見出し語の時代情報を付与した電子化辞書の構築

鴻野知暁　小木曽智信
{tkouno, togiso}@ninjal.ac.jp
人間文化研究機構　国立国語研究所

1. はじめに

国立国語研究所では『現代日本語書き言葉均衡コーパス（BCCWJ）』の完成を受けて、「日本語歴史コーパス」（通時コーパス）の構築が進められている（近藤 2012）。通時コーパスの構築にあたって、上代から近代に及ぶ広範な時代の資料に対して形態論情報を付与する必要がある。古典語を対象とした形態素解析を行うための辞書を開発するには、辞書のデータベースを整備する必要がある。発表者らは現代語用の形態素解析辞書 UniDic に古典語の見出しを追加して、辞書データを拡充してきた（小木曽 2013, 小木曽・小町・松本 2013）。さまざまな古典資料を出典とする見出し語を追加したことにより、特定の時代でしか使用しない見出し語が増え、形態素解析辞書の機械学習にかかる時間が大幅に増大するという問題が生じていた。また、特定の時代の資料にとって不要な見出し語が辞書に存在することは、解析時の曖昧性の増大にも通じ、機械解析の精度の低下を招く恐れがある。解析対象とする古典資料の性質に応じて、解析に用いる見出し語を絞り込む工夫が求められるのである。

2. 見出し語の制限

形態素解析に使われる見出し語を制限する方法として、次の二つが考えられる。一つは、それがいかなる文体の資料に現れるかという位相の情報を、各見出し語に付与するものである。これにより、和文で現れる語や、和漢混淆文に現れる語などを区別することができる。もう一つは、それがいかなる時代の資料で使用されるかという時代情報を、見出し語ごとに付与するものである。時代情報は、その見出し語がどの時代から使われるか（開始時代）と、どの時代まで使われるか（終了時代）との二つの情報の組によって指定される。本稿は主として、この時代情報によって使用項目を制限する方法について述べる。

3. 時代区分

日本語の歴史は、日本語史研究においておおむね図 1 のように時代区分される[1]。電子化辞書での使用年代もこれにあわせて、上代・中古・中世・近世・近代・現代の 6 分類を考える。開始時代・終了時代の属性値はこれらのうちのいずれかの値をとる。

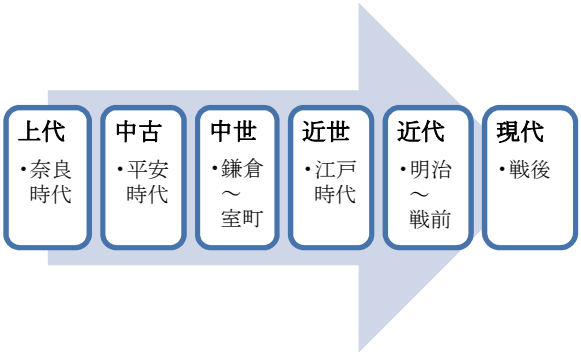

図 1　日本語の歴史的区分

4. 見出し語への時代情報の付与

4.1 見出し語の階層

UniDic では、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ように、見出し語を語彙素・語形・書字形・発音形の 4 つのレベルで階層的に管理している（伝ほか 2007）。

図 2　UniDic 見出し語の階層構造

Copyright(C) 2014 The Association for Natural Language Processing.
All Rights Reserved.

図 3 は「何処（イズコ）」の例であるが、代表となる辞書見出しとして語彙素「何処」(語彙素読みイズコ)を立て、その下に異なる語形として「イズコ」、「イドコ」を配している。表記の違いが書字形として各語形の下に位置づけられる。

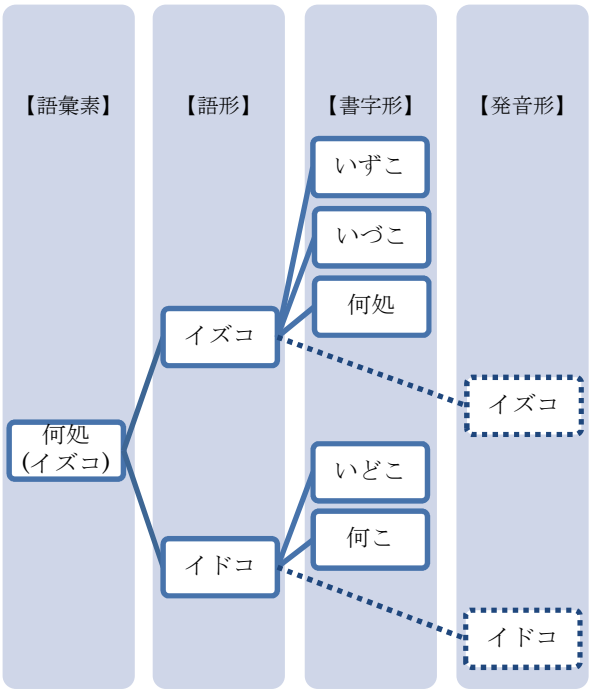

図 3　「何処」（イズコ）の階層構造

4.2　時代情報の整合性

見出し語の使用年代は、語彙素のレベルで決まるものから、書字形のレベルで決まるものまで様々であるため、各レベルで使用年代を付与できるようにしている。ただし、見出し語全体の時代情報の整合性を保つため、子見出しはデフォルトでは親見出しの時代情報を継承し、より狭い時代範囲に絞り込むことのみを可能にしている。

5.　時代情報の入力の実例

5.1　語彙素レベル

語彙素レベルの時代情報は、人手修正を経た学習用コーパスの実際の用例、また、『日本国語大辞典』などにあがっている用例を参照して入力される。

中古資料の『法華百座聞書抄』を時代情報が未入力の辞書で解析すると、図 6 のようなエラーが生じている。出現形が平仮名表記であり、しかも品詞が同じであるために、語彙素判別を誤っている。しかし、「蜀黍」が近世以降に使われる語であることを図 5 のように時代情報として指定すれば[2]、当該の中古資料の解析にこの語を使用しないように制限できる。

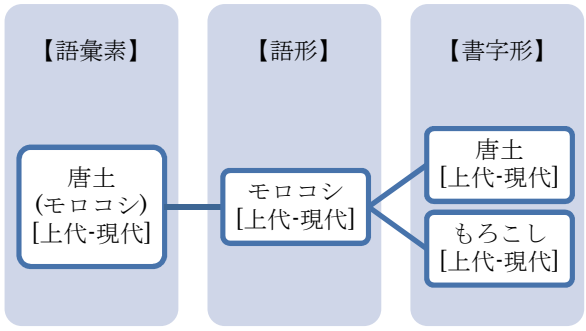

図 4　「唐土」（モロコシ）の時代情報

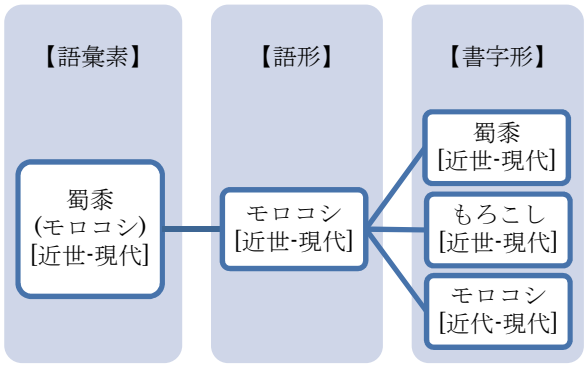

図 5　「蜀黍」（モロコシ）の時代情報

	前文脈	キー	後文脈	語彙素読み	語彙素	出現発音形	品詞	解析活用型	活用形
（誤）	昔、	もろこし	に、道をまかりけるもの、	モロコシ	蜀黍	モロコシ	名詞-普通名詞-一般		
（正）	昔、	もろこし	に、道をまかりけるもの、	モロコシ	唐土	モロコシ	名詞-普通名詞-一般		

図 6　語彙素レベルのエラー例（『法華百座聞書抄』）

Copyright(C) 2014 The Association for Natural Language Processing.
All Rights Reserved.

5.2 語形レベル

語形レベルは語形の揺れ、変異を区別するものであり、語形（音韻）変化をカバーする。また、活用型の属性を情報として持ち、活用型を区別する。音韻・活用型の情報は、『日本国語大辞典』や『日葡辞書』などの辞典類の記述を元にして入力される。

(a) 語形（音韻）変化

「ダレ（誰）」は古くは清音の「タレ」であった（現代でも文語調の文章ではタレの読みのことがある）。通時コーパスでは、「誰」と漢字表記で出現した場合、発音形が「ダレ」かそれとも「タレ」なのかという曖昧性が一般に生じる。しかし、図 7 のように UniDic の見出し語に時代情報を入力すれば、上代から中世の資料においては「ダレ」という読みは排除できる。

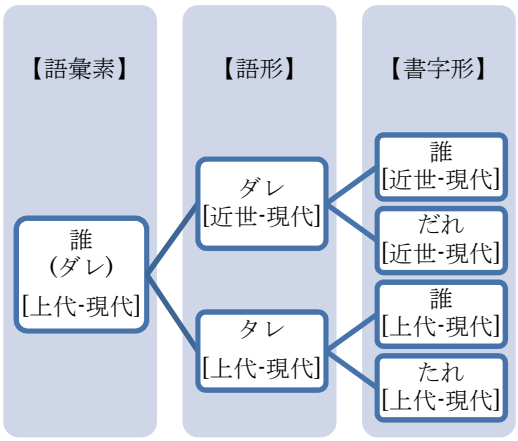

図 7 「誰」（ダレ）の時代情報

(b) 活用型

「文語形容詞-シク」や「文語四段」などの文語活用の動詞・形容詞の活用型は、時代情報を「近代」以前に限定した。逆に「形容詞」や「五段」などの口語活用の動詞・形容詞の活用型は、「近世」以降に限った。

活用語である動詞は、時代とともに活用型が変化する場合がある。この史的変遷の有り様は、一つの語彙素に、異なった活用型を持つ複数の語形をぶらさげ、それらに適切な時代情報を付すことによって表される。

「くしゃみをする」という意味の動詞「嚏ひる（ハナヒル）」は、上代では上二段活用として使われていたが、中古以降は上一段となった。

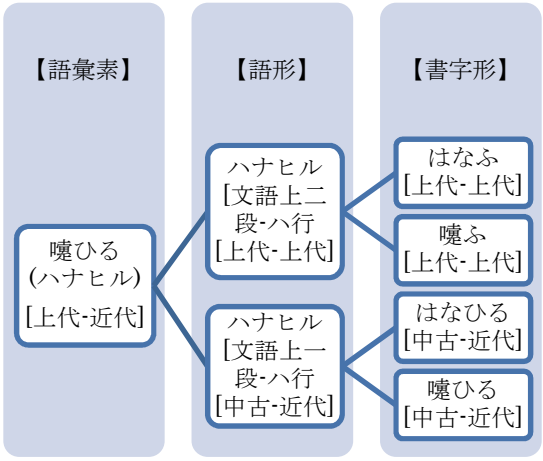

図 8 「嚏ひる」（ハナヒル）の時代情報

時代情報が未入力の辞書での中古作品『枕草子』の誤解析例を図 9 に示す。「嚏ひる」の連用形は上二段と上一段で同形となり、エラーでは活用型の情報という一点のみが異なっている。しかも「嚏ひる」は出現度数がきわめて低く、このような誤りは人手による修正でも見落とされやすい。しかし、図 8 のように時代情報を辞書側で管理すれば、機械解析の時点で正しい語形を選択することができる。

5.3 書字形レベル

既存の辞典類には書字形（表記）に関する言及は少ない。書字形レベルの時代情報は、学習用コーパスに現れた用例や、表記に関する研究書を参考にして入力される。

書字形のレベルで使用年代を制限したものは、仮名遣いや漢字の異体字に関するものである。現代では用いられない旧字形の「讀書」（読書）、「辯明」（弁明）、「聯絡」（連絡）のようなものは「近代」までとし、逆に「キレる」「イタい」のような現代語特有の表記は「現代」専用（[現代-現代]）としている。

	前文脈	キー	後文脈	語彙素読み	語彙素	出現発音形	品詞	解析活用型	活用形
（誤）	男主ならでは高く	鼻ひ	たる、いとにくし。	ハナヒル	嚏ひる	ハナヒ	動詞-一般	文語上二段-ハ行	連用形-一般
（正）	男主ならでは高く	鼻ひ	たる、いとにくし。	ハナヒル	嚏ひる	ハナヒ	動詞-一般	文語上一段-ハ行	連用形-一般

図 9 語形レベルのエラー例（『枕草子』）

Copyright(C) 2014 The Association for Natural Language Processing.
All Rights Reserved.

中世から近世の資料では、語頭のワ（ア）行音をハ行の仮名で書く独特な表記が見られる。中世末期から近世の口語資料である狂言では、「はるひ」（悪い）、「ひたす」（致す）などといった例がある[3]。他の時代にはこのような表記は現れないので（「ひたす」は普通「浸す」の意）、時代情報でこれらを[中世-近世]と制限する必要がある。

さらに、書字形を活用展開した形である出現書字形にも、時代情報を付与した。たとえば、形容詞「早い」の書字形「はやい」を連用形ウ音便として展開した場合、一般的な表記形は「はよう」であるが、近世資料では「はやふ」という表記で現れることがある。このようなものには[近世-近世]という時代情報が与えられている。

6. 形態素解析辞書の見出し語数の削減効果

時代情報を入力することにより各時代の見出し語数がどれだけ制限されるかを示したものが表 1 である[4]。書字形の活用展開後の数を見ると、時代情報の制限がない場合と比べ、特に中古・中世・現代の時代で見出し語数が相当に減少している。これは中世までは口語活用の活用型が、現代では文語活用の活用型が、それぞれ使用されないためである。

表 1　時代情報の制限を付けた各時代の見出し語数

	語彙素	語形	書字形	活用展開済み
全語彙	230787	258585	420989	1375692
中古	209615	221105	299936	718554
中世	209734	221425	301245	720046
近世	209864	235910	338453	1319427
近代	212692	239443	342142	1347219
現代	216313	231531	313215	895420

このような見出し語数の削減が、中古語コーパスの学習・解析にどのように影響したかを最後に述べる。中古の見出し語と MeCab 0.993 を用いて、CHJ 平安時代編のコーパスで学習を行った結果、見出し語数の削減効果は表 2 のように、辞書サイズの縮小として確認できた。

表 2　辞書サイズの比較

	中古語彙のみ	全語彙
配布用ソース辞書（Lex.csv）	145MB	295MB
解析用のバイナリ辞書全体	253MB	686MB

また、機械学習にかかる時間は 8 時間 9 分（全語彙）から 2 時間 21 分（中古語彙のみ）へと大幅に短縮された。解析精度については、表 3 に示すとおり、ほぼ同じ結果が得られた。

表 3　解析精度の比較

	中古語彙のみ	全語彙
境界認定	0.9949	0.9948
品詞認定	0.9824	0.9824
語彙素認定	0.9733	0.9735
発音形認定	0.9709	0.9709

7. おわりに

電子化辞書の見出し語に時代情報を付与することにより、各時代で使用する見出し語を制限でき、その結果として、機械学習の時間の短縮、辞書サイズの縮小という効果が得られた。今後、時代情報を更に精密に整備することにより、新規資料の解析精度が向上することも期待される。

参考文献

遠藤邦基（2010）『国語表記史と解釈音韻論』和泉書院.

小木曽智信（2013）「中古仮名文学作品の形態素解析」『日本語の研究』9-4, pp. 49-62.

小木曽智信・小町守・松本裕治（2013）「歴史的日本語資料を対象とした形態素解析」『自然言語処理』20-5, pp. 727-748.

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, pp. 101-123.

近藤泰弘（2012）「日本語通時コーパスの設計について」『国語研プロジェクトレビュー』3-2, pp.84–92.

参考 URL

MeCab: Yet Another Part-of-Speech and Morphological Analyzer https://code.google.com/p/mecab/

[1] この時代区分は日本語自体の変化によって規定されるべきものであるから、日本史の歴史区分と厳密に一致するものではなく、それぞれの境界も必ずしも明確なものではない。

[2] 本稿では、開始時代が「近世」で終了時代が「現代」であることを[近世-現代]と表す。

[3] 遠藤（2010）第 3 章参照。

[4] 2014 年 1 月時点。

Copyright(C) 2014 The Association for Natural Language Processing.
All Rights Reserved.