J-UniMorph: 日本語の形態論における意味分類の体系化

松﨑孝介 [1]　谷口雅弥 [2]　乾健太郎 [3,1,2]　坂口慶祐 [1,2]
[1] 東北大学　[2] 理化学研究所　[3]MBZUAI
matsuzaki.kosuke.r7@dc.tohoku.ac.jp　masaya.taniguchi@riken.jp
kentaro.inui@mbzuai.ac.ae　keisuke.sakaguchi@tohoku.ac.jp

概要

我々は日本語の動詞における語形変化とその意味をまとめたデータセットを作成した．本データセット（J-UniMorph）は，世界中のあらゆる言語の語形に共通の基準で特徴ラベルを付与する UniMorph スキーマに準拠しており，107 種類の動詞に対して合計 12,533 組の語形と特徴ラベルを掲載している．UniMorph ラベルを付与した日本語のデータセットにより，他言語との比較・分析や言語横断的なデータの解析が見込めるほか，語形変化を意味から捉えられるようになるため，日本語学習者への支援など言語教育等のアプリケーション開発が期待できる．

github.com/cl-tohoku/J-UniMorph

1　はじめに

UniMorph は，世界中のあらゆる言語の語形に共通の特徴ラベルを付与するプロジェクトである．UniMorph では，以下のような「原形，語形，特徴ラベル」の 3 つ組のデータを，さまざまな単語について網羅的に生成する．

走る　走らない　V;PRS;IPFV;NEG

この例では，「走る」という原形に対して「走らない」が V（Verb：動詞），PRS（Present：現在），IPFV（Imperfective：未完了），NEG（Negative：否定）の特徴を有することを表す．このようなデータセットが，英語やフランス語をはじめ 169 以上の言語で作成されている [1)] にもかかわらず，人手で作成した日本語のデータセットはこれまで存在しなかった．

我々は，UniMorph の特徴ラベルと日本語の動詞の語形変化の対応を検討し，上記のような 3 つ組を 12,533 組生成した．そして，作成したデータセット（J-UniMorph と呼ぶ）の分析や，UniMorph では表せない日本語の表現についての検討を行った．

膠着語である日本語は，動詞の語幹に多種多様な接辞（接頭辞と接尾辞）や語尾を加えることで，さまざまなに語形が変化し，多様な意味を表現できる．本論文では，「た」「なかった」などの接辞や語尾あるいはそれらの組み合わせを**語形変化**と呼び，「走った」「走らなかった」など実際に語形変化した表現を**語形**と呼ぶ．

データセット作成の手順は，図 1 に示すように，(1) **語形の生成**，(2) **ラベルの決定**，(3) **フィルタリング**の大きく 3 つで構成される．(1) では，107 種類の動詞の原形に対して最大 125 個の語形を生成する．(2) では，それぞれの語形変化に適切な UniMorph ラベルを決定して，生成した語形に付与する．(3) では，Google の完全一致検索でのヒット件数を用いて，使用頻度が低いと考えられる語形を除去する．

このようなデータセットを整備することにより，日本語の語形変化の意味を国際的なスキーマに基づいた特徴ラベルに対応付けることができる．

2　関連研究

動詞の語形変化については，言語学的な分析・研究 [1, 2] がなされており，電子的な資源としては JUMAN [3] や MeCab [4] のような形態素解析器の辞書や，広く語形変化を扱った HaoriBricks3 [5]，日本語機能表現辞書 [6] などのリソースがある．HaoriBricks3 は，日本語文を合成するためのドメイン特化言語であり，どのような日本語文を合成するかを Ruby コードで記述したものである．日本語機能表現辞書は，機能表現（機能語と複合辞を合わせたもの）の辞書である．「にたいして」や「なければならない」のような複合辞を今後 UniMorph に登録する際には，利用できる可能性がある．

本研究の目的は，これら先行研究で蓄積された知見を国際的な共通スキーマに接続することで，言語横断的に利用できる資源を構築することである．

1）https://unimorph.github.io/

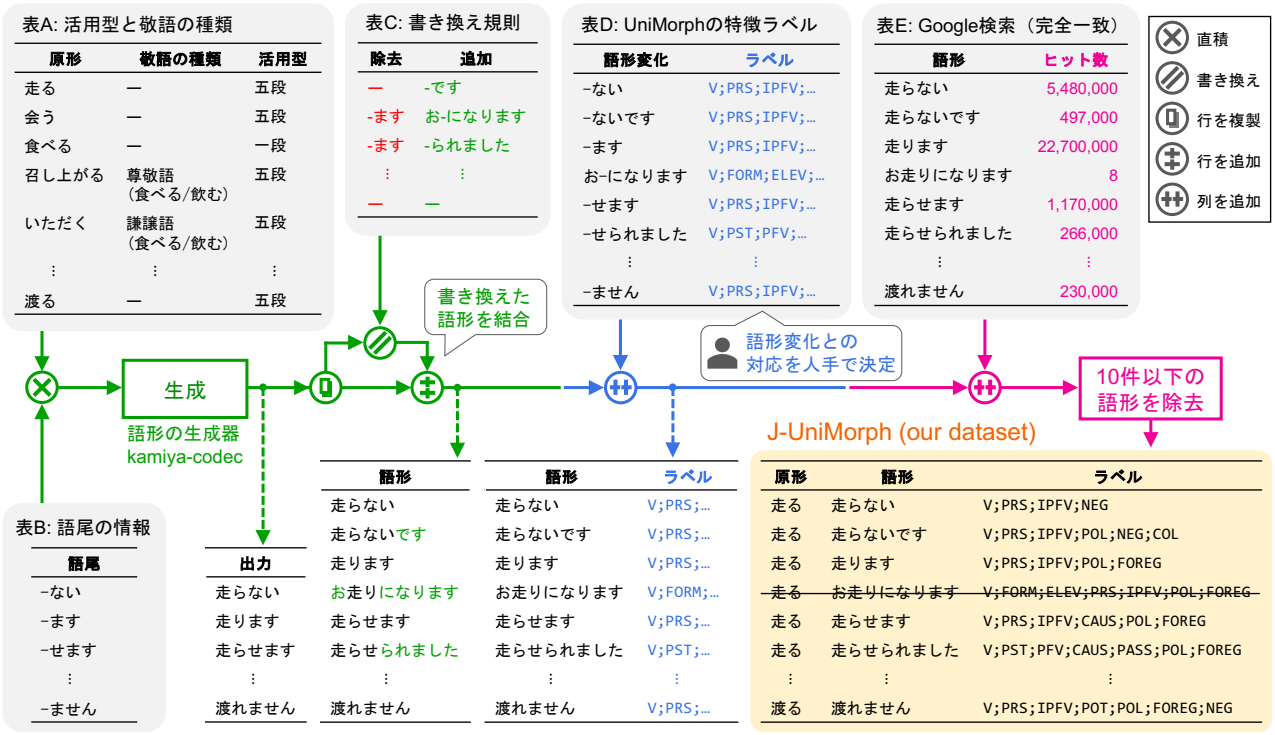

図 1　データセット作成の流れ．(1) 語形の生成：動詞の情報（表 A）と語尾の情報（表 B）をもとに，生成プログラム kamiya-codec を用いて語形を生成．加えて，出力の一部を書き換える（表 C）ことで，kamiya-codec では生成できない語形を追加．変格活用動詞は手動で生成．(2) ラベルの決定：UniMorph の特徴ラベルと語形変化の対応を決定（表 D）．各語形に付与．(3) フィルタリング：Google の完全一致検索でのヒット件数（表 E）が 10 以下であった語形を除去．

3　データセットの構成

以下では，本データセット（J-UniMorph）が準拠する UniMorph スキーマ [7] の概要を述べたあと，本データセットが扱う動詞の種類と語形変化や，特殊な敬語の扱いについて述べる．

3.1　UniMorph スキーマの概要

UniMorph は，23 の意味の次元と 212 を超える特徴ラベルで構成される [7]．意味の次元には品詞，人称，丁寧さ（politeness），時制（tense），法（mood），相（aspect），態（voice）などがあり，そのそれぞれに複数の特徴ラベルが用意されている．例えば，時制の次元には PRS（Present）や PST（Past）などの特徴ラベルがある．

UniMorph では語形単独の意味を扱い，副詞や感嘆詞など周辺の語を伴うことで生じる特殊な意味は扱わない．例えば，「あっ，ここに鍵があった．」のような過去形の特殊な用法（発見）[8] や，「もうすぐ学校に着く．」のような，副詞によって，言及する動作のタイミングが未来であると定まる用法が，扱わない意味に該当する．

3.2　生成する動詞の基準

本研究で構築するデータセットには，生成する動詞に関して以下の 2 つの基準を設けた．

(1) 日本語能力試験（JLPT）5 級程度の基本的な動詞 107 語 [2)]
(2) 敬語に関する書籍 [9] に掲載されていて，(1) の動詞と対応が取れる尊敬語 19 語，謙譲語 21 語

J-UniMorph では動詞の屈折（inflection）のみを扱い，品詞が変わる派生（derivation）は扱わない．なお，動作性名詞に軽動詞「する」を付けて動詞化したもの（「勉強する」等）は派生とみなす．

3.3　生成する語形変化の基準

日本語の動詞は，表 1 のように大きく 3 種類の活用型に分類される．[3)] 変格活用動詞は「来る」と「する」[4)] だけである．敬語の分類には，尊敬語・謙譲

2)　https://nihongoichiban.com/2011/08/21/list-of-all-verbs-for-the-jlpt-n5/
3)　学校文法では，一段活用を上一段活用と下一段活用に，変格活用はカ行変格活用とサ行変格活用に分けている．
4)　「する」には，「勉強する」のような動作性名詞のあとの軽動詞や，「愛する」「信ずる」などの「〜する・ずる」を含む．

表 1　活用型ごとの動詞の例

活用型による動詞の分類	例
五段活用動詞（I 型動詞）	会う，書く，走る
一段活用動詞（II 型動詞）	着る，食べる，見る
変格活用動詞（不規則に活用）	来る，する

語・丁寧語の 3 分類 [10] を採用する．

掲載する語形変化は，原形を含めた基本的な形，推量（〜だろう，でしょう），提案（〜よう，ましょう），願望（〜たい，たがる），命令（〜なさい，〜ください等），可能（〜られる等），受動（〜れる・られる），尊敬（〜れる・られる，お〜になる [5]），謙譲（お〜する [6]），許可の謙譲（〜せていただく），使役（〜せる・させる），縮約された使役（〜す・さす），使役・受動（〜せられる・させられる），縮約された使役・受動（〜される [7]）を基本とした．このそれぞれで「過去」「丁寧」「否定」の有無を変えた $2^3 = 8$ 通りの表現が可能か確認した．複数の意味がある語形は，その各々の特徴ラベルを掲載した．

登録を見送った主な語形変化には，補助動詞（〜ている，〜ておく等），ら抜き言葉，い抜き言葉，さ入れ言葉，疑問の表現（〜か），語彙的使役動詞（「寝る」に対して「寝かせる」等），複合辞（〜かもしれない等）がある．詳細な理由は付録 A に示すが，語形変化の数が膨大になることや，意味とラベルの対応が複雑であることが主な理由である．

以上により生成する語形変化を決定した結果，1 語あたりに生成する語形変化と特徴ラベルの組数は最大で，普通の動詞では 125 組，特殊な尊敬語では 102 組，特殊な謙譲語では 91 組となった．

3.4　特殊な敬語の扱い

尊敬語と謙譲語には「召し上がる」や「いただく」のように，敬意を表すために語そのものが変化する動詞がある．これらは語形の範囲を超えるが，特殊な敬語は原形自体に敬意が含まれると考え，原形は「食べる」などの対応する普通の動詞とした．

普通の動詞と特殊な敬語の対応は 1 対 1 に定まらないため，以下のようにそれぞれ掲載する．尊敬表現は `FORM;ELEV`（Formal，Referent Elevating）で表す．

食べる	召し上がった	`V;FORM;ELEV;PST;PFV`
飲む	召し上がった	`V;FORM;ELEV;PST;PFV`

5）「お〜くださる」など他の尊敬表現もあるが扱わない．
6）「お〜いたす」など他の謙譲表現もあるが扱わない．
7）縮約された使役・受動は五段活用でのみ生成する．

4　データセットの構築方法

データセットの作成手順は，(1) **語形の生成**，(2) **ラベルの決定**，(3) **フィルタリング**（使用頻度の低い語形の除去），の大きく 3 つであり，以下ではこれらの各手順について述べる．参照している表 B〜表 E は，図 1 中の表に対応している．

4.1　語形の生成

まず，UniMorph に掲載する語形変化を，日本語学習者向けの教科書 [11] などを参考にしながらリストアップし（表 B），kamiya-codec [8] をベースに各動詞の語形を生成した．kamiya-codec は，原形と語尾の情報を入力して語形を出力するコードであり，上記の教科書に基づいて作成されている．

その後，kamiya-codec の生成能力を超える語形を，作成した書き換えリスト（表 C）によって生成した．不規則動詞は手動で生成した．「召し上がる」などの特殊な敬語は，それを原形として生成したあと，前述の敬語対応に基づいて原形を置き換えた．

4.2　ラベルの決定

次に，UniMorph スキーマ [7] の特徴ラベルから，各語形変化に対応する特徴ラベルを決定した（表 D）．ラベルは，スキーマの記載内容や日本語の教科書 [8–13] を確認し，いくつかの動詞で意味や用法を吟味して決定した．

ラベル検討の際には，形態素ごとのラベル付与ではなく，語形全体の意味を考慮した．例えば，「食べないです」という表現は，語尾が「ない＋です」のように 2 つの形態素に分割されるが，否定（「ない」）と丁寧（「です」）の意味に加え，口語的な印象も含まれる．このように，語形全体によって新たな意味が生じることがある．

同じ語形変化には，活用型ごとに同じラベルを付与した．例えば，一段活用動詞の「れる・られる」形には可能，受動，尊敬の意味があるが，自動詞など，これらの意味が不自然な動詞もある．しかし，あえて削除はせずそのまま残している．

4.3　フィルタリング

最後に，生成した語形の整合性を確認するために，Google の検索結果を取得できるサービス [9] を

8）https://github.com/fasiha/kamiya-codec
9）https://serpapi.com/

使用して，完全一致検索でのヒット件数を調査した（表 E）．詳細な理由は付録 B に示すが，ヒット件数が 10 件以下であった語形は統計上信頼がおけないと判断し，除去した．加えて，日常で注意が必要な表現として 16 個を恣意的に除去した．[10)]

使用用途によっては，生成したすべての語形を扱いたいという場合が想定されるため，フィルタリング前のデータセットも公開している．

5 データセットの分析

多くの言語の語形変化に対応できるモデルを開発するコンペとして，UniMorph の原形と特徴ラベルから語形を予測する性能を競う Shared Task が，2016 年 [14] から毎年行われている．Wiktionary から自動生成された日本語のデータセットは，Shared Task 向けに公開された．[11)] これを Wiktionary 版と呼ぶ．

以下では，作成したデータセット（J-UniMorph）を Wiktionary 版と比較したあと，今後の方針やデータセットの応用先と貢献について述べる．

5.1 既存データセットとの比較

2023 年の Shared task で，Wiktionary 版を用いてモデルの評価が行われた [15]．J-UniMorph との内容構成の比較を表 2 にまとめる．

表 2 に示すように，J-UniMorph は Wiktionary 版に対して語形の総数 † で上回り，1 単語あたりの語形数 ♣ は約 10 倍である．したがって，J-UniMorph は多様な語形・表現を扱えている．

Wiktionary 版では動作性名詞 ♦ が全体の約 7 割を占めているが，J-UniMorph では動作性名詞を含めなかった．J-UniMorph は，語形が同じように変化する単独の「する」を掲載することで，動作性名詞の語形変化を網羅した．

5.2 J-UniMorph の今後

本データセット（J-UniMorph）に関する今後の方針や，考えられる応用先と貢献を述べる．

データセットの拡張　J-UniMorph の今後の拡張方針として，対象の動詞や語形変化を拡大することや，形容詞や形容動詞，名詞を追加することを検討している．なお，ラベル検討の際に，現状の UniMorph スキーマでは表せない日本語の表現を確認した．それらは進行相や複合辞などであり，一部を付録 C に示す．UniMorph スキーマの改善点については，UniMorph のプロジェクト側にラベルやスキーマの拡張を提案することを検討している．

応用　J-UniMorph の応用として，日本語学習者のための支援ツールの開発を検討している．これまでに開発された形態素解析器は形から意味を捉えるものであるが，日本語の語形変化がわからない日本語学習者は，意味を理解するうえで従来の日本語文法体系を使用することが難しい状況にある．しかし，本データセットを活用すれば，表現したい意味から語形を知ることができるようになる．現在，ユーザーが入力した語形に対応する UniMorph ラベルを抽出することで，学習者が語形の意味を調べたり，知っている語形から別の意味の表現を取得したりできるツールを試作している．

貢献　本研究が社会に与えるインパクトや貢献は，日本語の語形変化の意味を国際的なスキーマに基づいた特徴ラベルに対応付けることにある．UniMorph が多くの言語で整備されると，言語ごとに構築され汎用性がなかった解析システムを共通化できる．これにより，将来的に増加するであろう言語横断的なデータ（複数の言語が混在した文書）を解析する際の基盤技術になることが期待できる．

表 2　Wiktionary 版と J-UniMorph の内容構成

	Wiktionary 版			J-UniMorph
	Train	*Dev*	*Test*	*(Ours)*
掲載語形の総数 †	10,000	1,000	1,000	12,533
語形数 / 単語 ♣	12.5	10.0	10.0	117.1
単語数	800	100	100	107
動詞	25%	27%	30%	100%
動作性名詞 ♦	72%	69%	67%	0%
助詞を含むもの	3%	2%	3%	0%
副詞	1%	2%	0%	0%

6 おわりに

本論文では，日本語の動詞における語形変化とその意味をまとめたデータセットを UniMorph に準拠して作成し，他言語との分析や言語横断的なデータの解析の可能性について述べた．今後は，扱う語形変化と特徴ラベルを 12,533 組から増大させ，日本語学習者の支援ツールの開発を行う．本データセット（J-UniMorph）はオンラインで公開している．

10）　これらは「死ぬ」の語形変化であり，人間に対しては一般的に「亡くなる」という柔らかい印象の表現が使用される．したがって，「お死にになる」や「死なれる」などは尊敬語として適切でない．他にも実際は使用されない表現があるが，これらは特に気を付けるべきであり，特別に削除した．

11）　https://github.com/sigmorphon/2023InflectionST/

謝辞

本研究は，理化学研究所の基礎科学特別研究員制度の支援と，JSPS 科研費 JP22H00524, JP21K21343 の助成を受けたものです．本研究を進めるにあたり，東北大学坂口・乾・徳久研究室の吉田遥音氏，羽根田賢和氏に有益なコメントをいただきました．ここに深く感謝申し上げます．

参考文献

[1] 日本語記述文法研究会. 現代日本語文法 1. くろしお出版, 2010.
[2] 益岡隆志, 田窪行則. 基礎日本語文法 ―改訂版―. くろしお出版, 1992.
[3] 黒橋禎夫, 河原大輔. JUMAN. https://nlp.ist.i.kyoto-u.ac.jp/?JUMAN.
[4] 工藤拓. MeCab. https://taku910.github.io/mecab/.
[5] 佐藤理史. Haoribricks3: 日本語文を合成するためのドメイン特化言語. 自然言語処理, Vol. 27, No. 2, pp. 411–444, 2020.
[6] 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.
[7] John Sylak-Glassman. The composition and use of the universal morphological feature schema (unimorph schema). 2016.
[8] 日本語記述文法研究会. 現代日本語文法 3. くろしお出版, 2007.
[9] 平林周祐, 浜由美子. 外国人のための日本語 例文・問題シリーズ 10 敬語. 荒竹出版, 1988.
[10] 日本語記述文法研究会. 現代日本語文法 7. くろしお出版, 2009.
[11] 神谷妙子. The handbook of Japanese verbs. 講談社, 2001.
[12] 日本語記述文法研究会. 現代日本語文法 2. くろしお出版, 2009.
[13] 高見健一. 受身と使役 ―その意味規則を探る―. 開拓社, 2011.
[14] Ryan Cotterell, Christo Kirov, John Sylak-Glassman, David Yarowsky, Jason Eisner, and Mans Hulden. The SIGMORPHON 2016 shared Task—Morphological reinflection. In Micha Elsner and Sandra Kuebler, editors, Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology, pp. 10–22, Berlin, Germany, August 2016. Association for Computational Linguistics.
[15] Omer Goldman, Khuyagbaatar Batsuren, Salam Khalifa, Aryaman Arora, Garrett Nicolai, Reut Tsarfaty, and Ekaterina Vylomova. Sigmorphon–unimorph 2023 shared task 0: Typologically diverse morphological inflection. In Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology, pp. 117–125, 2023.
[16] 森田良行, 松木正恵. 日本語表現文型. アルク, 1989.

A　扱わない語形変化

3.2 節にて，扱わない語形変化があることを述べたが，ここではそれらの詳細を述べる．

補助動詞　「〜ている」や「〜ておく」など，動詞のあとに付けて意味を変える補助動詞の扱いは見送る．理由としては，補助動詞自身が語形変化を持つため生成数が指数関数的に増大すること（例：「書いている」は「いる」と同じだけ語形がある），意味変化があり複雑なこと（例：「いる」単独の意味と異なり「書いている」として用いると動作継続や結果継続を表す），複数の補助動詞を付けるとどの順番が文法的に正しいか判断が難しいこと（例：「書いてみておく」と「書いておいてみる」）が挙げられる．

インフォーマルな表現　日本語における口語的な表現として，ら抜き言葉，い抜き言葉，さ入れ言葉がある．これらは近年，口語で使う人が増えているため，今後，扱うことを検討している．

疑問の表現　疑問を表す接尾辞「か」は，意味がそのまま保たれるとは限らないため，扱いを見送る．例えば，「食べません」は否定を意味するが，「食べませんか？」とすると提案を表す．なお，口語では語尾のイントネーションを上げるだけで疑問表現になるため，扱い方は検討の必要がある．

語彙的使役動詞　これまでに述べたような「せる・させる」を付けて使役を表す動詞（迂言的使役動詞）以外にも，それ自体で使役過程とその結果生じる出来事の両方を表す他動詞（語彙的使役動詞）がある [13]．以下に例を示す．

(1) お母さんは，子供を寝させた．
(2) お母さんは，子供を寝かした／寝かせた．

(1) はこれまでに述べた使役形だが，(2) の「寝かした／寝かせた」の原形は「寝る」ではなく，「寝かす／寝かせる」という別の他動詞である [13]．そのため，これを動詞の語形変化とは捉えない．

複合辞　複合辞とは，いくつかの独立した意味を持つ語が複合して新たな意味を持つ表現のことである [16]．例えば，「かもしれない」は「かも＋しれ＋ない」と分析されるが，分けてしまうと「かもしれない」全体で表される固有の意味や機能を十分に説明できない．このような複合辞は多様な表現があって複雑なことに加え，付録 C で示すような，UniMorph の構造で表せる意味の範囲を超える表現もあるため，扱わない．

B　使用頻度が低い語形の除去基準

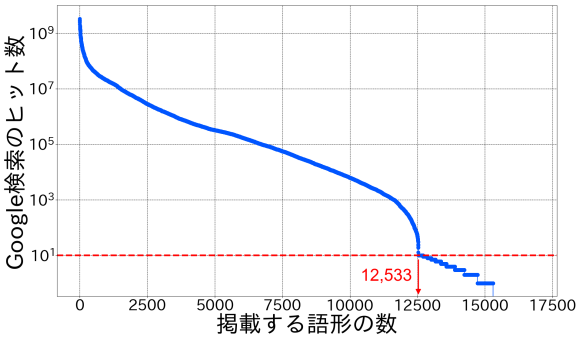

図 2　Google 検索のヒット数と掲載する語形の数の関係．赤い点線は我々が設定したしきい値（10^1）であり，傾向が変わる 10 以下の語形を除去した．矢印は J-UniMorph の語形の総数が 12,533 語であることを示している．

図 2 に，完全一致の検索ヒット数と単語数の関係を示す．この図が Long tail となっていることは，ヒット数が 10 件以下の語形が多く現れたことを示している．出現頻度の少ない語形が大量に現れたことは，統計分布上信頼がおけないものであるから，これらは掲載する語形から除外することにした．なお，我々は言語学的な信頼性よりも，統計的な信頼性による分布を採用した．

C　UniMorph の改善点

5.2 節で触れた今後の改善点の一部を説明する．

進行相　進行相を表す「ている」形に該当するラベルは確認できず，スキーマの改善が必要と考える．なお，UniMorph スキーマ [7] の PROG（Progressive）ラベルは起動相に用いられることが読み取れ，これは「つつある」の表現に対応すると我々は考えた．

複合辞　複合辞を扱う場合には，UniMorph では表せない順序を考慮したラベルが必要となる．以下はどちらも V;PST;PFV;LKLY（Verb，Past，Perfective，Likely）のラベルが適切と考えた例である．

(1) 彼はリンゴを食べたかもしれない。
(2) 彼はリンゴを食べるかもしれなかった。

(1) からは「彼はリンゴを食べた可能性が高い」ことが想像され，LKLY が全体の意味を握る．また，主語を「私」にすると「私はリンゴを食べた事実を忘れた」ことが想像される．(2) からは「彼はリンゴを食べる予定だったが，何かしらの原因でそうならなかった」ことが想像され，PST;PFV が全体の意味を握る．このように，ラベルの優先順位や文脈によって意味が変化する例があり，改善が必要と考える．