語種情報を用いた同表記異音語の解消

伝 康晴 * 　　中村 純平 † 　　小木曽 智信 ‡ 　　小椋 秀樹 ‡

* 千葉大学文学部 　† 東京農工大学大学院工学府 　‡ 国立国語研究所研究開発部門

1　はじめに

コーパス中のテキストを単語に分割し、見出し語・品詞・語種などの形態論情報を与えることは、語彙・語法の研究や品詞の分布などから言語資料の特徴を明らかにしようという研究にとって欠かせない。近年では、このような作業に形態素解析システムを利用することが不可欠となり、JUMAN や ChaSen などのフリーソフトが広く用いられている。その一方で、日本語研究への応用を考えた場合、従来の形態素解析システムにはさまざまな不都合・不十分な点がある。

形態素解析は一般に 2 つの処理からなる。1 つは、入力文字列を単位列に分節化することであり、もう 1 つは、個々の単位に品詞を付与することである。しかし、日本語研究ではしばしば、これ以上の情報が必要になる。たとえば、日本語の語には表記の揺れがある。「表わす」という語は「表す」とも「あらわす」とも書かれるし、「桜」という語は「さくら」とも「サクラ」とも書かれる。テキスト中の語の頻度を計数するにあたって、日本語研究では、これらの表記の揺れを同一視する。日本語にはさらに、同表記異語がたくさんある。たとえば、「生物」という表記には、「ナマモノ」と「セイブツ」という意味の異なる語が対応する。日本語研究では、これらの同表記異語を異なる語として区別する必要がある。

これらの異表記同語と同表記異語の問題は本質的には同じ問題である。すなわち、分節化された単位に対して、辞書的見出しを与えるという問題である。我々は、異表記同語を同一視し、同表記異語を区別するこのような見出しを**語彙素**とよぶ。形態素解析システムを日本語研究に応用するためには、語彙素同定処理を扱わなければならない。このような取り組みは、テキスト音声合成の分野で読み付与の問題として扱っている研究が一部みられる (たとえば、長野ほか, 2006) ものの、従来の形態素解析システムの研究ではほとんどなされていない。

本稿では、形態素解析における語彙素同定の問題に対する我々のアプローチについて述べる。とくに、同表記異音語に注目し、形態素解析辞書に記述された語種情報を用いて同表記異音語を解消する手法を提案する。まず、2 節では、我々が開発している形態素解析辞書 UniDic における異表記同語・同表記異語の扱いについて述べる。3 節では、語種情報を用いた同表記異音語の解消のアイデアについて述べる。4 節では、コーパス分析に基づいて本手法が適用できる事例の範囲の見積もりを与え、5 節では、形態素解析システム MeCab を用いて本手法を実装し、その性能を評価する。最後に、6 節で、本手法の意義と今後の方向性を議論する。

2　形態素解析用電子化辞書 UniDic

筆者らは、コーパス日本語学への応用を指向した形態素解析用電子化辞書 UniDic を開発している (伝ほか, 2007)。UniDic の特徴は以下のようにまとめられる。

- 国立国語研究所で規定した「短単位」という揺れがない斉一な単位で設計されている。
- 語彙素・語形・書字形・発音形の階層構造を持ち、表記の揺れや語形の変異にかかわらず同一の見出しを与えることができる。
- アクセントや音変化の情報を付与することができ、音声処理の研究に利用することができる。

本稿と関係が深いのは 2 番目の項目である。UniDic における階層的見出しの例を図 1 に示す（本稿の議論と関係しない発音形の階層は省略する）。**語彙素**は概ね、国語辞典の見出しに相当し、語形・表記・発音の変異を同一視し、意味・文法機能が同一とみなしうるものに同一の見出しを与えたものである。**語形**は、同一の語彙素に対して、形態の違いを区別したものである。たとえば、「ヤハリ【矢張り】」という語彙素には、「ヤハリ」「ヤッパリ」「ヤッパシ」「ヤッパ」など、いくつかの異形態がある。語彙素の下に語形の階層を設けることにより、これらの異形態を区別する。**書字形**は、同一の語形に対して、表記の違いを区別したものであり、テキスト

語彙素	語形	書字形
ヤハリ【矢張り】	ヤハリ	矢張り
		やはり
	ヤッパリ	やっぱり
キョウゾン【共存】	キョウゾン	共存
	キョウソン	共存
カケル【掛ける】	カケル	掛ける
		かける
カケル【欠ける】	カケル	欠ける
		かける
ナマモノ【生物】	ナマモノ	生物
		なま物
セイブツ【生物】	セイブツ	生物

図 1　階層的見出しの例

中に出現する語の実現形を見出しとして掲げたものである。たとえば、「ヤハリ」という語形には、「矢張り」「やはり」という異表記がある。語形の下に書字形の階層を設けることにより、これらの異表記を区別する。

このような階層的見出し設計によって、異表記同語の一部は自然に解消される。すなわち、テキスト中で「矢張り」や「やはり」と分節化された単位に対して、同一の語彙素を与えることができる。その一方で、「掛ける」と「かける」に同一の語彙素を与えるためには、当該の「かける」に対応する語彙素が「カケル【欠ける】」ではなく、「カケル【掛ける】」であることを同定しなければならない。これには、同表記異語の問題が関わっており、辞書設計だけでは解決できない。

図 1 には、2 通りのパターンの同表記異語がある。1 つは、「かける」に「カケル【掛ける】」「カケル【欠ける】」の 2 つの語彙素が対応するというパターンで、語彙素の読みは同じだが、代表的な表記が異なる場合である。これを**同表記同音異語**とよぶ。これには、「ライト【ライト-light】」「ライト【ライト-right】」のように、原語における綴りが異なる場合も含める。もう 1 つは、「生物」に「ナマモノ【生物】」「セイブツ【生物】」の 2 つの語彙素が対応するというパターンで、語彙素の読みが異なる場合である。これを**同表記異音異語**、あるいは、より簡単に**同表記異音語**とよぶ。これには、「良人」に対する「リョウジン【良人】」「オット【夫】」のように、代表的な表記が異なる場合も含める。本稿では、とくに、後者の同表記異音語に焦点をあてる。

なお、図 1 には、「共存」に「キョウゾン」「キョウソン」の 2 つの語形が対応するというパターンもある。これは、語彙素同定のさらに先の段階として、発音形を同定する段階で問題になることであるが、任意性も高いため、本稿では取り上げない。

3　同表記異音語と語種

UniDic には、従来の形態素解析システムの辞書よりもはるかに多くの情報が記述されている。その 1 つに語種情報がある。これを同表記異音語の解消に利用しようというのが本稿で提案する手法のアイデアである。

語種は、和語・漢語・外来語・混種語など、語の出自を分類したものである。同表記異音語の一部は語種の違いに帰着できる。たとえば、「ナマモノ」は和語であり、「セイブツ」は漢語である。語種によって造語力が異なることはよく知られている (たとえば、 野村, 1973)。漢語の短単位は漢語同士で複合しやすいが、和語の短単位は一般に複合しにくい。比較的長い見出し語が多い三省堂『大辞林 第 2 版』で「ナマモノ」「セイブツ」を構成要素に含む語を検索すると、「セイブツ～」は「生物学」「生物時計」など 30 項目、「～セイブツ」は「微生物」「浮遊生物」など 16 項目があるのに対して、「ナマモノ～」「～ナマモノ」は 1 つもない。そこで、隣接要素（「学」「時計」「微」「浮遊」など）が漢語であることがわかれば、「生物」は和語の「ナマモノ」ではなく、漢語の「セイブツ」である可能性が高いことがわかる。

このような語種同士や語種と品詞の連接情報（bi-gram）を利用して、同表記異音語を解消しようというのが本手法の骨子である。

4　コーパス分析

同表記異音語のうち、どれくらいの範囲が語種情報によって解消できるのか見積もるために、コーパスから同表記異音語を持ちうる事例を抽出し、語種情報による曖昧性解消の可否を判断した。

4.1　辞書

語種情報を記載した辞書として UniDic の開発版を用いた。これは、公開中の 1.3.5 版を拡充・整備し、語種情報を追加したものである (小椋ほか, 2008)。各階層の異なり見出しの数は表 1 の通りである。語種情報は語彙素の階層に記述し、和語・漢語・外来語・混種語・固有名・記号・不明の 7 種類を区別した。

表 1　辞書の異なり見出し数

階層	見出し数
語彙素	107,414
語形	111,744
書字形	153,154

表 2　対象コーパス（短単位数）

コーパス	総単位数	学習用	評価用
RWCP	899,350	802,959	96,391
CSJ	458,759	413,167	45,592
白書	228,219	113,688	114,531
計	1,586,328	1,329,814	256,514

4.2　コーパス

分析対象コーパスとして、『RWCP テキストコーパス』(RWCP, 1998)、『日本語話し言葉コーパス（CSJ）』(前川, 2004)、および、『現代日本語書き言葉均衡コーパス』(山崎, 2007) に含まれる白書データ（人手修正済み）を用いた。コーパスの規模を表 2「総単位数」に示す。

4.3　辞書中の同表記異語の分類

辞書中で、書字形および品詞・活用型が同一の項目で、2 つ以上の語彙素に対応するものを同表記異語とした。153,154 項目の書字形のうち、同表記異語は 7,472 項目（4.9%）（同音異語: 1,214 項目（0.8%）、異音異語: 6,258 項目（4.1%））あった。これらを、語彙素の語種に応じて、以下のように分類した。

同:　すべての語彙素の語種が同一で固有名以外

固:　すべての語彙素の語種が固有名

異:　すべての語彙素の語種が互いに異なる

混:　一部の語彙素の語種が同一で、他は異なる

表 3 に、同音異語と異音異語ごとに、これらの内訳を示す。同表記同音異語では、大半が同一の語種を持つのに対し、同表記異音語では、1/3（33.1%）が語種情報によって語彙素を区別できることがわかる。

4.4　コーパス中の同表記異語の分類

次に、コーパスから同表記異語を抽出し、上記と同様の分類を行なった（表 4）。1,586,328 単位のうち、同表記異語は 297,846 単位（18.8%）（同音異語: 45,010 単位（2.8%）、異音異語: 252,836 単位（15.9%））あった。同表記異音語の割合は、白書データが 45.0% ともっとも高く、CSJ が 22.2%、RWCP が 5.4% だった。これらのうちの 6 割以上（62.9%）が語種情報によって語彙素を区別でき、部分的に語彙素を絞り込める「混」を含めるとその割合は 3/4（76.5%）にも達する。語種情報で同表記異音語を解消できる割合は、CSJ が 65.5% ともっとも高く、RWCP が 63.1%、白書が 60.2% だった。

表 3　同表記異語の分類（辞書）

分類	同音異語（比率）		異音異語（比率）	
同	1,112	(91.6%)	2,302	(36.8%)
固	24	(2.0%)	1,514	(24.2%)
異	72	(5.9%)	2,071	(33.1%)
混	6	(0.5%)	371	(5.9%)
計	1,214	(100.0%)	6,258	(100.0%)

表 4　同表記異語の分類（コーパス）

分類	同音異語（比率）		異音異語（比率）	
同	44,734	(99.4%)	55,222	(21.8%)
固	10	(0.0%)	4,162	(1.6%)
異	205	(0.5%)	159,005	(62.9%)
混	61	(0.1%)	34,447	(13.6%)
計	45,010	(100.0%)	252,836	(100.0%)

5　実験

コーパス分析に基づく見積りから、同表記異音語の 6 割以上が、語種情報によって解消できる可能性があることがわかった。そこで、このアイデアを形態素解析システムに実装し、その性能を評価した。

5.1　形態素解析システム

公開中の UniDic は、隠れマルコフモデル（HMM）に基づく形態素解析システム ChaSen 上で動作する。しかし、HMM で語種情報を利用するためには、品詞と語種の組に対する bi-gram を学習する必要があり、データスパースネスを引き起こす可能性が高い。そこで、多数の素性を柔軟に取り入れることができる条件付き確率場（CRF）に基づく形態素解析システム MeCab (Kudo et al., 2004) を用いて実装した。素性として、品詞・活用型・活用形に加え、書字形（出現形と基本形）・語彙素・語種の uni-gram/bi-gram 素性を用いた。比較のため、語種情報を用いない ChaSen および MeCab でも実装した。

5.2　学習・評価コーパス

表 2 のコーパスの一部を学習用（RWCP・CSJ は約 90%、白書データは約 50%）に、残りを評価用に用いた（表 2「学習用」「評価用」）。ただし、未知語処理の性能の影響を避けるため、語彙はコーパス全体から取得した。

表 5　分節化・品詞付与・語彙素同定の性能（F 値）

	RWCP			CSJ			白書		
	分節化	品詞	語彙素	分節化	品詞	語彙素	分節化	品詞	語彙素
ChaSen（語種無）	99.27	97.89	97.12	99.28	97.39	96.57	99.64	98.95	98.69
MeCab（語種無）	99.64	98.77	97.97	99.64	98.18	97.35	99.86	99.26	98.93
MeCab（語種有）	99.65	98.80	98.51	99.65	98.28	97.78	99.87	99.30	99.16

表 6　同表記異音語の分類ごとの正解率

分類	RWCP	CSJ	白書
同	99.26%	99.33%	99.86%
固	87.22%	99.90%	99.69%
異	99.40%	99.86%	99.78%
混	98.47%	99.73%	99.43%
全体	99.18%	99.74%	99.75%

5.3　結果

語種情報を用いない ChaSen/MeCab、および、本稿で提案した語種情報を用いた MeCab による分節化・品詞付与・語彙素同定の F 値（再現率と精度の調和平均）を表 5 に示す。語種情報を用いることによって、語彙素同定の性能が、ChaSen と比べて 0.47〜1.39%、MeCab（語種無）と比べて 0.23〜0.54% 改善している。

表 4 の同表記異音語の分類ごとの正解率を表 6 に挙げる。RWCP の固有名の曖昧性を除いて、いずれもきわめて高い正解率を得ている。語彙素の語種がすべて同一の場合でも高い正解率となったのは、語彙素の uni-gram 素性だけで解消できる例が多かったためと思われる。

6　考察

語種情報を用いることで、語彙素同定の性能がいずれのコーパスにおいても改善した。これは、語種によって曖昧性の解消が可能な同表記異音語の解析精度が向上したためと考えられる。

長野ほか (2006) は、品詞と読みを組み合わせた *n*-gram を用いることで、同表記異音語の解消に取り組んでいるが、そこでの分節化・読み付与の正解率は 97.69% であり、本手法の性能はそれよりも高い（ただし、長野ほか (2006) の実験は未知語を含む）。これは、語種情報というより有益な情報を利用したことと、CRF というより洗練された統計モデルを採用したことによる。このことは、言語資源の開発における精緻な言語学的情報の記述と、自然言語処理技術における洗練された手法の利用とを融合することで、最適な性能を引き出せる可能性を示唆している。

その一方で、本手法にはまだ改善の余地がある。本手法では、MeCab を用いて、分節化・品詞付与・語彙素同定を同時にモデル化した。しかし、語彙素の曖昧性を保留したままで分節化・品詞付与を行ない、後処理によって、語彙素同定処理を行なうという方略も考えられる。この場合には、bi-gram よりも広い範囲の情報を利用することも可能であり、より性能向上に貢献する要因を発見できる可能性がある。このような方略について、今後検討する予定である。

参考文献

伝康晴・小木曽智信・小椋秀樹・山田篤・峯松信明・内元清貴・小磯花絵. (2007). コーパス日本語学のための言語資源: 形態素解析用電子化辞書の開発とその応用. 日本語科学, *22*, 101–123.

Kudo, T., Yamamoto, K., & Matsumoto, Y. (2004). Applying conditional random fields to Japanese morphological analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing* (pp. 230–237). Barcelona, Spain.

前川喜久雄. (2004). 『日本語話し言葉コーパス』の概要. 日本語科学, *15*, 111–133.

長野徹・森信介・西村雅史. (2006). *N*-gram モデルを用いた音声合成のための読みおよびアクセントの同時推定. 情報処理学会論文誌, *47*, 1793–1801.

野村雅昭. (1973). 複次結合語の構造. 国立国語研究所報告 49: 電子計算機による国語研究 5 (pp. 72–93). 東京: 国立国語研究所.

小椋秀樹・小木曽智信・原裕・小磯花絵・冨士池優美. (2008). 形態素解析用辞書 UniDic への語種情報の実装と政府刊行白書の語種比率の分析. 言語処理学会第 14 回年次大会発表論文集.

新情報処理開発機構 (RWCP) テキスト・サブ・ワーキンググループ. (1998). 研究開発用知的資源: タグ付きテキストコーパス報告書.

山崎誠. (2007).「現代日本語書き言葉均衡コーパス」の基本設計について. 特定領域「日本語コーパス」平成 18 年度公開ワークショップ（研究成果報告会）予稿集 (pp. 127–136).