

深層学習時代のコーパス分析： コーパス検索では見えないこと、 深層学習で見えること

英語コーパス学会 第49回大会（30周年記念大会）
2023年9月9日（土）@関西大学

東京大学大学院 総合文化研究科 言語情報科学専攻
川崎 義史

ykawasaki@g.ecc.u-tokyo.ac.jp

応用例2

- 第一言語習得
- 第二言語習得
- 通時意味論
- 通時形態論
- まとめ

応用例2

- 第一言語習得
- 第二言語習得
- 通時意味論
- 通時形態論
- まとめ

Recurrent Neural Networks in Linguistic Theory: Revisiting Pinker and Prince (1988) and the Past Tense Debate

Christo Kirov

Johns Hopkins University
Baltimore, MD

Ryan Cotterell

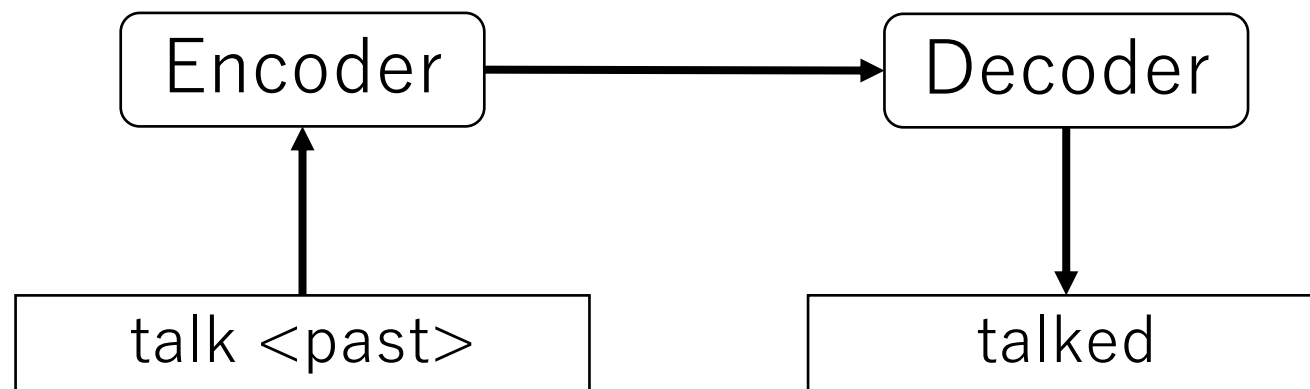
Johns Hopkins University
Baltimore, MD

深層学習で見えること
L1習得の認知モデリング

Past tense mapping problem

come	->	came
talk	->	talked
pad	->	padded

- Rumelhart & McClelland (1986) : ニューラルモデル提案
- Pinker & Prince (1988) : 実証性が不十分と反駁
- 本論文Kirov & Cotterell (2018) : Encoder-Decoder modelを用いた実験を行い, ニューラルモデルの利用可能性を再提示



Kirov & Cotterell (2018)

- 全体正解率95%（規則動詞99%，不規則動詞29%）
 - R&Mでは67%
- 人間に類似した振る舞いを示す
 - ミクロU字型学習曲線（came~comedの揺れ）を示すことがある
 - 不規則形と規則形の融合形（eat -> ated）を生成しない

→ニューラルモデルを認知モデリング利用できる（かも）！

応用例2

- 第一言語習得
- 第二言語習得
- 通時意味論
- 通時形態論
- まとめ

Second Language Acquisition of Neural Language Models

Miyu Oba¹ **Tatsuki Kuribayashi^{2,3}** **Hiroki Ouchi^{1,4}** **Taro Watanabe¹**

¹Nara Institute of Science and Technology

²MBZUAI ³Tohoku University ⁴RIKEN

深層学習で見えること
L2習得の認知モデリング

(Oba+ 2023) 1

- RQ : L1 (仏独露日) はL2 (英語) 習得にどう影響するか？
- 手法 : L1を学習した言語モデル (BERT) にL2を学習させる
→ L2に関する文法テスト (形態, 統語, 意味) を実施
- 人間が言語習得に必要とされる量のデータのみを使用し, 認知モデリングを目指す

a. Many teenagers were helping **themselves**.
b. *Many teenagers were helping **herself**.



(Oba+ 2023) 2

L1	L2の文法テストの全体正解率 (事前学習の効果)	文法 カテゴリー	事前学習の効果
	58.0% (+5.3%)	形態	+3.5%
	61.1% (+5.2%)	統語	+4.8%
	52.8% (+0.7%)	意味	+1.6%
	56.2% (+1.5%)		

- L1の知識（事前学習）は，L2の習得に役立つ
- 仏独の知識は露日よりも，英語学習に効果的
- L1の知識は，意味よりも形態や統語の学習に効果的
- L1による差は少ない？
- 人間の振る舞いとは必ずしも一致しないが，面白いアプローチ！

応用例2

- 第一言語習得
- 第二言語習得
- 通時意味論
- 通時形態論
- まとめ

Revisiting Statistical Laws of Semantic Shift in Romance Cognates

Yoshifumi Kawasaki

The University of Tokyo

y Kawasaki@g.ecc.u-tokyo.ac.jp

Maëlys Salingre

Mejiro University

m.salingre@mejiro.ac.jp

Marzena Karpinska

University of Massachusetts Amherst

mkarpinska@cs.umass.edu

Hiroya Takamura

AIST

takamura.hiroya@aist.go.jp

Ryo Nagata

Konan University

nagata-coling2022@ml.hyogo-u.ac.jp

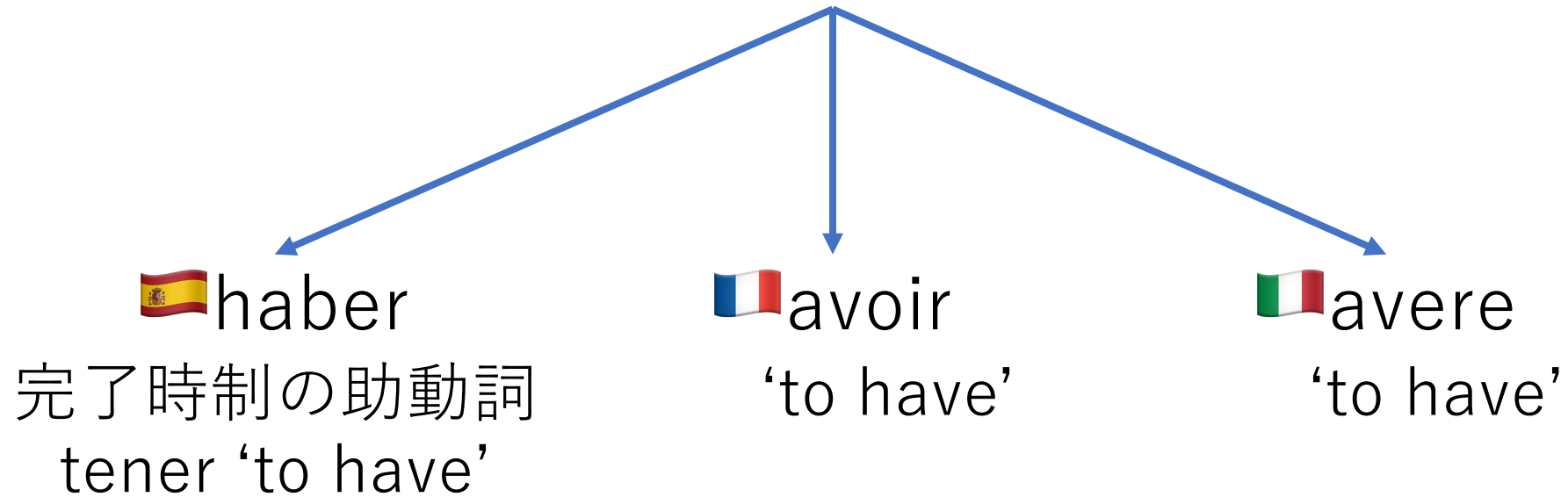
深層学習で見えること
単語間の意味の類似度

俗ラテン語から派生
した言語の総称



ロマンス語同源語

Lat. habere 'to have'



🇫🇷 avoirと🇮🇹 avereは似ているが、🇪🇸 haberはやや異なる

研究の目的

- ロマンズ語同源語の意味・用法の近さに影響する言語的要因を回帰により分析する
- 従属変数：ロマンズ語同源語間の非類似度（単語ベクトル間のコサイン距離）
- 説明変数：頻度，多義性，単語長，編集距離などの言語的要因

アウトライン

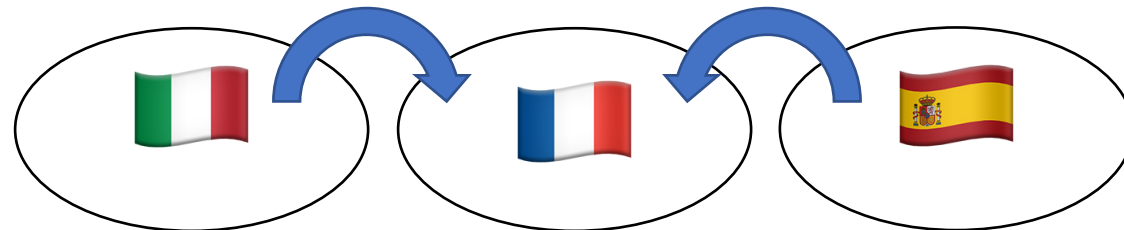
- 多言語単語ベクトル
- 回帰分析
- 考察
- まとめ

アウトライン

- 多言語単語ベクトル
- 回帰分析
- 考察
- まとめ

多言語単語ベクトルの獲得

- 使用言語：フランス語，イタリア語，スペイン語
- 各言語の単語ベクトルはWikipediaで学習
- ある言語（仏）の空間に他の言語（伊西）の単語ベクトルを写像



アウトライン

- 多言語単語ベクトル
- 回帰分析
- 考察
- まとめ

従属変数

ロマンス語同源語ペアのコサイン距離の対数

$$DIST_{rom} \equiv \log(1 - \cos(cog1, cog2))$$

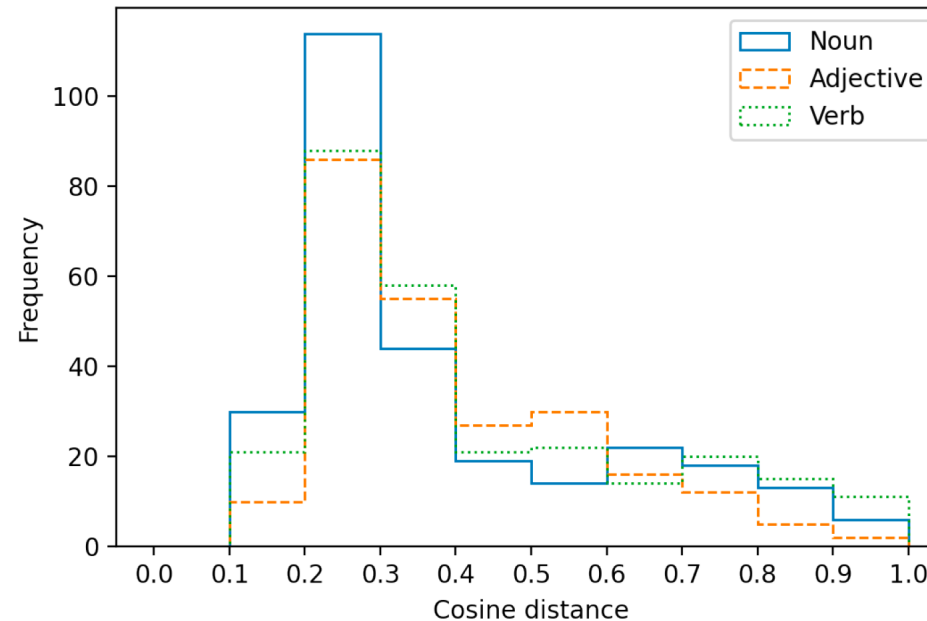


Figure 3: Distribution of cosine distance scores for French–Spanish cognate pairs in different word classes.

説明変数

ラテン語に関する言語的要因

- $FREQ_{lat}$: 語源の相対頻度の対数
- $POLY_{lat}$: 語源の語義数の対数
- LEN_{lat} : 語源の単語長（形態的複雑性の指標）

ロマンス語に関する言語的要因

- $FREQ_{rom}$: 同源語ペアの相対頻度の調和平均の対数
- $NORM_{rom}$: 同源語ペアの単語ベクトルのノルムの算術平均
- $EDIT$: ラテン語語源と同源語ペアの正規化編集距離の算術平均（継承語・借用語の指標）

アウトライン

- 多言語単語ベクトル
- 回帰分析
- 考察
- まとめ

考察

	Coef.	SE	t	$p > t $	
Intercept	0.00	0.03	0.00	1.00	
FREQ _{lat}	-0.08	0.04	-1.82	0.07	
POLY _{lat}	0.10	0.04	2.28	0.02	: 多義語→意味変化 大
LEN _{lat}	-0.21	0.03	-6.29	0.00	: 単語長大→意味変化 小
FREQ _{rom}	-0.54	0.03	-18.40	0.00	: 頻度大→意味変化 小
NORM _{rom}	—	—	—	—	
EDIT	0.13	0.03	4.07	0.00	: 継承語→意味変化 大

Table 3: Results of regression analysis on distance scores of French–Spanish cognate pairs ($N = 794$, $Adj.R^2 = 0.35$). NORM_{rom} was kept out by model selection methods.

アウトライン

- 多言語単語ベクトル
- 回帰分析
- 考察
- まとめ

まとめ

- ロマンズ語同源語の意味・用法の近さに影響する言語的要因を回帰により分析
- 単語ベクトルを用いることで、単語間の意味・用法の類似度を連続変数として扱える（→分析の幅が広がる）

応用例2

- 第一言語習得
- 第二言語習得
- 通時意味論
- 通時形態論
- まとめ

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

深層学習で見えること
通時的言語変化のモデリング

ラテン語の動詞活用

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， $\bar{\text{}}$ 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

$$2\text{態} \times (\text{直説法6時制} + \text{接続法4時制}) \times (3\text{人称} \times 2\text{数}) \\ = 120 \text{ 変化} + \text{不定詞} \cdot \text{分詞}$$

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。 27

俗ラテン語から派生
した言語の総称



ロマンス語の動詞活用

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	✓				

→未来や受動を表現するロマンス語独自の活用が発達

→ロマンス語に継承された活用でも類推作用により平準化が進む

ラテン語からロマンス語への変化は予測・再現できるか？

研究の目的

- 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。
 - 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？

アウトライン

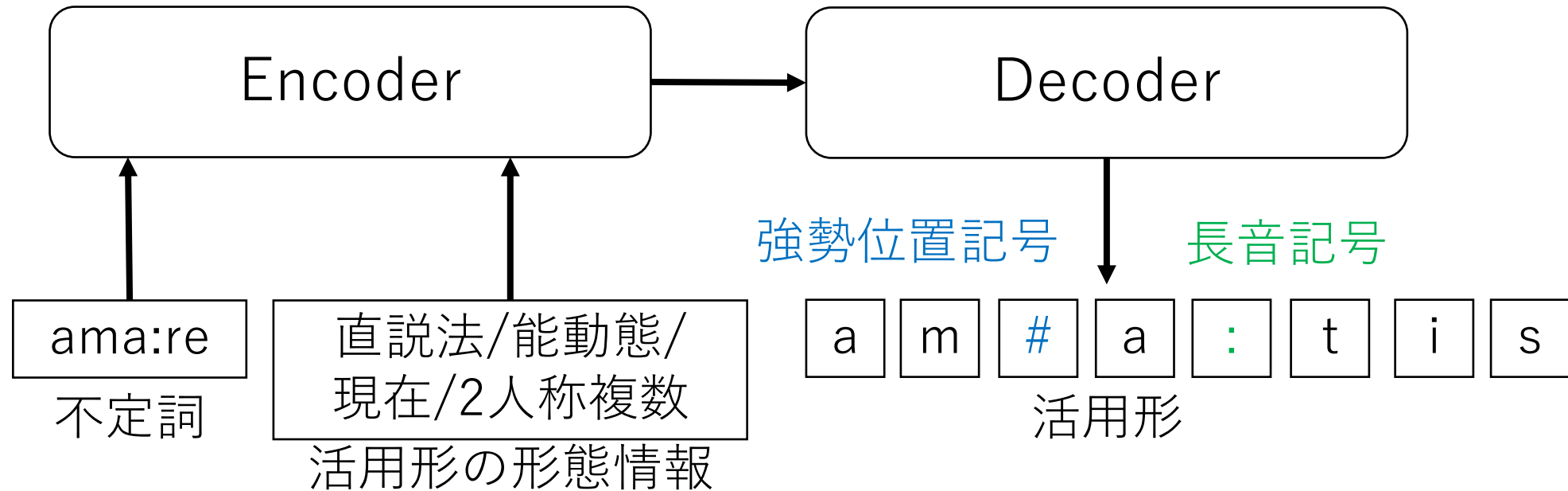
- 手法 & 実験
- 考察
- まとめ

アウトライン

- 手法 & 実験
- 考察
- まとめ

手法 & 実験

「不定詞 + 活用形の形態情報 → 活用形」
の写像をRNNでモデル化



コーパス頻度に基づき動詞（約600種類）をサンプリング
→ 検証データ正解率を約90%に抑えて，出力形を分析

アウトライン

- 手法 & 実験
- 考察
- まとめ

意味は能動だが活用は受動態を用いる動詞

形式受動態動詞

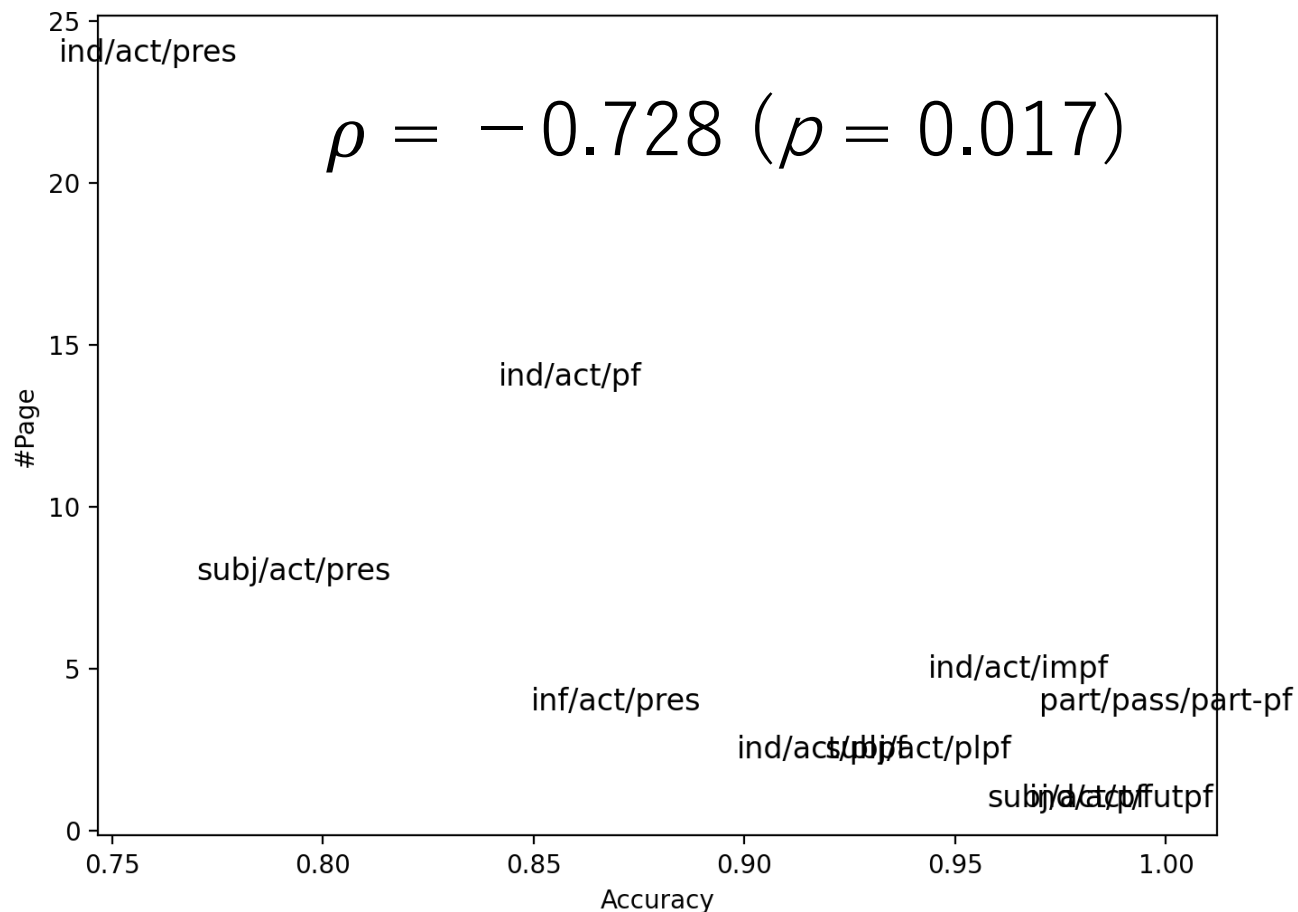
正解率		$(p < 0.001)$
形式受動態動詞	0.559 (62/111)	
それ以外の動詞	0.912 (1723/1889)	

全体に占める割合の小ささ & 「意味は能動だが活用は受動態」という特殊性→学習が著しく困難

→実際に、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

≡[4]で各活用の説明に割かれている紙数

正解率と実際の形態的変化の大きさ



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

→モデルの出力が，現実の通時変化の傾向と合致

ただし，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（直説法・能動態・未来，接続法・能動態・未完了や受動態）

→形態面の学習困難性以外の他の要因の検討が必要

ロマンス語形の出現

- 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）

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！

アウトライン

- 手法 & 実験
- 考察
- まとめ

まとめ

- 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
→ RNN は現実の言語変化を部分的に再現
- 計算機で通時的言語変化をモデリングできるかも？

応用例2

- 第一言語習得
- 第二言語習得
- 通時意味論
- 通時形態論
- まとめ

コーパス検索では見えないこと、 深層学習で見えること

- 単語や文をベクトルとして表すことで、分析の幅が広がる
- 現象の再現を含むモデル化やシミュレーションができる
- コーパスの全データ + 深層学習で、新たな分析ができる

ご清聴ありがとうございました。

* 本研究はJSPS科研費JP18K12361, JP18K11456の助成を受けたものです。