分散表現を用いた
ロマンス語同源語の意味変化の分析

「現代語の意味の変化に対する計算的・統計力学的アプローチ」
シンポジウム

川崎 義史
東京大学大学院 総合文化研究科 言語情報科学専攻
ykawasaki@g.ecc.u-tokyo.ac.jp

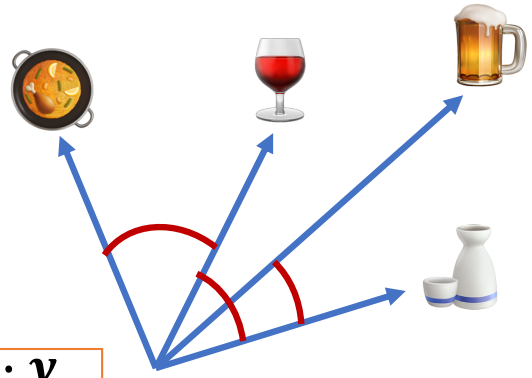
＝分布仮説（Harris 1954; Firth 1957）に基づく，密なベクトル表現

単語の分散表現

同様の文脈に出現する単語同士は「意味」が類似

バルで**ワイン**を飲む
パブで**ビール**を飲む
居酒屋で**日本酒**を飲む
レストランで**パエリア**を食べる

学習



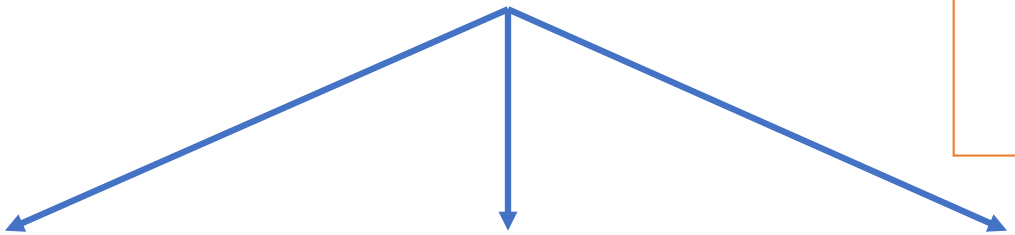
$$cos(\boldsymbol{x},\boldsymbol{y}) = \frac{\boldsymbol{x}\cdot\boldsymbol{y}}{\|\boldsymbol{x}\|\|\boldsymbol{y}\|}$$
$$-1 \le cos(\boldsymbol{x},\boldsymbol{y}) \le 1$$

コサイン類似度**大**（ベクトルの向きが同じ）＝「意味」が類似
コサイン類似度**小**（ベクトルの向きが違う）＝「意味」が異なる

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’



ロマンス語同源語ペア
🇫🇷avoir–🇪🇸haber
🇪🇸haber–🇮🇹avere
🇮🇹avere–🇫🇷avoir

🇪🇸haber
❌Yo he hambre.
⭕Yo **tengo** hambre.
< Lat. tenere ‘to hold’

🇫🇷avoir
J’ai faim.

🇮🇹avere
Io ho fame.
「私は空腹を持っています」

🇫🇷avoirと🇮🇹avereは似ている，🇪🇸haberは少し違う
→同源語の意味変化に影響を与える言語内要因を特定したい

アウトライン

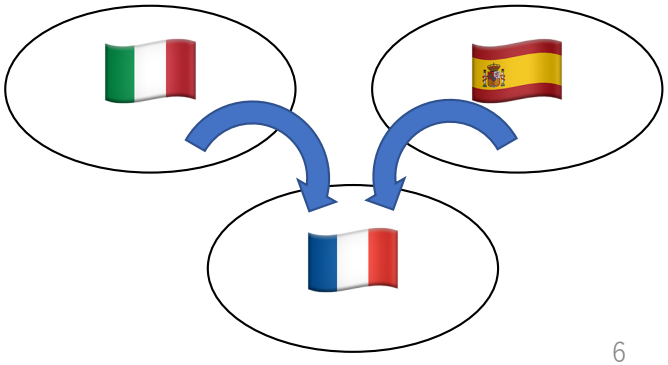
- 分散表現の学習
- 分析方法
- 実験結果
- まとめ

アウトライン

- 分散表現の学習
- 分析方法
- 実験結果
- まとめ

多言語分散表現の獲得

- 各言語，レンマ化済みのWikipediaダンプ（2018年12月時点）で分散表現（600次元）の学習（gensimのword2vecを利用）
- ある言語の空間に，それ以外の言語の分散表現を線形写像
  - バイリンガル辞書MUSEから抽出したシード語ペア（約1万語）の二乗誤差の和が最小になるように線形変換
  - 写像先の言語による大きな影響は見られない



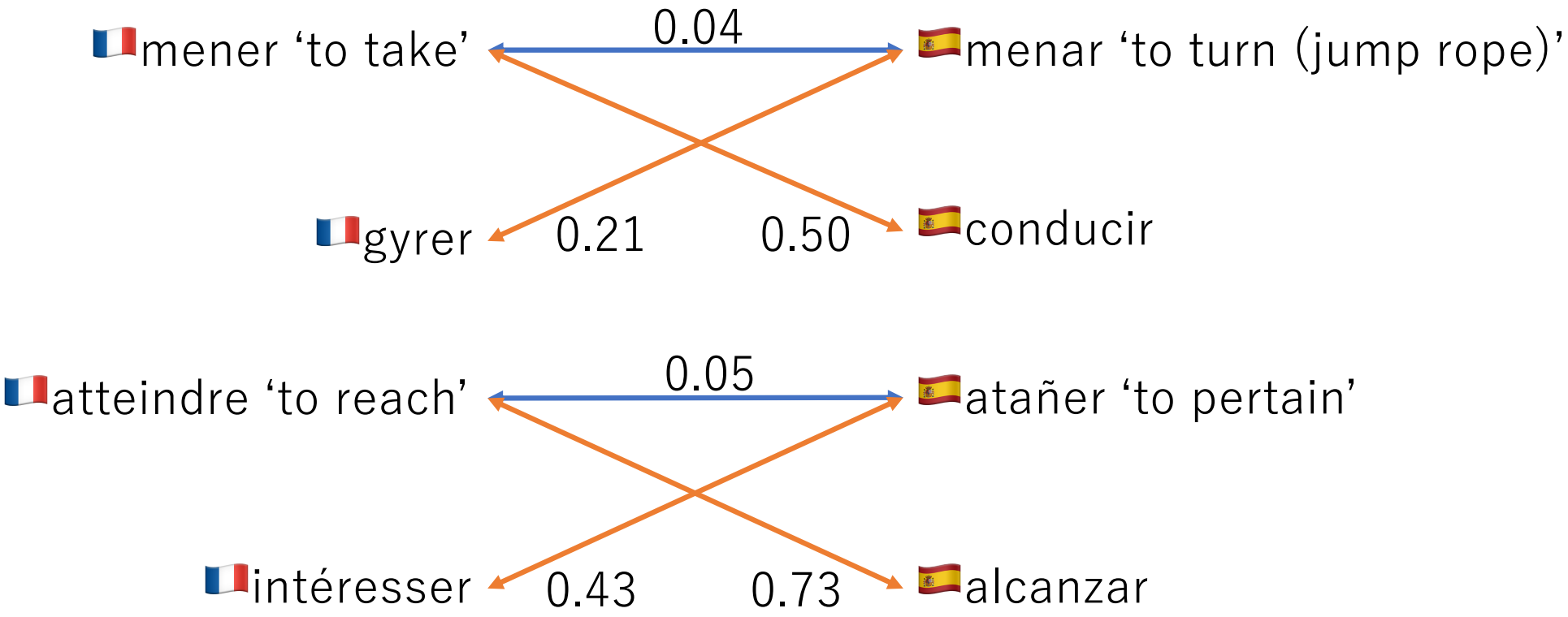
単語ペア類似度

	🇫🇷	🇪🇸	類似度
類似度最大	construire ‘to construct’	constuir ‘id.’	0.87
	provoquer ‘to provoke’	provocar ‘id.’	0.87
	détruire ‘to destroy’	destruir ‘id.’	0.86
	assassiner ‘to assassinate’	asesinar ‘id.’	0.85
	chanter ‘to sing’	cantar ‘id.’	0.84
類似度最小	falloir ‘to be necessary’	fallir ‘to fail’	0.07
	revenir ‘to return’	revenir ‘id.’	0.06
	avoyer ‘to set saw’	aviar ‘to prepare’	0.05
	atteindre ‘to reach’	atañer ‘to pertain’	0.05
	mener ‘to take’	menar ‘to turn (jump rope)’	0.04

→直感に合う結果

2言語間において，同語源で形は類似しているが意味が異なる単語ペア

偽りの友（false friend）



→意味的に対応する語を検出可能

アウトライン

- 分散表現の学習
- 分析方法
- 実験結果
- まとめ

先行研究

- 頻度**大**→意味変化**小**（Hamilton+ 2016　英語）
- 頻度**小**→意味変化**小**（Uban+ 2019　ロマンス語同源語）
  - ペアの片方の言語の頻度のみ→本研究では両方の頻度を考慮

- 多義性**大**→意味変化**大**（Hamilton+ 2016; Uban+ 2019）
  - 多義性＝分散表現と同じベクトル空間上の最近隣語のクラスタ係数

  →データ内の頻度と大きな相関があるという問題点（Dubossarsky+ 2017）

  →本研究では，分散表現と独立に，しかも変化前の多義性を使用

本研究の新規性

- 意味変化**前**のラテン語の頻度・多義性を考慮（Uban+ (2019)では，意味変化**後**の頻度と多義性）
- ロマンス語のペア頻度の考慮（Uban+ (2019）では片方の頻度のみ）
- ラテン語の語源の単語長や，ロマンス語との編集距離の考慮
- 分散表現のノルムの考慮

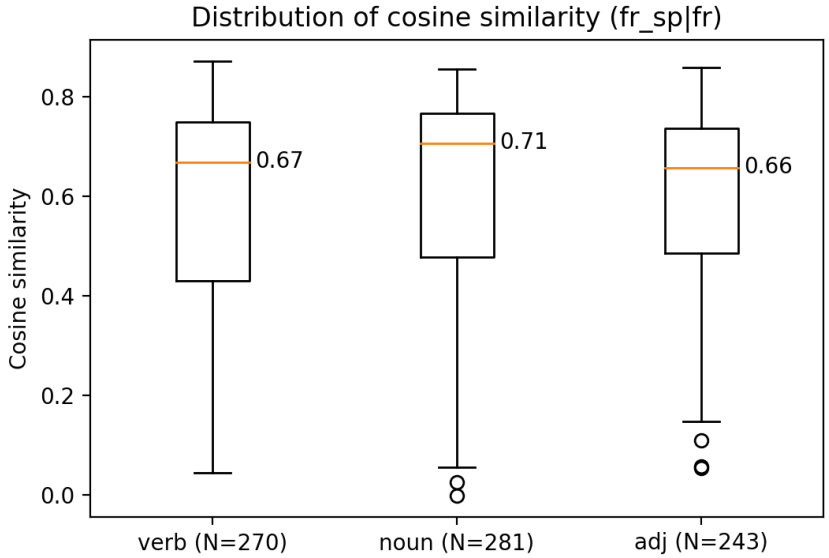
目的変数

コサイン類似度**大**→意味変化**小**
コサイン類似度**小**→意味変化**大**

動詞，名詞，形容詞のそれぞれに関して，仏伊西いずれかの言語のデータにおいて最頻上位300に含まれる同源語ペアのコサイン類似度

- 動詞：493組
- 名詞：487組
- 形容詞：477組

→ラテン語の頻度，多義性，語源形を取得できたのは半数程度



本研究の説明変数

意味変化前

- ラテン語の頻度（FREQ_LAT）：PhiloLogic4から取得した，ラテン語の語源の1万語あたりの頻度を対数変換したもの。
- ラテン語の多義性（POLY_LAT）：Oxford Latin Dictionaryに掲載されているラテン語の語源の語義の見出し数。
- ラテン語の語源の単語長（LEN_LAT）：ラテン語の語源の文字数。形態的複雑性の指標（接辞付加や派生→意味の限定・単語長大：e.g. Lat. praevidere ‘to foresee’ <prae- ‘ahead’ + videre ‘to see’ ）

意味変化後

- ロマンス語同源語ペアの頻度（FREQ_ROM）：ロマンス語同源語ペアの各言語データ内での相対頻度（例えば，🇫🇷avoirと🇪🇸haberの各言語での相対頻度）の調和平均を対数変換したもの。
- ロマンス語同源語ペアのノルム（NORM_ROM）：ロマンス語同源語ペアの各々の分散表現のノルムの算術平均。頻度と多義性の指標（頻度大＆特定の文脈のみで使用→ノルム大）
- ラテン語の語源とロマンス語同源語ペアとの編集距離（EDIT）：ラテン語の語源 l とロマンス語同源語ペア(r, r')の各々の編集距離$edit(\cdot)$を単語長$|\cdot|$で正規化したものの平均。民衆語・教養語（Penny 2002）の指標

→線形回帰分析で，意味変化に影響する言語内要因を特定する

ロマンス語同源語ペアの頻度 (FREQ_ROM)

ロマンス語同源語ペアの各言語データ内での相対頻度（例えば，🇫🇷avoirと🇪🇸haberの各言語での相対頻度）の調和平均を対数変換したもの

$$\frac{2}{\frac{1}{freq_{🇫🇷}}+\frac{1}{freq_{🇪🇸}}}=\frac{2\times freq_{🇫🇷}\times freq_{🇪🇸}}{freq_{🇫🇷}+freq_{🇪🇸}}$$

cf. 算術平均

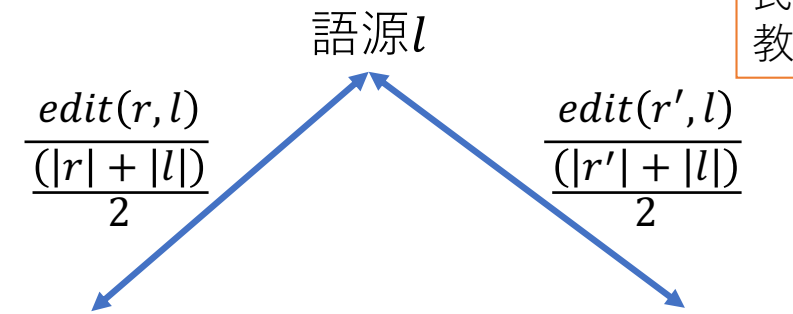
$$\frac{freq_{🇫🇷}+freq_{🇪🇸}}{2}$$

調和平均は，ペアの頻度がどちらも大きい場合にのみ大きくなり，ペアのいずれかが低頻度の場合は小さくなる。

ラテン語の語源とロマンス語同源語ペアとの編集距離（EDIT）

ラテン語の語源lとロマンス語同源語ペア(r, r')の各々の編集距離$edit(\cdot)$を単語長$|\cdot|$で正規化したものの平均。民衆語・教養語（Penny 2002）の指標

＊比較対象は，動詞は現在能動不定詞，名詞と形容詞は単数対格形。ロマンス語の母音のアクセント記号等は削除した（á, à > a）。子音はそのまま。



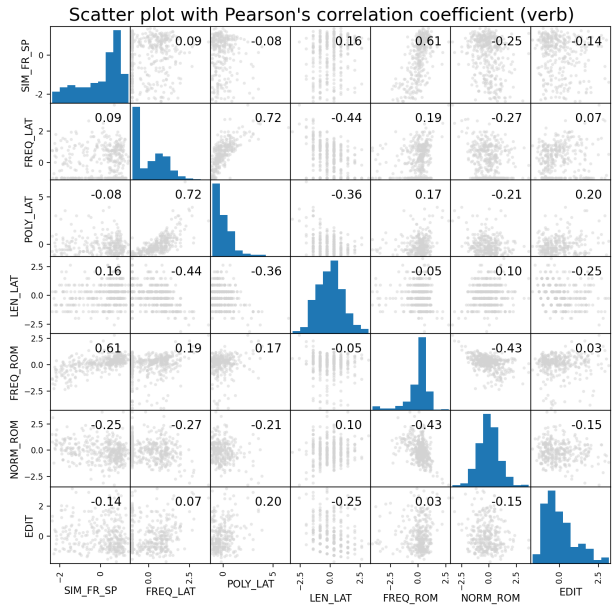
$$\rightarrow \frac{1}{2}\left(\frac{edit(r,l)}{\frac{(|r|+|l|)}{2}}+\frac{edit(r',l)}{\frac{(|r'|+|l|)}{2}}\right)$$

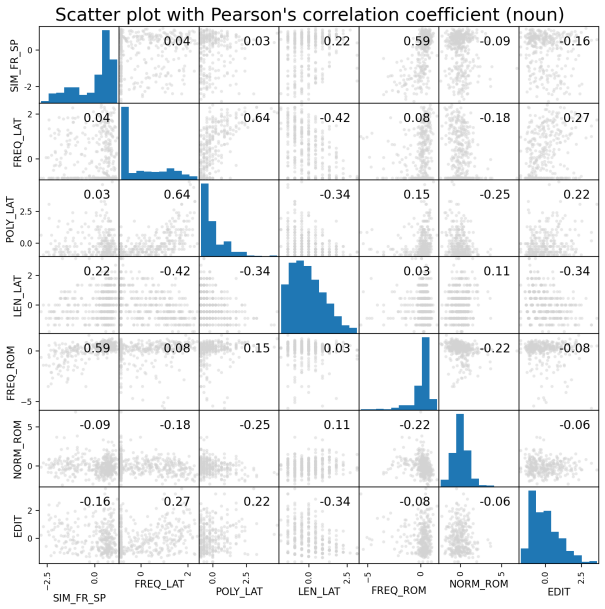
民衆語：俗ラテン語からの継承語→音変化大（編集距離大）
教養語：比較的最近の借用語→音変化小（編集距離小）

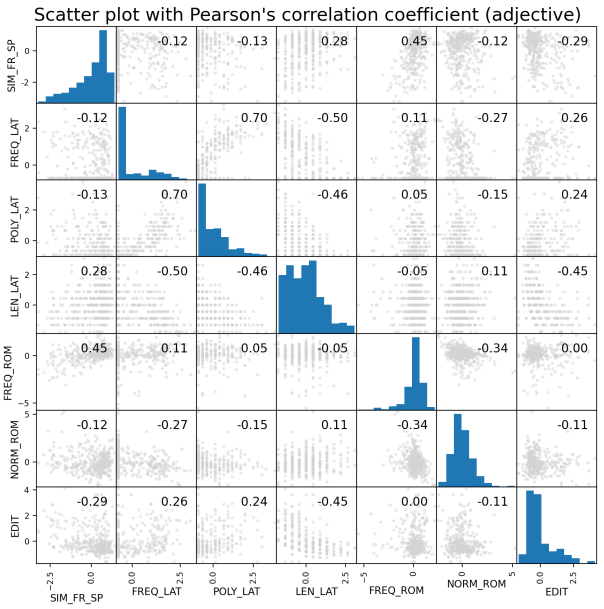
Lat. intervenire ‘to intervene’
🇪🇸intervenir
🇫🇷intervenir
🇮🇹intervenire
→編集距離小（教養語）

Lat. videre ‘to see’
🇪🇸ver
🇫🇷voir
🇮🇹vedere
→編集距離大（民衆語）

仏・西同源語ペアの類似度と説明変数の散布図と相関係数







変数は全て標準化

アウトライン

- 分散表現の学習
- 分析方法
- 実験結果
- まとめ

実験結果（全体）

仏・西ペアのコサイン類似度を目的変数とした回帰分析の結果
（全変数を標準化，AICによる変数選択あり）

	回帰係数	標準誤差	t	p > |t|
切片	0.00	0.03	0.00	1.00
FREQ_LAT	0.09	0.04	2.30	0.02
POLY_LAT	**-0.12**	**0.04**	**-3.00**	**0.00**
LEN_LAT	**0.20**	**0.03**	**6.19**	**0.00**
FREQ_ROM	**0.57**	**0.03**	**20.00**	**0.00**
NORM_ROM	–	–	–	–
EDIT	**-0.13**	**0.03**	**-4.32**	**0.00**
		N=794		
		Adj. R2=0.39		

調和平均の代わりに算術平均を利用
→Adj. R2=0.11

- ラテン語の多義性**大**→意味変化**大**
- ラテン語の単語長**大**→意味変化**小**
- ロマンス語の頻度**大**→意味変化**小**
- 編集距離**大**→意味変化**大**

実験結果（品詞別）

	動詞				名詞				形容詞			
	回帰係数	標準誤差	t	p > |t|	回帰係数	標準誤差	t	p > |t|	回帰係数	標準誤差	t	p > |t|
切片	0.00	0.04	0.00	1.00	0.00	0.05	0.00	1.00	0.00	0.05	0.00	1.00
FREQ_LAT	0.13	0.07	1.94	0.05	–	–	–	–	–	–	–	–
POLY_LAT	**-0.24**	**0.07**	**-3.77**	**0.00**	–	–	–	–	–	–	–	–
LEN_LAT	**0.17**	**0.05**	**3.43**	**0.00**	**0.24**	**0.05**	**5.19**	**0.00**	**0.19**	**0.06**	**3.20**	**0.00**
FREQ_ROM	**0.66**	**0.05**	**14.21**	**0.00**	**0.58**	**0.05**	**12.28**	**0.00**	**0.48**	**0.05**	**9.01**	**0.00**
NORM_ROM	–	–	–	–	–	–	–	–	–	–	–	–
EDIT	**-0.14**	**0.05**	**-2.95**	**0.00**	–	–	–	–	**-0.19**	**0.06**	**-3.14**	**0.00**
	N=270				N=281				N=243			
	Adj. R2=0.49				Adj. R2=0.38				Adj. R2=0.31			

Adj. R2：動詞＞名詞＞形容詞
名詞や形容詞は，説明変数以外の影響（言語外要因など）を受けやすい？

考察

- ロマンス語の頻度**大**→語と意味の結びつきが弱化しにくい（Bybee 2015）→意味変化**小**
  - ロマンス語の頻度は意味変化**後**のもの。意味変化の結果？原因？
- ラテン語の多義性**大**→言語ごとに，異なる意味が優勢になる可能性が高い→意味変化**大**
  - e.g. Lat. trahere ‘to drag’（意味数22）＞ 🇮🇹trarre ‘to draw’, 🇫🇷traire ‘to milk’, 🇪🇸traer ‘to bring’
- ラテン語の単語長**大**（接辞付加や派生）→意味の限定→意味変化**小**
- 編集距離**大**（ロマンス語での使用期間が長い）→意味変化**大**

意味変化前のラテン語の変数のみで回帰

	全体			
	回帰係数	標準誤差	t	p > |t|
切片	0.00	0.03	0.00	1.00
FREQ_LAT	**0.16**	**0.05**	**3.28**	**0.00**
POLY_LAT	-0.07	0.05	-1.5	0.14
LEN_LAT	**0.27**	**0.04**	**7.00**	**0.00**
	N=794			
	Adj. R2=0.06			

- 単語長**大**→意味変化**小**
- 頻度**大**→意味変化**小**

有意だが，Adj. R2は微小・・・

	動詞				名詞				形容詞			
	回帰係数	標準誤差	t	p > |t|	回帰係数	標準誤差	t	p > |t|	回帰係数	標準誤差	t	p > |t|
切片	0.00	0.06	0.00	1.00	0.00	0.06	0.00	1.00	0.00	0.06	0.00	1.00
FREQ_LAT	**0.34**	**0.09**	**3.88**	**0.00**	0.14	0.08	1.72	0.09	0.06	0.09	0.66	0.51
POLY_LAT	**-0.26**	**0.09**	**-3.01**	**0.00**	0.05	0.08	0.67	0.50	-0.06	0.09	-0.66	0.51
LEN_LAT	**0.23**	**0.06**	**3.62**	**0.00**	**0.31**	**0.06**	**4.82**	**0.00**	**0.26**	**0.07**	**3.72**	**0.00**
	N=270				N=281				N=243			
	Adj. R2=0.07				Adj. R2=0.07				Adj. R2=0.06			

アウトライン

- 分散表現の学習
- 分析方法
- 実験結果
- まとめ

まとめ

- ラテン語の多義性**大**→意味変化**大**
- ラテン語の単語長**大**→意味変化**小**
- ロマンス語の頻度**大**→意味変化**小**
- 編集距離**大**→意味変化**大**

- 品詞ごとに異なる振る舞い？

- 今後の課題
  - 意味変化の種類（特殊化・一般化・良化・悪化など）の特定（Traugott & Dasher 2005）
  - 多義性に関して，典型性や上位・下位関係など質的違いの考慮

参考文献 1

- Alkire, T., & Rosen, C. (2010). *Romance Languages: A Historical Introduction*. New York: Cambridge University Press.
- Bybee, J. (2015). *Language Change*. Cambridge: Cambridge University Press.
- Dubossarsky, H., Grossman, E., & Weinshall, D. (2017). Outta Control: Laws of Semantic Change and Inherent Biases in Word Representation Models. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 1136–1145. https://doi.org/10.18653/v1/d17-1118
- Dubossarsky, H., Weinshall, D., & Grossman, E. (2016). Verbs Change More Than Nouns: A Bottom-up Computational Approach to Semantic Change. *Lingue e Linguaggio*, *15*(1), 5–25. https://doi.org/10.1418/83652
- Firth, J. (1957). A synopsis of linguistic theory 1930-1955. *Studies in Linguistic Analysis*. Oxford Philological Society. 1-32.
- Glare, P. G. W. (Ed.). (2012). *Oxford Latin Dictionary* (2nd ed.). Oxford: Oxford University Press.
- Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016a). Cultural Shift or Linguistic Drift? Comparing Two Computational Measures of Semantic Change. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2116–2121. https://doi.org/10.18653/v1/d16-1229
- Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016b). Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 1489–1501. https://doi.org/10.18653/v1/p16-1141
- Harris, Z. S. (1954). Distributional structure. *Word*. 10:2-3, 146-162.

参考文献 2

- Kutuzov, A., Øvrelid, L., Szymanski, T., & Velldal, E. (2018). Diachronic Word Embeddings and Semantic Shifts: A Survey. *Proceedings of the 27th International Conference on Computational Linguistics*, 1384–1397. Santa Fe, New Mexico, USA.
- Mikolov, T., Le, Q. V, & Sutskever, I. (2013). Exploiting Similarities among Languages for Machine Translation. *CoRR Abs/1309.4*. Retrieved from http://arxiv.org/abs/1309.4168
- Oniga, R. (2014). *Latin: A Linguistic Introduction* (N. Schifano, ed.). New York: Oxford University Press.
- Penny, R. (2002). *A History of the Spanish Language*. Cambridge: Cambridge University Press.
- Řehůřek, R., & Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45–50. Valletta, Malta: ELRA.
- Schlechtweg, D., McGillivray, B., Hengchen, S., Dubossarsky, H., & Tahmasebi, N. (2021). SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection. *Proceedings of the 14th International Workshop on Semantic Evaluation*, 1–23. https://doi.org/10.18653/v1/2020.semeval-1.1
- Takamura, H., Nagata, R., & Kawasaki, Y. (2017). Analyzing Semantic Changes in Japanese Loanwords. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, *1*, 1195–1204. https://doi.org/10.18653/v1/e17-1112
- Traugott, E. C., & Dasher, R. B. (2005). *Regularity in Semantic Change*. Cambridge: Cambridge University Press.
- Uban, A. S., Ciobanu, A. M., & Dinu, L. P. (2019). Studying Laws of Semantic Divergence across Languages using Cognate Sets. *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, 161–166. https://doi.org/10.18653/v1/w19-4720

ご清聴ありがとうございました。

＊本研究はJSPS科研費18K11456の助成を受けたものです。