし易さが低下していると推測される．この理由として，書き換えの際に語彙を制限しているために，使えない語を別の語や節，句で置き換えることが必要となり，結果，記述が冗長になったとことがあげられる．

文献 [25] では，文に含まれる用言の数は，文の長さとともに文記述の簡潔さと関わり，用言数が少ない程，分かり易い簡潔な文であると主張している．そこで，書き換え前後の基本語の語義文についても調査を行った．表 7 中の “用言” とは，1 文に含まれる用言 (動詞，形容詞，形容動詞) の平均数を示している．語義文書き換え結果については，平均文長と同様に，わずかであるが，用言の数が増えている．

表 7　基本語の語義文の理解し易さに関わる統計量

	original	edited
平均文長 (文字)	13.8	15.0
平均文長 (単語)	8.4	9.3
用言	1.3	1.4
平均文長 (文節)	3.8	4.4

文の理解し易さとしては，文の記述の簡潔さだけではなく，文の表す意味をイメージすることが容易であるかも大きく関わる．しかし，これを直接測定することは困難である．そこで，文の意味のイメージし易さと，文を構成する個々の単語の意味のイメージし易さの平均値の間の相関は高いと仮定し，語義文の分析を行った．

高橋らは，文の表す意味をイメージし易いかどうかについて，(「-的」,「-正」,「-度」) のような接辞及び形態素辞書より得られる抽象性に関する情報より得られる抽象語句の割合を指標として用いている．しかし，これらは人間の心理的属性とどの程度一致しているかは検証されていない．そこで本稿では，心理実験に基づいて得られた単語の心的属性である，単語親密度と心像性に基づいた評価を行う．単語親密度とは，基本語の選定において説明した通り，単語のなじみの度合いを表す尺度であり，なじみのあまり無い語で記述された語義文はその意味を理解しにくいと考えられる．一方，心像性とは，単語の心的イメージの想起し易さを表す心的属性であり，約 5 万語について，30 名の被験者による 7 段階の評定実験 (1:非常にイメージしにくい〜7:非常にイメージしやすい) 結果が集計されている [28]．この文献では，単語親密度との関係について調査されており，0.78 の高い相関を示すが，親密度が高い語の一部には心像性が低い語 (例,「経済」,「政府」) が存在すると報告されている．そこで，語義文を構成する自立語の異なりと延べ語彙について，単語親密度と心像性の両方の値の平均値を調べた (表 8)．

単語親密度に関しては，異なり語彙では，語義文の書き換えによって単語親密度平均は 5.07 から 5.60 と，1 割以上高くなっている．これは，書き換え前の語義文にはなじみの無い語も多く出現しているが，これらなじみのある語に書き換えられたためと考えられる．そのため書き換えによって，理解し易い語義文となったと思われる．一方，語義文を構成する自立語の出現に関する延べの単語親密度を見ると，どちらも親密度が 6 近くで差はわずかであり，語義文全体を見た場合には，単語のなじみの程度では差がほとんどない．

一方，単語の心像性の値は，述べ平均も異なりの平均も，語義文の書き換え前後で大きな差が無かった．そのため，単語のイメージのし易さという点では，ほぼ同等であったと考えられる．

以上の結果より，語義文の書き換えによって文を構成する単語は，全体的に見て，なじみの度合いもイメージのし易さもほぼ同等であることがわかった．よって，元の語義文に現れるなじみが無い自立語をなじみのある基本語で書き換えることができたといえる．

表 8　基本語の語義文の理解し易さに関わる統計量

出現自立語彙 (異なり)	original	edited
語彙数	26,560	16,914
親密度平均	5.07	5.60
心像性平均	4.34	4.47
出現自立語彙 (延べ)	original	edited
語彙数	315,339	323,555
親密度平均	5.91	5.88
心像性平均	4.61	4.57

5.3　主観評価

評価対象とする語義文は，2.8 万の基本語の 4.5 万語義よりランダムに抽出した 300 の語義の語義文である．個々の語義について，書き換え前の語義文と書き換え後の語義文を 2 名の評価者に提示し，どちらか一方の表現が分かり易いか，あるいは，同等に分かり易いかを判定させた．また，評価者は，300 語義について評価を行った後に，一致しなかった判定についての理由を話し合い，再判定した．判定結果を表 9 に示す．

“一致語義数” とは，300 件の語義中で 2 名の作業者の判断が一致した語義を表し，“original(%)” は，その内で書き換え前が分かり易いと判断された割合，“edited(%)” は，制限語彙による書き換えた文の方が分かり易いと判断された割合を示す．また，“同等”

は，一致語義の中でどちらの語義文も同じ程度分かり易いと判断された割合である．また，“fine” は，3 種類の判断が独立であるとみなした場合の結果であり，“coarse” は，一方の作業者が “同等” と判断し，他方が “原文” か “編集文” のいずれかを選定した場合には後者の選定で一致しているとみなす場合の結果である．

2 人の作業者が話し合いを行う前に行った評価の一致件数は fine で 147 件 (49%) であり，分かり易さに関する作業者の評価尺度が十分一致していなかった．それに対して，話し合いを行った後の一致件数は 224 件 (75%) であり，話し合いにより作業者間の評価尺度の揺れが少なくなったといえる．

次に，語義文の原文と編集文の分かり易さの評価値を比較すると，作業者間の話し合いを行う前後，及び，一致の条件に関わらず，書き換えられた編集文が分かり易いと判断された割合が 6 割前後，国語辞典の元の語義文が分かり易いと判断された割合が 3 割強という結果となった．これにより，書き換えによって語義文が分かり易く表現されるようになったと言える．

表 9　語義文の分かり易さの主観評価結果

	話し合い前		話し合い後	
	fine	coarse	fine	coarse
一致語義数 (件)	147	245	224	264
original(%)	32.7	34.3	34.4	36.0
edited(%)	59.9	61.2	58.9	58.3
同等 (%)	7.5	4.5	6.7	5.7

6　おわりに

本稿では，意味に関する情報処理の基盤となりうる基本語の言語知識ベースである「基本語彙知識ベース」について，その構想を述べた．そして，知識ベースの中核となる，基本語の意味記述である「基本語意味データベース」Lexeed の構築状況について説明した．2.8 万語の基本語の選定を完了し，その語義文について，国語辞典を初期値として書き換えを行い，書き換えの効果を分かり易さの観点より評価した．今後は，基本語の用例文の作成を進め，心理実験により基本語義を選定し，これを元に意味，統語，語用の側面より知識ベースを構築する予定である．

参考文献

[1] Berners-Lee, T., Hendler, J. and Lassila, O.: The Semantic Web, *Scientific American*, pp. 34–43 (2001).

[2] Schutze, H. and Pedersen, J.: Information retrieval based on word senses, *Fourth Annual Symp. on Document Analysis and Information Retrieval*, pp. 161–175 (1995).

[3] 熊本, 島田, 加藤: 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—, 情処研報, SIG-ICS 115, pp. 9 – 16 (1999).

[4] 池原, 宮崎, 白井, 横尾, 中岩, 小倉, 大山, 林 (編): 日本語語彙大系, 岩波書店 (1997).

[5] Fellbaum, C. *editor*: *WordNet : an electronic lexical database*, The MIT Press (1998).

[6] Proctor, P., *editor*: *Longman Dictionary of Contemporary English*, Longman (1978).

[7] 日本電子化辞書研究所: *EDR Electronic Dictionary Technical Guide*, tr-042 edition (1993).

[8] 天野, 笠原: 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, 16, No. 2C3-09 (2002).

[9] 金杉, 笠原, 稲子, 天野: 単語親密度に基づく基本的語彙の推定, 情処研報, 2002-NL-150, pp. 119–124 (2002).

[10] Bond, 藤田, 橋本, 笠原, 成山, Nichols, 大谷, 田中, 天野: 日本語ツリーバンク「檜」: 言語理解のためのコーパス, 情処研報, 2003-NL-159 (掲載予定).

[11] Bond, 藤田, 橋本, 成山, Nichols., 大谷, 田中: 精細な文法に基づいたツリーバンク「檜」の構築, 情処研報, 2003-NL-159 (掲載予定).

[12] 国立国語研究所: 語彙の研究と教育 (上), 大蔵省印刷局 (1984).

[13] 小池, 小林, 細川, 犬飼 (編): 日本語学キーワード事典, 朝倉書店 (1997).

[14] Kucera, H. and Francis, W. N.: *Computational analysis of present-day American English. Province, RI.*, Brown University Press (1967).

[15] 国立国語研究所: 現代雑誌九十種の用語用字 (1), 秀英出版 (1962).

[16] 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).

[17] 阪本: 読みと作文の心理, 牧書店 (1955).

[18] 井上: 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).

[19] 天野, 近藤: 日本語の語彙特性 第 1 巻 単語親密度, 三省堂 (1999).

[20] 天野, 近藤: 日本語の語彙特性, 第 7 巻 頻度, 三省堂 (2000).

[21] 金田一, 池田 (編): 学研 国語大辞典 第二版, 学習研究社 (1988).

[22] Amano, S. and Kondo, T.: Estimation of mental lexicon size with word familiarity database, *Proc. of Intl. Conf. on Spoken Language Processing*, Vol. 5, pp. 2119 – 2122 (1998).

[23] 鷲坂, 山崎, 廣津, 尾内: 情報検索のための高速日本語形態素解析システム「すもも」, 情全大, No. 2, pp. 59–60 (1999).

[24] 松田, 松本: 品詞タグ付きコーパス作成支援 GUI ツール VisualMorphs, 情処研報, 2000-NL-137, p. 98 (2000).

[25] 高橋, 牛島: 計算機マニュアルのわかりやすさの定量的評価方法, 情報処理学会論文誌, Vol. 32, No. 4 (1991).

[26] 天野, 近藤: 日本語の語彙特性 第 5 巻 文字特性, 三省堂 (2000).

[27] NAIST: CaboCha/南 瓜: Yet Another Japanese Dependency Structure Analyzer, `http://cl.aist-nara.ac.jp/~taku-ku/software/cabocha/`.

[28] 佐久間, 伊集院, 伏見, 辰巳, 田中, 天野, 近藤: 文字単語、音声単語の大規模な心像性評価, 認知神経心理学研究会 (1999).