4 XML 構造化タグを利用した形態素解析 v2.1

当て字を振り仮名の通りに自動形態素解析するため，提案手法ではまず当て字が付された本行文字列に対し，以下の前処理を施す．具体的には，図2のようなXMLに対し，振り仮名タグ（r）でくくられた文字列を当該タグの属性（rt）の値（振り仮名文字列）で置換しておく（以下，この操作を**ルビを開く**と表現）．これを前処理とし，次段の処理として自動形態素解析を実施する．ルビを開く対象タグはあらかじめ3.3節で述べた基準で選定し，**type**属性と，その属性値として**当て字**を新しく付与した．この追加アノテーションは，今回すべて人手で実施したが，文献[15]で提案された当て字の自動検出手法を導入することで，将来的な自動化も検討している．

文献[16]では，洒落本の自動形態素解析を地の文・会話文の文体別に分け，それぞれで別個の形態素解析用辞書を用意している（洒落本の地の文用UniDic，洒落本の会話文用UniDic）．本研究でも同様に，地の文・会話文それぞれ専用のUniDicを用意する．辞書に収録する短単位の選定基準は文献[17]と同じであるが，ルビを開いた後の平仮名表記の解析にも対応するため，辞書の仮名形出現形のフィールドから当該の書字形出現形の仮名表記（片仮名）を取り出し，それを平仮名に置換した後，辞書のキーである表層形，および書字形出現形のフィールドに格納，これを新たなエントリとして辞書に追加を行なった（図5）[2]．これにより辞書のエントリ数（キー数）はおよそ2倍となった（約300万）．また各エントリに対し，新たに1フィールド（列）を追加し，そこに置換前の漢字表記での書字形出現形を残した．このフィールドを以降，**オリジナル表記**と呼ぶ．また元々辞書に登録されていたエントリのオリジナル表記には，表層形と同じ文字列が格納される．

以下の短単位情報アノテーション済み洒落本コーパス（21作品）を使い，地の文・会話文の文体別で，前述の辞書のMeCab v0.996[19]のコスト学習を実施する．表3に洒落本21作品の総文数，総短単位数，総文数の内訳を示す．

- 甲駅新話
- 阿闍陀鏡
- 北華通情
- 興斗月
- 新月花余情
- 陽台遺編・［女+壮］閣秘言
- 風流裸人形
- 異本郭中奇譚
- 箱まくら
- 粋の曙
- 花街鑑
- 花街寿々女
- 跖婦人伝
- 遊子方言
- 傾城買四十八手
- 繁千話
- 傾城買二筋道
- 郭中奇譚
- 侠者方言
- 聖遊廓
- 月花余情

辞書に新たに追加した平仮名表層形のエントリの学習のため，コーパスは通常の本行表記だけでなく，本行表記を（平仮名化した）仮名形出現形で置き換えた文も併用し，性能評価のため訓練9：評価1の割合で，文単位の分割を行なった．分割の結果，地の文の訓練用コーパスは12,447文，107,321短単位，評価用コーパスは1,383文，12,351短単位となった．また会話文の訓練用コーパスは13,176文，165,534短単位，評価用コーパスは1,464文，17,274短単位となった．学習時のCRF[18]の正規化項のハイパーパラメータC (=σ^2) は2.4に設定した[3]．

[2]すべての書字形について，その平仮名表記がキーとして網羅的に登録されていないため，この処理を実施している．

[3]UniDicの学習では伝統的にこの値が使用され続けており，由来は恐らく文献[19]と思われる．

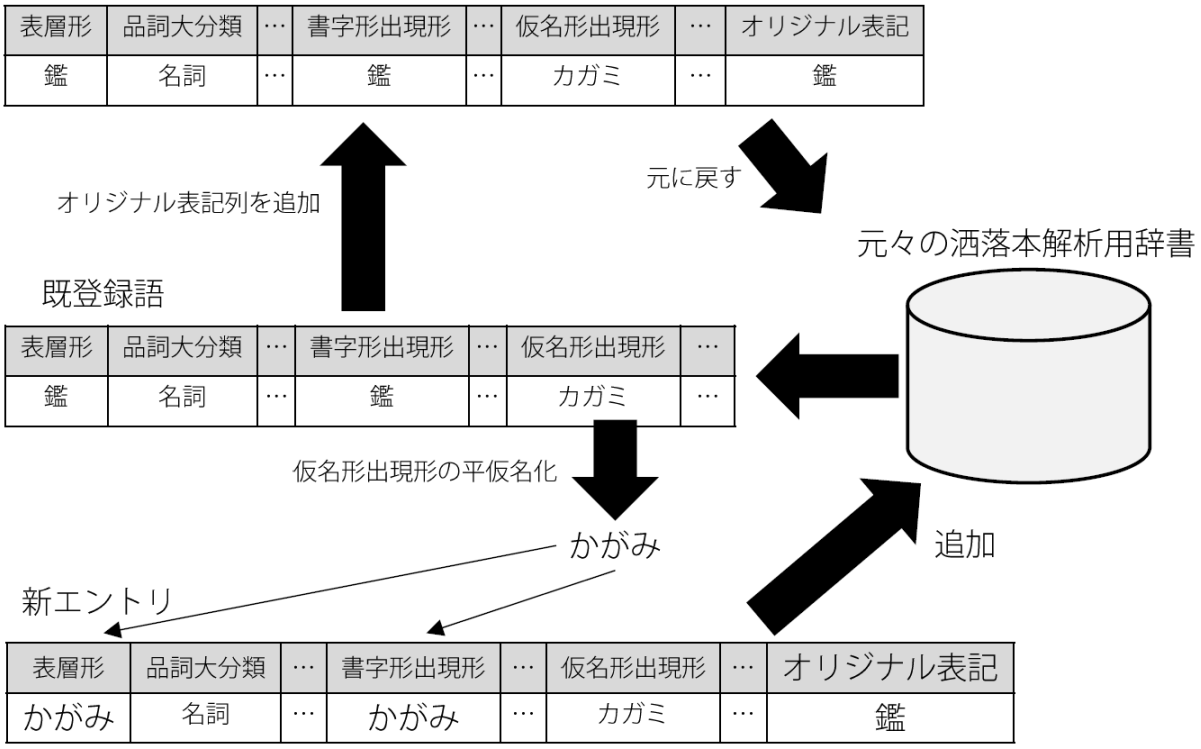

図 5: 解析用辞書の拡張：平仮名表層形エントリ及びオリジナル表記列の追加.

洒落本コーパス内でのインドメインの解析結果の性能評価を表 4 に示す．評価には形態素解析器性能評価ツール MevAL(Beta 版)[4]を使用した．評価は文献 [16, 17] と同じく，境界認定，品詞認定，語彙素認定，発音認定の 4 段階評価を実施する．境界認定は，文中での開始位置と終了位置が両方正しく認定できた（正解データと一致した）短単位数を評価している．品詞認定は，境界認定をパスした短単位の内，品詞大分類，中分類，小分類，細分類，活用型，活用形がすべて正しく認定できた短単位数を評価している．語彙素認定は，品詞認定をパスした短単位の内，語彙素読み，語彙素の両方が正しく認定できた短単位数を評価している．発音認定は，語彙素認定をパスした短単位の内，発音形出現形が正しく認定できた短単位数を評価している．

表 4 の結果を見ると，「語彙素レベルでの未知語作成」した場合の方が，「書字形レベルでの未知語作成」した場合よりも性能が高くなっているが，これは書字形レベルで未知語を作成しただけでは，同一語彙素の別の語として解析される場合が多く，その場合，誤って選ばれた語は正解の表記よりも短い表記で，そこで過分割が生じたためである[5]．実際，短単位分割を短単位境界判別の 2 値分類問題として捉えた場合，False Positive 数 (過分割境界数) は「書字形レベルでの未知語作成」で 527 であったが，「語彙素レベルでの未知語作成」では 191 であった（他方，False Negative 数（分割すべきであるのに未分割の境界数）には大きな差は見られなかった）.

作成した辞書を使い，type 属性が付与され，そ

表 3: 辞書の学習・評価に使用したコーパス（洒落本 21 作品）の総文数，総短単位数，総文字数.

	総文数	総短単位数	総文字数
地の文	6,915	59,836	98,890
会話文	7,320	91,404	153,282
計	14,235	151,240	252,172

[4] https://teru-oka-1933.github.io/meval/

[5] 近世は送り仮名が漢字中に完全に引っ込んだような活用語がキーとして多数登録されているので，その影響と思われる.

表 4: 自動形態素解析結果の性能（F1 値）.「ALL」は辞書内のすべてのエントリを学習・評価に使った場合.「書字形レベルでの未知語作成」は評価用コーパスにしか現れない書字形出現形を辞書から取り除いた場合.「語彙素レベルでの未知語作成」は UniDic の階層構造を利用し，評価用コーパスにしか現れない語彙素 ID を持つエントリを辞書からすべて削除した場合である．2 つの未知語作成はそれぞれ,「辞書に登録はされているが，その表記とは（若干）異なっている**未知表記**（=辞書未登録表記）」が含まれる場合の評価と,「辞書に未登録の完全な**未知語**（=辞書未登録語）」が含まれる場合の評価に相当する．またここでの評価ではルビを開く処理を実施していないことに注意してほしい.

		境界認定	品詞認定	語彙素認定	発音認定
地の文	ALL	96.36	92.39	91.20	90.71
	書字形レベルでの未知語作成	91.48	85.59	83.81	83.14
	語彙素レベルでの未知語作成	95.52	91.39	90.14	89.71
口語	ALL	97.06	93.53	92.73	92.43
	書字形レベルでの未知語作成	94.15	89.36	88.39	88.01
	語彙素レベルでの未知語作成	96.81	93.25	92.43	92.14

こからルビを開いた XML のテキスト部の解析を実施した[6]．対象は学習にも使った洒落本コーパスより「花街鑑」（インドメイン），学習には使っていない人情本コーパスより「比翼連理花廼志満台」（アウトドメイン）[7]である．自動解析の際は，N-best (N=100) 出力を行い，解析結果のオリジナル表記列から再構築した文が，ルビを開く前の本行表記と編集距離最小となる解析結果を選択した．編集距離の最小で同一の再構築文が複数あった場合には，コストが最も低い解析結果を選択した.

以上の処理により，辞書未登録の「口諚（こうぢやう）」のような特殊な漢字表記を「口上，名詞-普通名詞-一般」のように解析できるようになった．また「誘引（さそは｜れ）」のような複数短単位にまたがる漢字表記も望んだ通りに解析できるようになった．ただし，この 2 例はインドメインの洒落本での事例である．アウトドメインの人情本の事例では,「看病（みとり）」を「かんびょう」でなく,「みとり」として解析できるようになったことを確認した.

また type 属性を設けた振り仮名タグに限定して自動形態素解析結果の正解率を計算した（表 5）．辞書にあらかじめ登録していた当て字のキーもあったため，従来手法でもわずかに想定通りの解析結果が得られていたが，それ以上に本手法によって，目的の解析結果に向けた性能が，格段に向上したことがわかる.

表 5: 「type=“当て字”」を設けた振り仮名タグに限定した自動形態素解析結果の正解率．この結果は全数，専門家のアノテータが人手チェックを実施している.

	花街鑑		比翼連理花廼志満台	
	正解数/タグ数	正解率	正解数/タグ数	正解率
従来手法	26/164	15.9%	3/40	7.5%
本手法	137/164	**83.5%**	31/40	**77.5%**

参考文献

[15] 岡照晃:文字単位多対多自動アライメントを用いた日本語歴史コーパスのルビアノテーションの自動修正，人文科学とコンピュータシンポジウム論文集（じんもんこん 2016）, pp.133-138 (2016).

[16] 市村太郎，小木曽智信:文書構造を利用した近世期洒落本の形態素解析，言語処理学会第 22 回年次大会 発表論文集，pp.107-110（2016）.

[6]ここで使用した辞書の訓練用コーパスには，前述の評価用コーパスもマージしている.

[7]今回は初編上巻のみで試行.

[17] 鴻野知暁，小木曽智信：見出し語の時代情報を付与した電子化辞書の構築，言語処理学会 第 20 回 年次大会 発表論文集，pp. 209-212 (2014).

[18] Lafferty, J., McCallum, A. and Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data, *Proc. the 18th International Conference on Machine Learning* (*ICML 2001*), pp. 282-289 (2001).

[19] Kudo, T., Yamamoto, K. and Matsumoto, Y.: Applying Conditional Random Fields to Japanese Morphological Analysis, *Proc. the 2004 Conference on Empirical Methods in Natural Language Processing* (*EMNLP 2004*), pp.230-237 (2004).