形態素解析辞書 UniDic における同語異語判別について

小椋秀樹　　原裕　　小木曽智信　　小磯花絵　　宮内佐夜香

人間文化研究機構 国立国語研究所

1. はじめに

国立国語研究所が中心となって構築を進めている『現代日本語書き言葉均衡コーパス』（Balanced Corpus of Contemporary Written Japanese, 以下 BCCWJ と略す。）は，国語学・情報工学等の幅広い分野での活用を目指したコーパスであり，そのために様々な研究用付加情報が付与される[1][2]。

筆者らは，この研究用付加情報のうち形態論情報の付与を行っている。BCCWJ では長単位・短単位という 2 種類の言語単位を採用し，長短両方の単位に対して見出し・品詞・語種等の情報を付与する。この形態論情報付与作業は，自動解析システムにより行い，短単位については，解析器に MeCab，解析用辞書に伝康晴氏（千葉大）や筆者らが開発を進めている UniDic を使用する[3]。(なお長単位解析には，短単位解析結果から長単位を自動構成する解析器を使用する[4]。)

UniDic では，表記や語形の違いにかかわらず，同じ語であれば，同一の見出しを与えるという方針の下，語の登録を行っている。しかし従来の UniDic では，語形の面から同じ語とするか否かを判断する基準は整備できていたが，意味の面から同じ語とするか否かを判断する基準には明確でない部分があり，見出しの与え方に不整合が生じていた。

そこで，筆者らは，国語辞典の見出しやブランチの立て方・漢字表記に関する注記，BCCWJ の頻度情報等を基に，UniDic に登録した和語動詞から同語異語判別に着手した。本稿では，この同語異語判別の方針を紹介するとともに，これまでに判別を行った語を幾つか取り上げ，同語異語判別の実例を示す。

2. 同語異語判別の必要性

UniDic では，表記や語形の違いにかかわらず，同じ語であれば，同一の見出しを与えるという方針を取り，語を階層化した形で登録している。この階層の最上位を語彙素（国語辞典の見出しに相当）と呼んでおり，この語彙素の下に語形（語形の違いを区別する層），更に語形の下に書字形（表記の違いを区別する層）という階層を設けている（図 1）。

語を，このような階層構造で登録した辞書を用いて形態素解析を行うことによって，例えば，ある語について，どのような語形の変異や表記のゆれが，どの程度あるのかという情報を容易に得ることができるなど，日本語研究の可能性が格段に広がることが期待される。

語彙素	語形	書字形
矢張り	ヤハリ	やはり
		矢張り
	ヤッパリ	やっぱり
		矢っ張り
	ヤッパ	やっぱ

図1 UniDicの階層構造の例

そのためには，ある語彙素の下に属する語形や書字形としてどの範囲までを含めるのかという基準，つまり同語異語判別の基準を立て，その基準に従って UniDic に語を登録する必要がある。この同語異語判別基準が明確になっていないと，本来，語形や表記のゆれとしてとらえるべきものを異なる語として処理したり，その反対に，異なる語としてとらえるべきものを同じ語の語形や表記のゆれとして処理したりするといった誤りが起こる。

UniDic における同語異語判別基準としては，語形の違いに関する基準を既に取りまとめている[5]。この基準では，清濁の差異及び濁音と半濁音との差異，語末長音の短呼形と元の形との差異など 10 種類の語形の差異を挙げ，このような差異を持つ任意の 2 語は，語源が同じで意味の違いを生じていない限り，一つの語彙素にまとめると規定している。例えば，語彙素「レンチュウ【連中】」の下には「レンチュウ」と「レンジュウ」とを，語彙素「センセイ【先生】」の下には「センセイ」と「センセ」とを登録する。

しかし，主として意味にかかわる問題については，これまで基準が未整備であった。動詞〈オサマル〉を例にすると，「収まる」と「治まる」とを一つの語の表記のゆれと見なすか，別語と見なすかということについて基準を整備できていなかったのである。

ここで動詞〈オサマル〉について，『岩波国語辞典』第 6 版（以下，『岩波』），『国語大辞典』（以下，『国語大』），『大辞林』第 2 版（以下，『大辞林』），『広辞苑』第 5 版（以下，『広辞苑』）における見出しの立て方を見ると，以下のようになっている。

岩　波　　収まる・納まる・修まる・治まる
国語大　　収まる・納まる・修まる・治まる
大辞林　収まる・納まる／修まる／治まる
広辞苑　　収まる・納まる・修まる・治まる

『岩波』『国語大』『広辞苑』は，一つの見出しにまとめるが，『大辞林』は「修まる」と「治まる」とを別立てにし，見出しを三つ立てる。

したがって，「収まる」「治まる」について言えば，『岩波』『国語大』『広辞苑』に従うと同じ語として扱う（同じ動詞〈オサマル〉の表記のゆれとして扱う）ことになるが，『大辞林』に従うと異なる語として扱う（語形は同じだが意味や表記の異なる別語として扱う）ことになる。このように見出しの立て方が辞書によって異なることからも，二つの語を同じ語と見なすか異なる語と見なすかには，様々な考え方があり，非常に判断の難しい問題と言うことができる。こうした問題の難しさから，従来の UniDic では，同語異語判別基準のうち意味の面に関する基準が整備できておらず，語によって語彙素の立て方に不整合が生じていた。

そこで，筆者らは，意味の面にかかわる同語異語判別の方針を立て，それに沿って UniDic の既登録語の語彙素を整理することとした。

3. UniDicにおける同語異語判別

3.1 基本的な方針

同語異語判別の先行事例としては，国立国語研究所がこれまでに行ってきた語彙調査が挙げられる。例えば，現代雑誌 90 種調査（1956 年刊行の雑誌 90 誌を対象とした用語用字調査）では，同語異語判別の結果を「判別実例一覧表」に示している[6]。

UniDic における同語異語判別でも現代雑誌 90 種調査等を参考にすることが考えられる。しかし UniDic の同語異語判別は，自動形態素解析を前提としたものであり，すべて人手で判別することを前提とした語彙調査をそのまま参考にするのは難しい。

自動形態素解析を前提とした同語異語判別で留意しなければならないのは，同じ書字形（表記）かつ同じ品詞でありながら，異なる語彙素（見出し）となるような語（同表記異語）をできる限り少なくするということである。

例えば，動詞〈オサマル〉において，漢字表記「収まる」「納まる」よりも平仮名表記が多く用いられており，しかも平仮名表記が「収まる」の意と「納まる」の意の両方で用いられ，用例数も両者で拮抗しているとする。このような場合，「収まる」と「納まる」とを異なる語彙素にすると，同表記（平仮名表記）の異語が多くなり，解析精度の低下を招く。

また，この例のように漢字表記よりも平仮名表記の方が多いということは，意味の差が微妙であるため漢字の書き分けがうまくできず，その結果，平仮名表記が選択されている可能性がある。このような人手でさえ区別の難しい，微妙な意味の差異によって語彙素を複数立てることは，形態素解析辞書の見出しの設定としてふさわしくない。

以上のことを踏まえ，UniDic における同語異語判別に関して，筆者らは次に挙げる二つの基本方針を立てた。

1. 同表記異語を生じさせるような語彙素の立て方はできる限り行わない。
2. 複数の語彙素に分ける場合は，明確な基準・理由をもってし，人手で正確に区別できないような語彙素の分割は行わない。

3.2 対象

現在公開している UniDic-1.3.12 には語彙素レベルで 15.7 万語が登録されている。そこで，同語異語判別の対象をひとまず絞り込むこととした。具体的には，UniDic に登録した和語の単純動詞のうち，現代雑誌 90 種調査の「判別実例一覧表」に掲げられている語（200 語）を中心に判別作業に着手した。ここで単純動詞を取り上げることとしたのは，意味による漢字の書き分けで問題となる語が多いためである。

なお，対象とした単純動詞の名詞形や対象とした単純動詞を語構成要素に持つ複合動詞については，単純動詞の同語異語判別が終わった後に，単純動詞の判別結果等を参考にして同語異語判別を行う。

3.3 方法

同語異語判別を行う際の検討資料として，以下の二つを用いた。

(1) 構築中の BCCWJ（約 5,500 万語）から取得した各書字形の頻度情報（表記別頻度）及び用例

基本方針 1 を踏まえ，現在 UniDic に登録されている語彙素がどのような書字形を持ち，その書字形がコーパス中に何例現れるのかを把握する必要があると考え，検討資料とした。できる限り同表記異語を生じさせないということから，特に平仮名表記例の頻度・使用状況に注目した。また，コーパス全体の頻度・用例等を参照するのに加えて，形態素解析システムの学習用コーパスとして作成している人手修正済みデータ（コアデータ：約 70 万語[7]）における頻度・用例も適宜参照した。

(2)『岩波』『国語大』『大辞林』『広辞苑』の 4 種の国語辞典

基本方針 2 にも掲げたように明確な基準・理由をもって同語異語判別を行うために国語辞典における見出しやブランチの立て方を参照することとした。ただし，先に見たように辞書によって見出しの立て

方に違いがあるため複数の辞書を参照することとし，上記の代表的な中型・小型辞書 4 種類を用いた。

検討資料を基に同語か否かを判別するに当たって，次の八つの観点を設けた。

《辞書の記述に基づく観点》

a. すべての辞書で見出しの立て方が一致し，それが一般的な漢字の書き分けの意識とも対応している場合，辞書の見出しに基づいて語彙素を立てる方向で検討する。
b. 各辞書の漢字表記に関する注記に違いがある場合，原則として一つの語彙素にまとめる。
c. 『国語大』に漢字表記に関する注記があり，他の辞書でもほぼ同一の書き分けを示している場合，『国語大』の注記に基づいて語彙素を立てる方向で検討する。
d. 『国語大』に漢字表記に関する注記があっても，他の辞書で示された書き分けと異なる場合，一つの語彙素にまとめることを優先する。
※『国語大』は漢字表記に関する注記を示すことに慎重な姿勢を取っていると思われるため，『国語大』の注記を重視する観点を設けた。

《動詞の性質等に基づく観点》

e. 自動詞と他動詞とで同形の語は，原則として異なる語彙素とする。
f. 移動動詞等，自然言語処理における応用面に深くかかわると思われる語の語彙素の統合については慎重に判断する。

《頻度・実例に基づく観点》

g. 漢字表記の頻度よりも仮名表記の頻度が非常に高い場合，一つの語彙素にまとめることを優先する。
h. 前接・後接語に違いがある場合，又は漢字の書き分けに混乱が見られない場合は，複数の語彙素に分ける方向で検討する。

上記八つを基準又は規則と言わずに観点と呼んでいるのは，意味にかかわる問題であるため，規則のように適用の優先順位等を決められないからである。

筆者らは，表記別頻度表・用例・4 種の国語辞典を基に八つの観点から同語とするか否か，1 語ずつ総合的に判断していった。現在までに 91 語について判別作業を終えた。

4. 同語異語判別の実例

ここでは，既に同語異語判別を行った動詞〈アウ〉〈アズカル〉〈オサマル〉を取り上げ，判別の実例を示す。

(1) アウ

UniDic-1.3.12 では，「会う」「合う」「遭う」の三つを語彙素とし，「逢う」は「会う」の書字形に，「遇う」は「遭う」の書字形に登録している。

辞書を見ると，『大辞林』が「合う」を別立てにしているが，他の辞書は一つの見出しにまとめている。ただし『国語大』『広辞苑』はブランチで「合」「会・逢・遭」と接尾的用法の「合」の三つに分けている。『国語大』『大辞林』『広辞苑』は，基本的に「合」と「会・逢・遭」とを分ける方針を取っていると考えられる。

表1 〈アウ〉の表記別頻度

あう	会う	合う	逢う
2,057	10,057	5,737	696
遭う	遇う		
89	913		

表記別頻度（表 1）を見ると，平仮名表記が 2,057 例と多い。この大半は「合う」の意の例と「遭う」の意の例であり，用例数はほぼ拮抗している。コアデータに限定して見ると，平仮名表記は 23 例あり，内訳は「合う」の意が 13 例，「遭う」の意が 8 例，「会う」の意が 2 例である。「合う」の意の例のうち 8 例は接尾的用法で前接語が動詞連用形である。「遭う」の意の例の前接語はすべて格助詞「に」である。平仮名表記例の大半を占める「合う」の意の例と「遭う」の意の例は，前接語に差異が認められる。

「会う」については，漢字表記も含めて見ると，格助詞「に」が前接語となることが多い点で「遭う」と類似しており，辞書でも「会う」と「遭う」とは同じ見出しにまとめられている。

以上のことから，「合う」「会う」の二つの語彙素とし，「逢う」「遭う」「遇う」は「会う」の書字形とした（図 2）。

語彙素	語形	書字形
合う	アウ	あう
		合う
会う	アウ	あう
		会う
		逢う
		遭う
		遇う

図2 〈アウ〉の語彙素・語形・書字形

(2) アズカル

UniDic-1.3.12 では，「与る」「預かる」の二つを語彙素としていた。

辞書を見ると，『大辞林』『広辞苑』は「与る」「預かる」の二つを見出しに立てる。『岩波』『国語大』は一つにまとめるが，その中を自動詞と他動詞の 2 項目に分け，自動詞に「与」，他動詞に「預」と注記する。

辞書の記述にあるとおり「与る」と「預かる」は自動詞・他動詞の違いがあり，意味についても「預かる」が金品等を手元に置き保管するといった意味

を表すのに対し，「与る」が関与するといった意味を表すというように違いがある。

表記別頻度（表 2）を見ると，平仮表記が「預かる」に次いで 489 例と多い。内訳は「与る」の意が 275 例，「預かる」の意が 214 例で，「与る」の意が約 3 割多くなっている。また，前接語を見ると，自他の違いを反映して，「与る」の意の例は格助詞「に」が来ることが多く，この点が「預かる」の意の例との違いとなっている。

表2 〈アズカル〉の表記別頻度

あずかる	与かる	与る	預かる	預る
489	18	63	848	145

意味が異なっていること，自他の区別があることなどから，UniDic-1.3.12 のまま「預かる」「与る」の二つの語彙素とした。

(3) オサマル

UniDic-1.3.12 では，「収まる」「治まる」「納まる」「修まる」の四つを語彙素としていた。

辞書の見出しの立て方は 2 節に示したとおりである。『国語大』は見出しの中を大きく「物事が安定した状態になる。ととのった状態になる。」と「物がきちんと中にはいる。また、物事が終わりになる。」の二つに分け，前者には「(治・修・収)」，後者には「(収・納)」と注記する。このように，いずれの項目でも複数の漢字が注記として付されていること，「収」については両方のブランチに漢字注記として付されていることから，明確な書き分けが示されているとは言えない。漢字注記に関しては，他の辞書も同様で，明確な書き分けの基準は示されていない。

表3 〈オサマル〉の表記別頻度

おさまる	修まる	収まる	収る	治まる
844	5	649	9	243
納まる	納る	蔵まる		
181	16	1		

表記別頻度（表 3）を見ると，平仮名表記が最も多くなっており，漢字の書き分けが相当難しいことがうかがわれる。

語彙素	語形	書字形
収まる	オサマル	おさまる
		修まる
		収まる
		収る
		治まる
		納まる
		納る
		蔵まる

図3 〈オサマル〉の語彙素・語形・書字形

以上のことから，「収まる」のみを語彙素として立て，「治まる」「納まる」「修まる」等はその書字形とした（図 3）。

5. 終わりに

本稿では，UniDic における意味の面からの同語異語判別の方針等について述べるとともに，動詞〈アウ〉〈アズカル〉〈オサマル〉を取り上げ，判別の実例を示した。

現在は，UniDic に登録した和語の単純動詞のうち，現代雑誌 90 種調査の「判別実例一覧表」に掲げられている語を中心に判別作業を進めている。今後，単純動詞の名詞形や複合動詞（その名詞形も含む。）の判別作業を進めていくことにしているが，更に他の品詞についても，範囲を区切りつつ継続して作業を進め，UniDic の語彙素について一通り整理していきたいと考えている。

また今後，同語異語判別結果を反映した UniDic を用いて解析を行い，精度の低下が見られないかといったことを確認しながら作業を進めていく必要がある。その際，現在の資料・観点で十分か否かも検討し，場合によっては，より望ましい同語異語判別が行えるよう資料・観点等を加えていくことも必要である。

参 考 文 献

[1] 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』,127-136.

[2] 前川喜久雄(2008)「KOTONOHA『現代日本語書き言葉均衡コーパス』の開発」『日本語の研究』4-1,82-95.

[3] 伝康晴ほか（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,国書刊行会.

[4] Uchimoto, K., & Isahara, H.（2007）. Morphological annotation of a large spontaneous speech corpus in Japanese, *Proceedings of IJCAI*, 1731-1737.

[5] 小椋秀樹ほか(2009)国立国語研究所内部報告書『『現代日本語書き言葉均衡コーパス』形態論情報規程集改定版』(LR-CCG-08-03),141-160.

[6] 国立国語研究所(1964)『現代雑誌九十種の用語用字(3)』,294-330.

[7] 小椋秀樹ほか(2009)「『現代日本語書き言葉均衡コーパス』における形態論情報付与作業の進捗状況」『特定領域「日本語コーパス」平成 20 年度公開ワークショップ（研究成果報告会）予稿集』,57-64.

関連URL

UniDic：http://download.unidic.org/
MeCab：http://mecab.sourceforge.net/

付記 本研究は，文部科学省科研費特定領域研究「日本語コーパス」による補助を得た。