Vol.2017-NL-231 No.9
Vol.2017-SLP-116 No.9
2017/5/15

日本語部分形態素アノテーションコーパスの構築

林部 祐太[1,a)]

概要：形態素解析誤りの大きな原因に未知語の問題があり，辞書を大規模して対処する研究が近年なされている．ところが，語を追加することで，以前正しく解析できていたものが誤った解析になってしまうリグレッション（退行）が起こりうる．そこでリグレッション検出のために，自然アノテーションを用いた半自動アノテーションと，文字列検索ツールを用いた手動アノテーションの 2 つのアプローチで，部分形態素アノテーションコーパスを構築した．本コーパスはオープンソースライセンスで公開予定である．

1. はじめに

形態素解析は，分かち書きされない言語の解析を行う上で欠かすことのできない基礎的な処理である．そのため，古くから様々な手法が提案されてきている [1].

日本語の形態素解析では，コストを形態素アノテーションコーパスで教師有り学習した辞書を事前に用意し，その辞書で形態素候補を列挙して解析する方式 [2] が広く用いられている．辞書にない形態素（未知語）[*1]が少ない新聞記事を対象とした形態素分割実験では，F 値で 99 を越える非常に高い精度が報告されている [3]．一方で，SNS やブログといった未知語を多く含むテキストでは解析精度が低いという問題が依然として残っている．

そのため，未知語を削減するために，最近では様々な研究がなされている．主なアプローチとして，辞書やコーパスといった言語資源を整備するアプローチと，解析時に未知語を既知語に帰着するアプローチ [4], [5], [6], [7] の 2 つが挙げられるが，開発コスト・保守コストの点から前者のアプローチが広く用いられている [8], [9]．特に，辞書を拡張するアプローチは容易で効果も高いことから，広く用いられている．

ところが，拡張後辞書は拡張前辞書と比べて常に良い解析を可能にするとは限らない．

例えば，JUMAN++が用いる辞書には，JUMAN が用いる辞書に無い形態素「物欲（ぶつよく）」が追加されている．ところが，(1a) に対して，JUMAN 7.01 では (1b) と正しく解析する[*2]一方で，JUMAN++ 1.02 で解析すると (1c) と誤解析してしまう．

(1) a. ついつい物欲（ものほ）しそうになってしまう
b. ついつい/物/欲し/そうに/なって/しまう
c. ついつい/物欲/し/そうに/なって/しまう

また，形態素解析器 MeCab[*3]用 IPA 辞書に語を大量に追加した辞書 mecab-ipadic-NEologd でも，(2a) に対して (2b) と正しく解析できていたものが，(2c) と誤解析する[*4]ようになってしまっている事例がある．（2017 年 3 月 20 日のバージョン）

(2) a. 鱗片の外側には細かい伏せた毛がある。
b. 鱗片/の/外側/に/は/細かい/伏せ/た/毛/が/ある/。
c. 鱗片/の/外側/に/は/細かい/伏せ/た毛/が/ある/。

このような，以前正しく解析できていたものが誤った解析になってしまうリグレッション（退行）を伴う変更は，安定した運用を行う上で非常に問題となる．

拡張辞書にリグレッションが生まれる原因の 1 つに，アノテーションコーパスのカバレッジが十分でないことが挙げられる．

例えば，JUMAN++の学習・評価に使われる京大コーパス・KWDLC は合わせて約 5.3 万文あるが，「物欲」という 2 文字は 1 回も出現しない．そのため，不具合として認知されなかったと思われる．

そこで，本研究では形態素解析器のリグレッションを検

[1] フェアリーデバイセズ株式会社
[a)] hayashibe@fairydevices.jp

[*1] 一般に見出し語化されていなければ未知語とよばれる．例えば，「カわイィ」「かわええ」といった非規範的な異表記・活用形は，それらが見出し語として辞書になければ，「かわいい」が辞書にあっても未知語である．

[*2] 本稿では形態素境界アノテーションを |, システムが出力した形態素境界を/で示す

[*3] http://taku910.github.io/mecab/

[*4] 「多毛」の異表記として「た毛」が追加されている

出するためのコーパスの構築を行う．これは，かな漢字変換システム [10] や商用形態素解析器 JMAT[*5][11] といった製品で，平均精度評価だけでなく回帰評価も製品出荷基準として用いられていることに着想を得ている．

これまでに公開されてきた形態素アノテーションコーパスといえば，形態素解析器の学習のために，文中の全ての語に対してアノテーション（フルアノテーション）を行ったコーパスがほとんどであった．これに対して，本研究ではリグレッションの検出に目的を絞り，高いコストがかかるフルアノテーションではなく，文中の一部の語に対してのアノテーション (部分アノテーション) を行なったコーパスの構築を行う．

本稿の以降の構成は次のとおりである．まず 2 節で現代日本語を対象とした主な形態素解析辞書とコーパスについて述べる．3 節で部分アノテーションの 2 つのアプローチについて述べ，4 節で構築したコーパスの分析を行う．最後に 5 節で今後の課題について述べる．

2. 関連研究

ここでは現代日本語を対象とした主な形態素解析辞書とコーパスについて述べる．

2.1 形態素解析辞書

特定ドメインに限らない汎用的な形態素解析辞書として以下の辞書が古くから広く用いられている．

- IPA 辞書[*6]: IPA 品詞体系に基づく辞書．2007 年の最終更新から少なくとも 10 年経過しており，公式でのメンテナンスはもう行われていないと思われるが，mecab-ipadic-NEologd の開発の一環で佐藤らによって修正作業が行われている [12].
- NAIST-jdic[*7]: IPA 辞書のライセンス問題をクリアし，表記ゆれ情報や複合語情報を付与した辞書．2011 年の最終更新から少なくとも 5 年経過しており，公式でのメンテナンスはもう行われていないと思われる．
- UniDic[*8]: 国語研短単位に基づく辞書．現代日本語書き言葉均衡コーパス (BCCWJ) のために開発された．2013 年以降更新されていないが，拡張が公式に予定されている [13], [14]．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られる．
- JUMAN 辞書[*9]: 益岡・田窪文法に基づく品詞体系を採用し，表記ゆれ情報や複合語情報を付与した辞書．形態素解析器 JUMAN に付属している．最新版は，新しい形態素解析器 JUMAN++[*10]の付属辞書として公開・開発されている．

最近ではオンライン百科事典 Wikipedia やオンライン辞書 Wiktionary から獲得した語 [3], [15] や，ウェブテキストから獲得した語 [9], [16], [17] を追加して，語彙を拡張する研究が行われている．それらの手法を用いて作られた辞書は，JUMAN++辞書（の一部）や NEologd[*11] としてフリーなライセンスで公開されており，広く用いられている．

専門ドメインに特化した辞書としては，農業用語辞書 [18]，医療用語辞書 [19]，忍殺語[*12]辞書[*13] などが人手で整備されている．

2.2 形態素アノテーションコーパス

フル形態素アノテーションコーパス

形態素解析器の学習・評価には，以下のような形態素情報が人手でフルアノテーションされたコーパスが用いられている．

- 京都大学テキストコーパス[*14]: 毎日新聞 1995 年版の記事約 2 万文と社説約 2 万文に形態素・構文情報を人手で付与したテキストコーパス [20]．そのうち 5,000 文に対しては，格関係，照応・省略関係，共参照の情報も付与されている [21].
- 京都大学ウェブ文書リードコーパス (KWDLC)[*15]: さまざまなウェブ文書の冒頭 3 文に各種言語情報（形態素・固有表現・構文・格関係，照応・省略関係，共参照，談話関係）を人手で付与したテキストコーパス [22] で，約 5,000 文書（約 1.5 万文）からなる．
- 現代日本語書き言葉均衡コーパス (BCCWJ)[*16] 新聞，雑誌，書籍，白書，Yahoo!知恵袋，Yahoo!ブログなどといったさまざまなレジスターからなる均衡コーパス．自動形態素解析された約 600 万文のうち，コアデータとよばれる約 6 万文が人手による確認・修正が行われている．

その他には，以下のようなコーパスもある．

- EDR 日本語コーパス[*17]
- 新聞記事 GDA コーパス 2004[*18]
- Kyoto-University and NTT Blog コーパス[*19][23]
- 日本語話し言葉コーパス (CSJ)[*20]

[*5] http://www.atok.com/biz/jmat.html
[*6] https://ja.osdn.net/projects/ipadic/
[*7] https://ja.osdn.net/projects/naist-jdic/
[*8] https://ja.osdn.net/projects/unidic/
[*9] http://nlp.ist.i.kyoto-u.ac.jp/?JUMAN
[*10] http://nlp.ist.i.kyoto-u.ac.jp/?JUMAN++
[*11] https://github.com/neologd
[*12] 小説「ニンジャスレイヤー」で用いられる独特の表現
[*13] https://twitter.com/njdict_Chado
[*14] http://nlp.ist.i.kyoto-u.ac.jp/?京都大学テキストコーパス
[*15] http://nlp.ist.i.kyoto-u.ac.jp/?KWDLC
[*16] http://pj.ninjal.ac.jp/corpus_center/bccwj/
[*17] http://www2.nict.go.jp/ipp/EDR/JPN/J_indexTop.html
[*18] http://www.gsk.or.jp/catalog/gsk2009-b/
[*19] http://nlp.ist.i.kyoto-u.ac.jp/kuntt/
[*20] http://pj.ninjal.ac.jp/corpus_center/csj/

- 首都大日本語 Twitter コーパス[*21][24]

部分形態素アノテーションコーパス

形態素解析器のドメイン適応の研究の一部として，医療マニュアル [25]，料理レシピ [26]，絵本 [27] などで部分アノテーションが行われている．

ウェブサイト「部分アノテーションの共有」[*22] では，JUMAN++が誤解析した部分文字列に対するアノテーションを収集しており，誰でも閲覧・投稿できる．例を (3), (4) に示す．2017 年 4 月現在，約 70 件が公開されている．

(3) 無理難題を周囲に | ほざく |

(4) ブラウスが | しわくちゃ | だ

また，複合辞や機能語といった機能表現の用例を収集したデータベースに，複合辞用例データベース “MUST1” [28][*23] や機能語用例データベース「はごろも」[29][*24] があり，これらも部分アノテーションコーパスとみなせる．

MUST1 は現代語複合辞用例集 [30] に収録されている複合辞 123 項目を 337 項目に細分化し，毎日新聞 1995 年版の文から各項目につき最大 50 文づつ収集し，6 種類に分類してラベルを付与している．例えば，立場を示す複合辞「にとり」について，(5) へは現代語複合辞用例集の用法と一致するとのラベルを付与している．また，(6) へは複合辞「にとり」の用例としては不適切とのラベルを付与し，補足コメントとして「に | とりつか」という分割情報も付与している．

(5) 日本とロシアの双方にとり、地域をどう安定させるかが問題だ。

(6) ここにも海のロマンにとりつかれた男がいた。

はごろもは，日本語教育への活用を目的として，1,849 項目の文法項目について難易度を付与し，話し言葉と書き言葉のコーパスから該当する用例を 4 種類の話し言葉コーパスと，4 種類の書き言葉コーパスから抽出したデータベースである．例えば，立場を示す複合辞「にとり」について，(7) のような用例を含んでいる．

(7) 多くの企業にとり、短期間の円急騰が予想外だったとすれば、…

3. 部分形態素アノテーションコーパスの構築アプローチ

コーパスのテキストには，幅広い話題もカバーしており[*25]，また再配布可能なライセンス (CC-BY-SA 3.0) であるオンライン百科事典 Wikipedia 日本語版[*26]を用いる．

そして，アノテーションには，次の 2 つのアプローチを用いる．1 つ目は自然アノテーションを用いた半自動アノテーションである．これは，大量のアノテーションを得ることを目的としており，Wikipedia のハイパーリンクと形態素解析器の解析結果を組み合わせて，アノテーションを行う．3.1 節で詳述する．

2 つ目は，文字列検索ツールを用いた手動アノテーションである．これは，半自動アノテーションでは得られにくい形態素に関する部分アノテーションを行うことを目的としている．3.2 節で詳述する．

なお，品詞体系は JUMAN 品詞体系を用いる．以下，断りがなければ JUMAN 品詞体系の用語を用いる．

3.1 自然アノテーションを用いた半自動アノテーション

3.1.1 自然アノテーションと確認・修正の必要性

中国語の単語分割タスクにおいて，[31] はハイパーリンクの両端は単語境界であると仮定して Wikipedia の約 390 万文を用いて自己学習を行い，精度向上を行った．このハイパーリンクのマークアップを，彼らは自然アノテーション (natural annotation) とよんだ．

本研究でも，大量の部分アノテーションを得るため，自然アノテーション用いることにする．まず，2017 年 1 月 20 日版の Wikipedia ダンプデータ[*27]から本文抽出・文分割・文字の正規化・不要文削除等を行い，14,671,896 文を得た．そのうち，ハイパーリンクを含む 3,507,562 文に対して，ハイパーリンクの始点と終点を形態素境界とみなす変換を行った．なお，ハイパーリンク内部の文字間には単語境界が存在しうることに留意されたい．

この約 350 万文には JUMAN 品詞体系の基準に一致しない形態素分割も含まれている．

例えば，(8a) では「| スレンダー | な」となっているが，「スレンダーな」はナ形容詞「スレンダーだ」のダ列基本連体形で 1 語であり，「| スレンダーな |」が正しい．

同様に，(8b) の「| 波 | 打った」は子音動詞「波打つ」のタ形で 1 語であり，「| 波打った |」が正しい．

(8) a. 街並み・橋梁／新中川をさわやかに吹き抜ける風をモチーフに、| スレンダー | な形状で軽快さを演出されている。

b. また、前線は直線的な場合が多いが、| 波 | 打ったような形をしていることもある。

修正は，ある程度はルールによる自動処理を行えるものの，最終的には人手によるチェックが必要である．

例えば (9a) と (9b) はともに「| ダイレクト | に」とアノテーションされている．(9a) は名詞の「ダイレクト」に助

[*21] https://github.com/tmu-nlp/TwitterCorpus
[*22] http://lotus.kuee.kyoto-u.ac.jp/~morita/JUMAN++/pannotation.html
[*23] http://nlp.iit.tsukuba.ac.jp/must/
[*24] http://hgrm.jpn.org/
[*25] 2017 年 4 月現在約 100 万記事が存在する
[*26] https://ja.wikipedia.org
[*27] https://dumps.wikimedia.org/jawiki/

詞の「に」が続いているため正しい．一方で，(9b) の「ダイレクトに」は，ナ形容詞「ダイレクトだ」のダ列基本連用形で 1 語であり，「| ダイレクトに |」が正しい．これらは，意味まで考慮しなければ修正できない．

(9) a. みずほ銀行のインターネットバンキングサービスである | みずほダイレクト | に申し込むと、自動的に宝くじラッキーラインの会員になる。
b. 恋愛の楽しさや片思いの切なさを、より | ダイレクト | に楽曲に反映させた作品にも仕上がっている。

さらには，ハイパーリンクが誤っている場合もある．(10a) は「| スピン |」が正しく，(10b) は誤ってリンクが作られていると思われる．

(10) a. このモーメントはその分子・原子における電子軌道での不対電子の | スピンか | ら生まれている。
b. この一覧ではアンコール放送分は除いて | い | る。

3.1.2 形態素解析器を用いたサンプリングと確認・修正

人的・時間的コストの観点から全ての文をチェックするのは困難である．そのため，形態素解析器が誤りやすいと思われるデータを一部抽出し，人手で確認・修正することにした．

まず，ハイパーリンクを含む約 350 万文に対して JUMAN++で形態素解析を行った．自然アノテーションで境界とされている箇所が JUMAN++の出力で形態素境界とならなかった（以下では「違反」とよぶ）文は 41,664 文であった．そして，その箇所を含む形態素の品詞ごとにグループ化する．

例えば，(11) の「あるいて座」では，自然アノテーションでは「る」と「い」の間に境界があるが，JUMAN++は動詞「あるいて」（歩いて）で 1 形態素として出力している．そのため，「動詞」のグループとして扱う．また，(12) の「はとこの」では，自然アノテーションでは「こ」と「の」の間に境界があるが，JUMAN++は指示詞「この」（此の）で 1 形態素として出力している．そのため，「指示詞」のグループとして扱う．なお，複数の違反が有る場合は，文の先頭に近い違反のみ考えることとする．

(11) 我々の銀河系の場合、全天に分布しているが、銀河中心の/ある | いて/座/| の方向に多く見られる。

(12) 奥州家の家督は義兄弟/で/| は/と/こ | の/業氏が継いだ。

そして，グループごとにいくつかランダムで選び，確認・修正作業を行った．グループごとに集計した数とコーパスに採録した用例数を表 1 に示す．

品詞大分類	品詞細分類	違反数	採録数
感動詞	*	4	4
形容詞	*	4281	62
指示詞	副詞形態指示詞	1	1
	連体詞形態指示詞	29	26
助詞	格助詞	8	6
	終助詞	3	3
	接続助詞	11	11
	副助詞	10	9
助動詞	*	29	16
接続詞	*	1	0
接尾辞	形容詞性述語接尾辞	109	20
	形容詞性名詞接尾辞	476	9
	動詞性接尾辞	22	0
	名詞性名詞助数辞	2227	0
	名詞性名詞接尾辞	3	0
動詞	*	634	307
判定詞	*	10	6
副詞	*	339	116
未定義語	その他	1072	54
	アルファベット	291	5
	カタカナ	7818	31
名詞	サ変名詞	406	2
	形式名詞	9	1
	固有名詞	86	1
	時相名詞	323	0
	人名	1619	1
	数詞	997	0
	組織名	136	1
	地名	2818	11
	普通名詞	17839	30
	副詞的名詞	4	0
連体詞	*	49	9
合計		41664	741

表 1 Juman++の Wikipedia 自然アノテーション違反数と採録数

また，JUMAN++で違反が無かった約 346 万文に対して，MeCab（辞書は UniDic 2.1.2）でも形態素解析を行い，違反する 22,930 文を抽出した．これらは，形態素解析器が誤る可能性が比較的高い自然アノテーションとみなせる．そして同様にグループ化を行い，JUMAN 品詞体系での確認・修正をグループごとに行った．集計した結果とそのうちの採録数を表 2 に示す．

なお，似たような文同士や，本文抽出エラーで非文となっている文等を除外しているため，実際に確認した数は，採録数よりも数割程度多い．

3.2 手動アノテーション

百科事典という性格上，Wikipedia のリンクの大半は名詞や複合名詞であり，機能表現には滅多にリンクが張られることはない．そのため，自然アノテーションを用いた半自動アノテーションだけでは，機能表現に対するアノテーションが不足する．[32] の形態素解析のエラー分析におい

図 1 文字列検索ツール

品詞大分類	品詞細分類	違反数	採録数
感動詞	フィラー	21	21
	一般	8	0
記号	一般	2	2
形状詞	一般	159	47
	助動詞語幹	1	1
形容詞	一般	14	11
助詞	格助詞	3	3
	接続助詞	4	3
	副助詞	24	20
助動詞	*	32	21
接続詞	*	3	3
接頭辞	*	4	2
接尾辞	動詞的	1	1
	名詞的	8	5
代名詞	*	6	3
動詞	一般	269	123
	非自立可能	22	8
副詞	*	212	60
補助記号	一般	1	0
名詞	固有名詞	2180	199
	数詞	12	4
	普通名詞	19840	22
連体詞	*	104	3
合計		22930	589

表 2 Juman++では違反せず，MeCab+UniDic で違反した Wikipedia 自然アノテーションの数と採録数

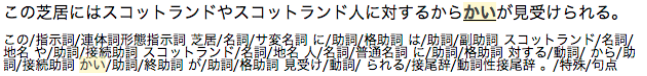

図 2 形態素解析結果表示ボタンの動作

て，辞書の拡張での誤りは機能表現に関する箇所に集まっているとしており，機能表現に対するリグレッションの検出は重要である．

そこで，手動で機能表現を中心として，部分アノテーションを行う．アノテーションは，ブラウザで動作するツール

(a) 表層での絞り込み　(b) 品詞での絞り込み

図 3 絞り込みボタンの動作．数字は該当文数を示す．

を生成し，それを使って行う．このツールはブラウザ上で動作し，入力した文字列（クエリ）を含む文を一覧表示する．「かい」を入力した例を**図 1** に示す．助詞の「かい」を含む (13) の他に，(14) のような形態素の部分文字列として「かい」を含む文も表示されている．

(13) 愛と平和と理解を信じる**かい**？

(14) 休日に運転されている特急**かい**おう 5 号が行き違い待ちで当駅に運転停車する。

検索結果は表形式で表示され，ID，テキスト等が 1 行に表示される，「ID」列はテキストの ID，「記事」列はテキストを含む記事の ID を示す．記事 ID には Wikipedia のページへリンクが張られている．「テキスト」列には，クエリが**太字**になった文が表示されている．

列「マッチ表層」と列「マッチ品詞」には，それぞれクエリを含む形態素の表層列（「テキスト」列では下線でマークされている）と品詞列が表示される．なお，形態素解析は MeCab と内製の JUMAN 品詞体系準拠辞書による．例えば，(14) は「特急/**か**/**い**おう/」と（誤）解析され，「かい」は助詞「か」と動詞「いおう」に含まれるので，「か–い

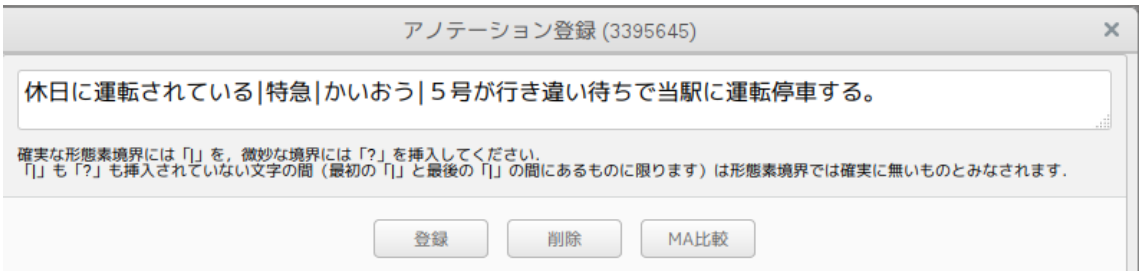

図 4 アノテーション登録画面

おう」と「助詞–動詞」がそれぞれ表示されている．

「Type」列には，解析結果が正しいか誤っているか等が入力できる．図 1 で “OK” となっているのは形態素境界が正しいと，“NG_SEG” となっているのは誤っているとアノテーションした文である．

図 1 の “toggleMA” ボタンをクリックすると**図 2** のように形態素解析結果を表示できる．また，右クリックメニューから，**図 4** に示したような画面を呼び出すことができ，アノテーションを登録できる．

なお，このツールでは，マッチする表層列や品詞列で検索結果を絞り込める．「マッチ表層」「マッチ品詞」のボタンを押すと，**図 3** のように，該当する表層列や品詞列の一覧が表示される．そして，1 つクリックすると，それに該当する文が絞りこめる．この機能は，クエリの用法ごとにアノテーションするのに役に立つ．

JUMAN++辞書の助詞等をクエリとしてこのツールを用い，640 文にアノテーションを行った．

4. 構築したコーパスの詳細と分析

4.1 アノテーション仕様

アノテーションは森ら [33] にならって，形態素境界を 3 値で表現する．森らは，形態素境界がある場合は |，無い場合は –，不明である場合は ␣ といった記号を文字間に挿入した．

本アノテーションでは，効率化のために，形態素境界がある場合は |，不明である場合は?を文字間に挿入し，無い場合は何も挿入しないことにした．ただし，この記法は最初に現れる | から最後に現れる | までのみ有効とする．すなわち，文頭から最初に現れる | までと，最後に現れる | から文末までは，たとえ文字間に何の記号が無かったとしても形態素境界が無いということを意味せず，未アノテーション扱いとする．

例えば，(23) のアノテーションは，「/は/かにカマボコ/」または「/は/かに/カマボコ/」という形態素分割が正しいことを示している．

4.2 アノテーションの例

表 3 に構築したコーパスの一部と，形態素解析器での解析結果を示す．形態素解析には JUMAN 7.01，JUMAN++ 1.02，MeCab 0.996 と UniDic 2.1.2，MeCab と mecab-ipadic-NEologd（2017 年 3 月 20 日のバージョン）の 4 種類を用いた．

なお，本アノテーションの分割は JUMAN 品詞体系の基準であり，UniDic や mecab-ipadic-NEologd での解析結果はあくまで参考情報であることに留意されたい．

(15)，(16) は「（人名）+ら」に対するアノテーションである．誤り例の「潜ら」「耽ら」のように，人名の末尾の漢字 1 文字と「ら」で，動詞の未然形として誤解析する事例は数多く見られた．

(17)，(18) は複合名詞に対するアノテーションの例である．(17) では UniDic が |地下?鉄|網| を/地下/鉄網/と誤解析し，(18) では JUMAN, JUMAN++が | 京都?大学 | 生理?学 | を/京都/大学生/理学/と誤解析し，複合名詞の意味が大きく変わっている．

(19)，(20) は助詞に対するアノテーションの例，(21)，(22) は人名に対するアノテーションの例である．

4.3 意味・知識処理が必要な例

アノテーションデータの中には，確実な解析を行うためには，意味・知識処理が必要となってくる例も見られた．表 4 に例を示す．

(23) から (28) の誤り例は，いずれも文法的な不自然さは感じにくいが，意味的な不自然や知識から誤っている判断できる例である．

例えば，(23) の「墓にカマボコがある」という状況や (24) の「妻であり，子でもある」という状況は一般に考えにくく，意味的に不自然な状況である．また，(25) は「輝ける闇」という小説は「開高健」が書いたという知識，(26) は「ヨハンナ・ベア」の結婚歴の知識，(27) は 80 年代の朝鮮事情の知識，(28) は「由利」と「たけし軍団」の関係の知識があれば，それぞれ正しく解析できると考える．

そのため，これらを正しく解析するには，文法的妥当さに加えて，意味的，知識的な妥当さも加味する必要がある．

なお，表 4 では，JUMAN++が多くの事例で正解しているが，用いている RNN 言語モデルの効果が大きいと考えられる．RNN 言語モデルは，形態素を意味的に汎化され

アノテーション	解析結果 J	J+	N	U	誤り例
(15) · · · 改組する形で | 片山?潜 | ら | と労働組合期成会を結成した。	✓	✗	✓	✓	/片山/潜ら/ (J+)
(16) 弟に陸景、陸玄、陸機、陸雲、| 陸?耽 | らがいる。	✓	✗	✗	✗	/陸/耽ら/ (J+, N, U)
(17) · · · 都市には１４路線ある | 地下?鉄 | 網が整備されている。	✓	✓	✓	✗	/地下/鉄網/ (U)
(18) １９５４年に | 京都?大学 | 生理?学 | 教室入室。	✗	✗	✓	✓	/京都/大学生/理学/ (J, J+)
(19) だが彼 | こそ | は高度な科学力とすぐれた肉体で · · ·	✓	✓	✓	✓	-
(20) · · · 学徒兵が外出 | がてら | に主人公の家で記念写真を撮る。	✗	✓	✗	✗	/外出/が/てら/に/ (J, N, U)
(21) · · ·| 藤堂?ユリカ |、藤原みやびの歌唱を担当。	✗	✓	✓	✗	/藤/堂/ユリカ/ (J), /藤堂/ユ/リカ/ (U)
(22) | 村川?梨衣 | にとって初のソロＤＶＤである。	✗	✓	✓	✗	/村/川/梨/衣/ (J, U)

表 3 コーパスの一部と，その形態素解析結果．J, J+, N, U はそれぞれ JUMAN, JUMAN++, mecab-ipadic-NEologd, mecab-unidic での解析結果を示す．

アノテーション	解析結果 J	J+	N	U	誤り例
(23) 食品関係で有名な例として | は | かに?カマボコ | がある。	✗	✓	✗	✗	/はか/に/カマボコ/ (J, N, U)
(24) · · · 中田重治の | 妻 | かつ子 | が死去した時には · · ·	✗	✓	✓	✓	/妻/かつ/子/
(25) 輝ける闇は、| 開高?健 | 作の小説。	✓	✓	✓	✗	/開高/健作/ (U)
(26) １８６８年に | またいとこ | のヨハンナ・ベアと結婚。	✗	✓	✓	✗	/また/いとこ/ (J), /またい/と/この/ (U)
(27) · · · ８０年代初頭の南北 | 朝鮮 | は一時緊張状態にあった。	✓	✓	✗	✗	/南北朝鮮/ (N), /南/北朝鮮/ (U)
(28) その後、由利の元を離れ | たけし?軍団 | に加入する。	✗	✗	✓	✗	/離れた/けし/軍団/ (J, J+), /離れ/た/けし/軍団/ (U)

表 4 形態素境界確定に意味処理や知識処理が必要な例と解析結果．略号は表 3 と同じ．

たベクトルとして扱い，大規模に自動解析したウェブコーパスから意味的に妥当なベクトル列を学習する言語モデルである．そして解析では，ラティスの素性のスコアだけでなく，RNN 言語モデルのスコアも使っている [3]．そのため，多くの事例でより自然な形態素分割が選択できたと考える．

5. おわりに

本稿では形態素解析器開発の際に生じるリグレッションの検出のために，部分形態素アノテーションコーパスの必要性を主張した．そして，自然アノテーションを用いた半自動アノテーションと文字列検索ツールを用いた手動アノテーションの 2 つのアプローチで，コーパスを構築した．

今後の課題としては，主に 3 つ挙げられる．

1 つ目は，コーパスのさらなる規模の拡大である．現時点で約 2,000 文を含んでいるが，特に機能表現に関しては，まだ分量が不足していると思われる．そのため，機能表現辞書「つつじ」[34][*28]等を参考に，用例を充実させていきたい．

2 つ目は，方言や口語的な表現を含んだ文に対するアノテーションである．Wikipedia は百科事典という性格上，会話文の引用などごく少量の例外はあるものの，非規範的な表記や方言など，いわゆる「くだけた表現」はほとんど出現しない．そこで，星空文庫[*29]においてフリーなライセンスで公開されている小説などを対象に，アノテーションを行いたいと考えている．

3 つ目は品詞などの形態素情報の付与である．今回構築したデータセットでは，形態素境界が合っているかだけの確認しか行えず，品詞や語彙素（JUMAN 辞書では概ね代表表記に相当する）まで正しく同定できているかの確認は行えない．そこで，曖昧性の高い形態素を中心にアノテーションを行いたいと考えている．

なお，本コーパスはオープンソースライセンスで公開予定である．オープンソースライセンスで公開された大規模な部分形態素アノテーションコーパスは，著者が知る限り本コーパスが初めてである．より品質の安定した形態素解析器の開発のため，回帰テスト等に広く用いられることを願っている．

謝辞 Wikipedia 日本語版の執筆者の皆様に感謝する．なお，本文中の例文は一部を除き全て Wikipedia 日本語版からの引用である．

[*28] http://www.cl.inf.uec.ac.jp/lr/tsutsuji/
[*29] https://slib.net/

参考文献

[1] 鍜治伸裕：日本語形態素解析とその周辺領域における最近の研究動向，日本知能情報ファジィ学会誌，Vol. 26, No. 6, pp. 174–183 (2013).
[2] Kudo, T., Yamamoto, K. and Matsumoto, Y.: Applying Conditional Random Fields to Japanese Morphological Analysis, *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pp. 89–96 (2004).
[3] Morita, H., Kawahara, D. and Kurohashi, S.: Morphological Analysis for Unsegmented Languages using Recurrent Neural Network Language Model, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 2292–2297 (2015).
[4] 風間淳一，光石　豊，牧野貴樹，鳥澤健太郎，松田晃一，辻井潤一：チャットのための日本語形態素解析，言語処

理学会第 5 回年次大会発表論文集，pp. 509–512 (1999).
[5] 工藤　拓，市川　宙，Talbot, D.，賀沢秀人：Web 上のひらがな交じり文に頑健な形態素解析，言語処理学会第 18 回年次大会予発表論文集，pp. 1272–1275 (2012).
[6] Sasano, R., Kurohashi, S. and Okumura, M.: A Simple Approach to Unknown Word Processing in Japanese Morphological Analysis, *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, Nagoya, Japan, pp. 162–170 (2013).
[7] 斉藤いつみ，貞光九月，浅野久子，松尾義博：文字列正規化パタンの獲得と崩れ表記正規化に基づく日本語形態素解析，自然言語処理，Vol. 24, No. 2, pp. 297–314 (2017).
[8] Mori, S. and Neubig, G.: Language Resource Addition: Dictionary or Corpus?, *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, pp. 1631–1636 (2014).
[9] 佐藤敏紀，橋本泰一，奥村　学：単語分かち書き用辞書生成システム NEologd の運用-文書分類を例にして-，情報処理学会第 229 回自然言語処理研究会予稿集，pp. 15:1–14 (2016).
[10] 工藤　拓，小松弘幸，花岡俊行，向井　淳，田畑悠介：統計的かな漢字変換システム Mozc，言語処理学会第 17 回年次大会発表論文集，pp. 948–951 (2011).
[11] 北浦雅子，紀伊馬章：拡張型 NLP『JMAT』における実利用に向けた形態素解析のリソースチューニング，語彙資源活用シンポジウム (2017). `https://www.slideshare.net/JSUXDesign/nlpjmat`.
[12] 佐藤敏紀，橋本泰一，奥村　学：単語分かち書き辞書 mecab-ipadic-NEologd の実装と情報検索における効果的な使用方法の検討，言語処理学会第 23 回年次大会発表論文集，pp. 875–878 (2017).
[13] 前川喜久雄：日本語の全体像を知るために―国立国語研究所による言語資源整備―，第 7 回産業日本語研究会・シンポジウム予稿集，pp. 3–6 (2017).
[14] 岡　照晃：『UniDic』の拡張計画，語彙資源活用シンポジウム (2017). `http://pj.ninjal.ac.jp/corpus_center/lrw/lrw2016_2016_03_06_11ok.pdf`.
[15] 柴田知秀，村脇有吾，黒橋禎夫，河原大輔：実テキスト解析をささえる語彙知識の自動獲得，言語処理学会第 18 回年次大会予発表論文集，pp. 81–84 (2012).
[16] 鍛治伸裕，福島健一，喜連川優：大規模ウェブテキストからの片仮名用言の自動獲得，電子情報通信学会論文誌，Vol. 92, No. 3, pp. 293–300 (2009).
[17] 村脇有吾，黒橋禎夫：形態論的制約を用いたオンライン未知語獲得，自然言語処理，Vol. 17, No. 1, pp. 55–75 (2011).
[18] 法隆大輔，深津時広，大塚　彰，木浦卓治，平藤雅之，二宮正士：農業関連文書用形態素解析サーバの開発とテキストの自動分類による検証，農業情報研究，Vol. 13, No. 2, pp. 127–137 (2004).
[19] 相良かおる，小野正子，小作浩美，鈴木隆弘，高崎光浩，嶋田　元：分かち書き用辞書 ComeJisyo の評価，医療情報学，Vol. 32, No. 6, pp. 301–307 (2012).
[20] Kurohashi, S. and Nagao, M.: Building a Japanese Parsed Corpus, *Treebanks: Building and Using Parsed Corpora*, Springer Netherlands, chapter 14, pp. 249–260 (2003).
[21] Kawahara, D., Kurohashi, S. and Hasida, K.: Construction of a Japanese Relevance-tagged Corpus, *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, pp. 2008–2013 (2002).
[22] 萩行正嗣，河原大輔，黒橋禎夫：多様な文書の書き始めに対する意味関係タグ付きコーパスの構築とその分析，自然言語処理，Vol. 21, No. 2, pp. 213–248 (2014).
[23] 橋本力，黒橋禎夫，河原大輔，新里圭司，永田昌明：構文・照応・評価情報つきブログコーパスの構築，自然言語処理，Vol. 18, No. 2, pp. 175–201 (2011).
[24] 大崎彩葉，唐口翔平，大迫拓矢，佐々木俊哉，北川善彬，堺澤勇也，小町　守：Twitter 日本語形態素解析のためのコーパス構築，言語処理学会第 22 回年次大会発表論文集，pp. 16–19 (2016).
[25] 坪井祐太，森　信介，鹿島久嗣，小田裕樹，松本裕治：日本語単語分割の分野適応のための部分的アノテーションを用いた条件付き確率場の学習，情報処理学会論文誌，Vol. 50, No. 6, pp. 1622–1635 (2009).
[26] Mori, S., Sasada, T., Yamakata, Y. and Yoshino, K.: A Machine Learning Approach to Recipe Text Processing, *Proceedings of the Cooking with Computers workshop*, pp. 29–34 (2012).
[27] 藤田早苗，平　博順，小林哲生，田中貴秋：絵本のテキストを対象とした形態素解析，自然言語処理，Vol. 21, No. 3, pp. 515–539 (2014).
[28] 土屋雅稔，宇津呂武仁，松吉　俊，佐藤理史，中川聖一：日本語複合辞用例データベースの作成と分析，情報処理学会論文誌，Vol. 47, No. 6, pp. 1728–1741 (2006).
[29] 堀　恵子，李　在鎬，長谷部陽一郎：機能語用例文データベース「はごろも」について，計量国語，Vol. 30, No. 5, pp. 275–285 (2016).
[30] 山崎誠，藤田保幸：現代語複合辞用例集，国立国語研究所 (2001).
[31] Jiang, W., Sun, M., Lü, Y., Yang, Y. and Liu, Q.: Discriminative Learning with Natural Annotations: Word Segmentation as a Case Study, *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, Sofia, Bulgaria, pp. 761–769 (2013).
[32] 鍛治伸裕，森　信介，高橋文彦，笹田鉄朗，斉藤いつみ，服部圭悟，村脇有吾，内海　慶：形態素解析のエラー分析，エラー分析ワークショップ (2015).
[33] 森　信介，小田裕樹：3 種類の辞書による自動単語分割の精度向上，自然言語処理，Vol. 18, No. 2, pp. 139–152 (2011).
[34] 松吉　俊，佐藤理史，宇津呂武仁：日本語機能表現辞書の編纂，自然言語処理，Vol. 14, No. 5, pp. 123–146 (2007).