RNN はラテン語からロマンス語への活用変化を再現するか？

川崎義史
東京大学
ykawasaki@g.ecc.u-tokyo.ac.jp

概要

本稿では，ラテン語からロマンス語への発達における動詞活用の変化を計算機上で再現し，計算歴史言語学の見地から分析を行った。具体的には，系列変換モデルにラテン語の動詞活用を学習させ，その出力形を分析した。その結果，以下のことが判明した：(1) 正解率の分布は予測と合致した；(2) 正解率は，ラテン語からロマンス語への変化の大きさと負の相関を示した；(3) 誤出力の中にはロマンス語を彷彿させる語形が見られた。

1 はじめに

ラテン語は，名詞・形容詞の曲用と動詞の活用を持つ屈折語である [1]。名詞は数・格に応じて $2\times6=12$ 変化，形容詞は性・数・格に応じて $3\times2\times6=36$ 変化する。動詞は，法・態・時制・人称・数に応じて一層複雑に活用する。能動態では，表 1 のチェックマークの記された法と時制において，主語の人称（1 人称，2 人称，3 人称）と数（単数，複数）に応じて最大 $3\times2=6$ 変化する。受動態でも同様に活用する。つまり，1 つの動詞は，能動態・受動態の直説法 6 時制と接続法 4 時制だけで，$2\times(6+4)\times6=120$ 変化する。さらに，命令法・不定詞・分詞の形も存在する。例として，表 2 は，AMĀRE「愛する」の直説法・能動態・現在と完了の活用を示している[1)]。太字は強勢母音を，¯は長母音を表す。

不規則動詞は除き，4 種類の動詞活用クラスがあり，それぞれ独自の語尾変化を持つ[2)]。動詞の正確な活用には，直説法・能動態・現在・1 人称単数，不定詞・能動態・現在，直説法・能動態・完了・1 人称単数，完了受動分詞・中性・単数の 4 つの語形が必須となる。例えば，AMĀRE の場合は，順に AMŌ, AMĀRE, AMĀVĪ, AMĀTUM となる。これらの語形の間には一定の対応関係が存在するものの，一般的には，ある形から他の形を導出することはできない。

表 1 ラテン語の能動態の動詞活用：チェックマークは該当する法や時制の活用が存在することを表す。

	直説法	接続法	命令法	不定詞	分詞
現在	✓	✓	✓	✓	✓
未完了	✓	✓			
未来	✓		✓	✓	✓
完了	✓	✓		✓	
過去完了	✓	✓			
未来完了	✓				

表 2 AMĀRE「愛する」の直説法・能動態・現在（左）と完了（右）の活用表：太字は強勢母音を，¯は長母音を表す。

直説法 能動態	現在		完了	
	単数	複数	単数	複数
1 人称	**AMŌ**	**AMĀMUS**	**AMĀVĪ**	**AMĀVIMUS**
2 人称	**AMĀS**	**AMĀTIS**	**AMĀVISTĪ**	**AMĀVISTIS**
3 人称	**AMAT**	**AMANT**	**AMĀVIT**	**AMĀVĒRUNT**

ラテン語から派生したイタリア語，スペイン語，フランス語，ポルトガル語，ルーマニア語等の言語は**ロマンス語**と総称される [2][3)]。ラテン語の直説法・能動態・未来や受動態の一部の活用などは，ロマンス語に継承されずに消滅した。その代わりに，ロマンス語は，未来や受動を表現するための独自の活用を発達させた [4]。ロマンス語に継承された活用では，類推作用 [5, 6, 7] により平準化が進み，活用形の不規則性は減少した。活用の消滅や平準化の進展などについて事後的な記述は可能であるが，これらの変化の蓋然性は自明ではない。

そこで，本稿では，「話者により体系が作られる」という考え [8] に基づき，ラテン語からロマンス語への発達における動詞活用の変化を計算機上で再現し，**計算歴史言語学**の見地から分析を行う。具体的には，Recurrent Neural Network（RNN）を利用する系列変換モデル [9] を話者に見立て，ラテン語の動詞活用の形態論を学習させる。そして，学習済みモ

1) 本稿では，能動態・現在の不定詞で動詞を代表させる。
2) 厳密には，第 3 活用には a と b の 2 種類がある。
3) 正確には，ロマンス語は，文語の（古典）ラテン語ではなく口語の俗ラテン語 [3] から発達したものである。しかし，後者の文字資料は存在しない。そのため，本稿では，前者をロマンス語の母体とみなして分析を行う。

デルの出力形の分析を行い，下記の問いに答える：

- 学習が困難な動詞や活用は存在するか？
- 各活用の正解率は，ラテン語からロマンス語への発達における変化の大きさを反映するか？
- 誤出力に何らかの特徴が見られるか？

本稿の構成は次の通りである。2 節で関連研究を概観する。3 節で手法の解説，4 節で実験設定の説明を行う。5 節で考察を行い，6 節で結論と今後の課題を述べる。

2 関連研究

本稿は，Paradigm Cell Filling Problem (PCFP)[10] と関連している。PCFP とは，名詞や動詞の活用形の一部が与えられた状況で，未知の活用形を予測するタスクである。[11, 12] は，系列変換モデルでこのタスクに取り組んだ。複数の言語での実験結果を報告しているが，出力結果の詳細な分析はない。

[13] は，英語の不規則な過去形の通時的な規則化 (-ed の付加) を計量的に分析し，規則化率 $\propto$ 使用頻度$^{-0.5}$ となることを発見した。[14] は，系列変換モデルで英語の過去形の習得をモデル化し，習得過程で見られる U 字カーブの再現を報告している。

[15] は，系列変換モデルで「言語もどき」の語順の学習と世代間継承をモデル化した。創発するパターンを分析し，長距離依存を嫌う傾向を発見した。対象は自然言語ではないが，系列変換モデルで言語変化をモデル化した点で特筆に値する。

本稿では，ラテン語の動詞活用全体を系列変換モデルに学習させ，ロマンス語とも対照しつつ，学習済みモデルの振る舞いを詳細に分析する。

3 手法

系列変換モデルに，ラテン語の動詞活用の形態論を学習させる。符号化器には Bi-LSTM，復号器には注意機構付きの LSTM を用いる。不定詞と活用形の形態情報を入力し，文字単位で活用形を出力させる。形態情報は，法/態/時制/人称・数を連結したものとする。文字と同様に，不定詞と形態情報も 1 つのトークンとする。図 1 に手法概略図を示す。例えば，`ama:re`，`ind/act/pres/2pl` ↦ `ama#a:tis` のような変換を学習させる。形態情報の `ind/act/pres/2pl` は，直説法/能動態/現在/2 人称複数を表す。また，`:` は長音を，`#`は直後の音節に強勢があることを表す記号とする。

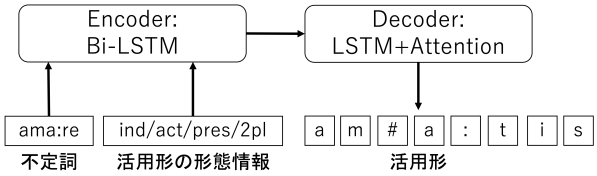

図 1 手法概略図：系列変換モデルにラテン語の動詞活用を学習させる。不定詞と活用形の形態情報を入力し，文字単位で活用形を出力させる。符号化器には Bi-LSTM，復号器には注意機構付きの LSTM を用いる。

本稿の狙いは，正解率の向上ではなく，学習が進んだ時点でのモデルの振る舞いの分析である。そのため，検証データの正解率が一定の値に達するまで訓練データで学習を行う。出力形と正解形が完全に一致する場合のみ，正解とする。その後，モデルが産出するテストデータの出力形を詳細に分析する。

4 実験

データセットは 2 万件の入出力ペアとし，8 : 1 : 1 で訓練データ，検証データ，テストデータに分割した。使用する動詞は，[16] 収録の 586 種類とした。母音の長短は [17] に従った。古典語コーパス PhiloLogic4[4] における使用頻度に基づいて動詞をサンプリングした。ただし，同一データの重複を避けたため，コーパス頻度に比べて滑らかな分布となった。法は，直説法：接続法：命令法：不定詞：分詞 = 6 : 4 : 1 : 1 : 1 の割合でサンプリングした。直説法の 6 時制，接続法の 4 時制に対し，命令法・不定詞・分詞の活用は一部の時制や人称・数に限られるため，割合を 1 とした（表 1 参照）。ただし，上記の割合はコーパス頻度を反映しておらず，暫定的なものである。態，時制，人称・数は，いずれのカテゴリも一様分布からサンプリングした。

埋め込み次元は 32，隠れ状態の次元は 64，エポック数は 100 とした。分析には一定量の誤出力を必要とするため，検証データの正解率は 9 割程度に抑えた。上記のパラメータ設定は，この基準を辛うじて満たすものだった。検証データとテストデータの正解率は，それぞれ，0.889，0.892 だった。

5 考察

本節では，テストデータの出力に基づき分析を行う。仮説検定の有意水準は 0.05 とする。

形式受動態動詞 ラテン語には，意味は能動だが活用は受動態を用いる形式受動態動詞と呼ば

4) https://artfl-project.uchicago.edu/philologic4

表 3 各形態情報（法/態のレベル）の正解率

法/態	正解率
直説法/能動態	0.891 (562/631)
直説法/受動態	0.949 (430/453)
接続法/能動態	0.905 (294/325)
接続法/受動態	0.953 (243/255)

表 4 未完了系・完了系の正解率

法/態/系	正解率
直説法/能動態/未完了系	0.865 (294/340)
直説法/能動態/完了系	0.921 (268/291)
直説法/受動態/未完了系	0.919 (216/235)
直説法/受動態/完了系	0.982 (214/218)
接続法/能動態/未完了系	0.845 (131/155)
接続法/能動態/完了系	0.959 (163/170)
接続法/受動態/未完了系	0.915 (108/118)
接続法/受動態/完了系	0.985 (135/137)

れる動詞が存在する [1]。これらの動詞の態は，意味に基づき能動態とした。形式受動態動詞の正解率は 0.559 (62/111) で，それ以外の動詞の正解率 0.912 (1723/1889) に比べ，有意に低かった ($p < 0.001$)。全体に占める割合の小ささと「意味は能動だが活用は受動態」という特殊性が，学習を困難にしていると考えられる。そのために，形式受動態動詞は，その他の動詞と同様の能動態動詞に変転しロマンス語に継承されたと推定される [4]。これ以降，形式受動態動詞は分析から除外する。

学習困難な活用　表 3 に，各形態情報（法/態のレベル）の正解率を示す。法は直説法と接続法に限定した。いずれの態でも，直説法と接続法の正解率に有意な差は見られなかった。一方，いずれの法でも，能動態の正解率は受動態よりも有意に低かった（直説法では $p < 0.001$，接続法では $p = 0.014$）。能動態の正解率の低さは，学習困難な形式受動態動詞や不規則動詞に起因すると考えられる。

次に，未完了系と完了系の振る舞いを分析する。ラテン語の動詞活用は，未完了系（現在・未完了・未来）と完了系（完了・過去完了・未来完了）に二分される（表 1 参照）。いずれの態でも，未完了系は，活用クラスごとに活用語尾が異なる。一方，完了系は，活用の仕様が全活用クラスで共通である。そのため，未完了系の方が完了系よりも学習が困難だと予想される。表 4 に示す実験結果は，この予想を支持する。直説法/能動態 ($p = 0.011$)，直説法/受動態 ($p = 0.001$)，接続法/能動態 ($p = 0.001$)，接続法/受動態 ($p = 0.004$) の全てで，未完了系の正解率は完了系よりも有意に低かった。

表 5 は，各形態情報（法/態/時制のレベル）の正解率を示している。この結果を，ラテン語からロマンス語への発達の過程で生じた形態的変化の大きさと対照してみたい。一つの方策として，[4] で各活用の説明に割かれている紙数を形態的変化の大きさの代理変数とみなすことにする[5)]。大きな変化を経た活用ほど，割かれる紙数も大きくなると予想される。分析対象は，ロマンス語に継承された以下の活用とした：直説法/能動態/現在，直説法/能動態/未完了，直説法/能動態/完了，直説法/能動態/過去完了，直説法/能動態/未来完了，接続法/能動態/現在，接続法/能動態/完了，接続法/能動態/過去完了，不定詞/能動態/現在，分詞/受動態/完了。ロマンス語に継承されなかった活用（直説法・能動態・未来や一部の受動態など）やラテン語に対応するものがない活用（ロマンス語の直説法・能動態・未来など）は除外した。正解率と紙数の順位相関係数は $\rho = -0.728$ ($p = 0.017$) で，統計的に有意な強い負の相関を示した。これは，学習困難なラテン語の活用ほど，ロマンス語発達の過程で大きな変化を受けたことを示唆する。この変化の実体は，類推作用による活用形の平準化である [4]。モデルの出力が，現実の通時変化の傾向と合致したことは興味深い。

しかしながら，正解率が高い活用が必ずしもロマンス語に継承されるわけではない。例えば，直説法・能動態・未来，接続法・能動態・未完了や受動態の正解率は低くはないが，ロマンス語に継承されなかった。これらの活用の消失は，形態面の学習困難性では説明できず，言語内外の他の要因の検討が必要となる。

活用クラス　表 6 は，各活用クラスの正解率を示している。活用 4 クラス間の正解率に有意な差は見られなかった。一方，不規則動詞の正解率は，いずれの活用クラスよりも，有意に低かった ($p < 0.01$)。

訓練データ頻度　各動詞の訓練データにおける出現頻度と正解率の順位相関係数は $\rho = 0.205$ ($p < 0.001$) で，有意な正の相関が見られた。低頻度の動詞ほど学習困難という傾向は直感と合致する。

一方，訓練データにおける形態情報（法/態/時制のレベル）の頻度と正解率の順位相関係数は $\rho = 0.134$ ($p = 0.436$) で，有意な相関は見られなかった。これは，頻度とは独立に，学習が容易（困難）な形態情報が存在することを示唆する。

5) 実際には，音変化により誘発された変化も多く [4]，純粋に形態的変化のみを定量化することは困難である。

ロマンス語形の出現　誤出力の中には，俗ラテン語やロマンス語を彷彿させる下記の語形が散見された。これらの語形はラテン語としては誤りであるが，ロマンス語としては尤もらしいものである。ラテン語の動詞活用を学習する過程で，モデルは類推作用 [5, 6, 7] を獲得し，これらの語形を生み出したと言える。つまり，ラテン語内部にロマンス語を生み出す契機が胚胎されていることを示唆している。

- 形式受動態動詞の能動動詞化：LOQUĪ「話す」の直説法/能動態/現在/2 人称単数の出力形 l#oquis（正解は l#oqueris）
- 不規則動詞の規則化：POSSE「できる」の直説法/能動態/未完了/3 人称複数の出力形 pot#e:bant（正解は p#oterant）
- 第 2 活用と第 3a 活用の混同：EXCĒDERE「去る」の直説法/能動態/現在/2 人称複数の出力形 exce:d#e:tis（正解は exc#e:ditis）
- 第 3a 活用と第 3b 活用の混同：DĒFICERE「不足する」の直説法/能動態/現在/3 人称複数の出力形 d#e:ficunt（正解は de:f#iciunt）
- 強勢位置の移動：EXCIPERE「除外する」の直説法/能動態/現在/3 人称単数の出力形 exc#ipit（正解は#excipit）

6　おわりに

本稿では，ラテン語からロマンス語への発達における動詞活用の変化を計算機上で再現し，計算歴史言語学の見地から分析を行った。具体的には，系列変換モデルにラテン語の動詞活用を学習させ，その出力形を分析した。その結果，以下のことが判明した：(1) 正解率の分布は予測と合致した；(2) 正解率は，ラテン語からロマンス語への変化の大きさと負の相関を示した；(3) 誤出力の中にはロマンス語を彷彿させる語形が見られた。

今後の課題の 1 つは，より現実を模倣した設定下で実験を行うことである。例えば，各形態情報の実際の使用頻度や，母音の長短などの音韻的区別の喪失 [4] を反映させることが考えられる。そのような設定下で実験を行うことで，新たな知見が得られる可能性がある。また，本稿の手法を発展させて，統語的変化や方言分化などの現象を計算歴史言語学の見地から検証することも興味深いと思われる。

表 5　各形態情報（法/態/時制のレベル）の正解率

能動態	
法/態/時制	正解率
直説法/能動態/現在	0.759 (88/116)
直説法/能動態/未完了	0.965 (110/114)
直説法/能動態/未来	0.873 (96/110)
直説法/能動態/完了	0.859 (85/99)
直説法/能動態/過去完了	0.918 (90/98)
直説法/能動態/未来完了	0.989 (93/94)
接続法/能動態/現在	0.793 (69/87)
接続法/能動態/未完了	0.912 (62/68)
接続法/能動態/完了	0.976 (83/85)
接続法/能動態/過去完了	0.941 (80/85)
命令法/能動態/現在	0.833 (5/6)
命令法/能動態/未来	0.944 (17/18)
不定詞/能動態/現在	0.870 (20/23)
不定詞/能動態/未来	0.667 (12/18)
不定詞/能動態/完了	0.923 (12/13)
分詞/能動態/現在	0.833 (20/24)
分詞/能動態/未来	0.833 (10/12)
分詞/能動態/動名詞	0.938 (15/16)
分詞/能動態/目的分詞	0.929 (13/14)
受動態	
法/態/時制	正解率
直説法/受動態/現在	0.878 (72/82)
直説法/受動態/未完了	1.000 (80/80)
直説法/受動態/未来	0.877 (64/73)
直説法/受動態/完了	0.984 (60/61)
直説法/受動態/過去完了	0.977 (84/86)
直説法/受動態/未来完了	0.986 (70/71)
接続法/受動態/現在	0.889 (56/63)
接続法/受動態/未完了	0.945 (52/55)
接続法/受動態/完了	0.984 (62/63)
接続法/受動態/過去完了	0.986 (73/74)
命令法/受動態/現在	0.833 (5/6)
命令法/受動態/未来	0.667 (10/15)
不定詞/受動態/現在	0.900 (9/10)
不定詞/受動態/未来	1.000 (14/14)
不定詞/受動態/完了	1.000 (15/15)
分詞/受動態/完了	1.000 (10/10)
分詞/受動態/動形容詞	0.636 (7/11)

表 6　各活用クラスと不規則動詞の正解率

活用クラス	正解率
第 1 活用	0.932 (354/380)
第 2 活用	0.913 (283/310)
第 3a 活用	0.928 (734/791)
第 3b 活用	0.910 (142/156)
第 4 活用	0.896 (103/115)
不規則動詞	0.781 (107/137)

謝辞

本研究は JSPS 科研費 JP18K12361 の助成を受けたものです

参考文献

[1] Renato Oniga. **Latin: A Linguistic Introduction**. Oxford University Press, 2014.

[2] Adam Ledgeway and Martin Maiden, editors. **The Oxford Guide to the Romance Languages**. Oxford University Press, 2016.

[3] Ralph Penny. **A History of the Spanish Language**. Cambridge University Press, 2002.

[4] Ti Alkire and Carol Rosen. **Romance Languages: A Historical Introduction**. Cambridge University Press, 2010.

[5] Javier Elvira. **El cambio analógico**. Gredos, 1998.

[6] James P. Blevins and Juliette Blevins. **Analogy in Grammar: Form and Acquisition**. Oxford University Press, 2009.

[7] Joan Bybee. **Language Change**. Cambridge University Press, 2015.

[8] Eugenio Coseriu. **Sincronía, diacronía e historia: El problema del cambio lingüístico**. Gredos, tercera ed edition, 1988.

[9] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. Vol. 27. Curran Associates, Inc., 2014.

[10] Farrell Ackerman, James P. Blevins, and Robert Malouf. Parts and wholes: Implicative patterns in inflectional paradigms, 2009.

[11] Robert Malouf. Abstractive morphological learning with a recurrent neural network. **Morphology**, Vol. 27, p. 431–458, 11 2017.

[12] Miikka Silfverberg and Mans Hulden. An encoder-decoder approach to the paradigm cell filling problem. p. 2883–2889. Association for Computational Linguistics, 2018.

[13] Erez Lieberman, Jean Baptiste Michel, Joe Jackson, Tina Tang, and Martin A. Nowak. Quantifying the evolutionary dynamics of language. **Nature**, Vol. 449, pp. 713–716, 10 2007.

[14] Christo Kirov and Ryan Cotterell. Recurrent neural networks in linguistic theory: Revisiting pinker and prince (1988) and the past tense debate. **Transactions of the Association for Computational Linguistics**, Vol. 6, pp. 651–666, 2018.

[15] Rahma Chaabouni, Eugene Kharitonov, Alessandro Lazaric, Emmanuel Dupoux, and Marco Baroni. Word-order biases in deep-agent emergent communication. p. 5166–5175. Association for Computational Linguistics, 2019.

[16] 有田潤. ラテン語基礎 1500 語. 大学書林, 1957.

[17] 水谷智洋（編）. 羅和辞典〈改訂版〉. 研究社, 2009.