

分散表現を用いたロマンス語同源語動詞の意味変化の分析

川崎 義史¹ Maëlys Salingre¹ Marzena Karpinska² 高村 大也³ 永田 亮⁴

¹ 東京大学 ²University of Massachusetts Amherst ³産業技術総合研究所 ⁴甲南大学
 ykawasaki@g.ecc.u-tokyo.ac.jp m_salingre@phiz.c.u-tokyo.ac.jp mkarpinska@cs.umass.edu
 takamura.hiroya@aist.go.jp nagata-nlp2022@ml.hyogo-u.ac.jp.

概要

ロマンス語とは、ラテン語から派生した姉妹言語の総称である。ロマンス諸語には、語源を共有する同源語が数多く存在するが、意味や用法が異なるものもある。本稿では、ロマンス語のうち仏伊西の動詞に焦点を絞り、意味変化の大きさと、言語内要因の関係を統計的に分析した。意味変化の大きさは単語の分散表現で測定した。分析の結果、意味変化の大きさには、ロマンス語の頻度、ラテン語の多義性、ロマンス語の単語長が影響することが判明した。

1. はじめに

ロマンス語とは、イタリア語、スペイン語、フランス語、ポルトガル語など、ラテン語から派生した姉妹言語の総称である[1]。ロマンス諸語には、語源を共有する**同源語**が数多く存在する。例えば、Lat. *habere* ‘to have’ は、Fr. *avoir*, It. *avere*, Sp. *haber* に対応する。しかし、イタリア語・フランス語では元々の「持つ」という意味を保持している一方、スペイン語ではこの意味は失われ、専ら完了時制を形成する助動詞として使用される。このように、語源的関係はあるものの、意味や用法が異なる場合がある。

本稿の目的は、ロマンス語のうち仏伊西の動詞に焦点を絞り、**意味変化**の大きさと、頻度や多義性などの言語内要因の関係を統計的に分析することである。計算的手法による通時的意味変化の検知は近年大きな注目を集めている[2], [3]。意味変化には、社会文化、政治状況、技術革新など言語外的要因も働くが[4]–[6]、本稿では扱わない。意味変化の大きさの測定には、その検出能力が知られている**単語の分散表現**を利用した[2], [3], [7], [8]。分析対象を動詞に限定したのは、名詞や形容詞よりも動詞の方が意味変化を受けやすく[8]、検知しやすいと考えたためである。

本研究の寄与は、次の3点である。まず、意味変化の大きさには、ロマンス語の頻度、ラテン語の多義性、ロマンス語の単語長が影響することを示した点である。2点目は、ラテン語の語源の頻度と多義性など、先行研究では考慮されていない変数を導入した点である。3点目は、**ロマンス語同源語ペア**の平均頻度の算出には、算術平均よりも調和平均の方が適していることを示した点である。ここで、ロマンス語同源語ペアとは、仏・西ペアの *avoir-haber* のように、共通のラテン語の語源を持つロマンス語の単語ペアのことを指す。

本稿の構成は次の通りである。2節でデータの説明、3節で実験結果の報告、4節で考察を行い、5節で結論と今後の課題を述べる。

2. 分析方法

分散表現学習用のデータとして、各言語、2018年12月時点の Wikipedia ダンプを使用した。まず、TreeTagger でレンマ化し、イタリア語・スペイン語は Wikiextractor で、フランス語は Wiki-cirrus でクリーニングを行ったⁱ。単純のため、同綴異義語の区別はしていない。

分散表現の学習は、gensim [9]の word2vec で行ったⁱⁱ。vector_size=600, min_count=50 とし、それ以外のハイパーパラメータはデフォルトの設定にした。

ロマンス語同源語ペアの類似度を測るためには、いずれかの言語のベクトル空間に、他の2言語の分散表現を写像する必要がある。写像は、シード語ペアの二乗誤差の和が最小になるように線形変換により行った[10]。シード語ペアはバイリンガル辞書 MUSE の supervised バージョンから抽出したⁱⁱⁱ。動詞はシード語から除外した。いずれの言語ペアの組み合わせでも、シード語の数は約1万語となった。後述の分析に、写像先の言語による大きな影響は見られなかった。以下の説明では、イタリア語・スペイ

ⁱ <https://github.com/attardi/wikiextractor>

ⁱⁱ <https://radimrehurek.com/gensim/models/word2vec.html>

ⁱⁱⁱ <https://ai.facebook.com/tools/muse/>

ン語がフランス語の空間に写像されたものとする。

分析対象は、仏伊西いずれかの言語のデータにおいて最頻動詞上位 300 に含まれるものとし、合計で 493 組となった。これらのうち、3 言語全てに同源語が存在し、後述のラテン語の頻度と多義性の数を取てきたのは、270 組だった。同源語の同定は、各言語の辞書の語源の記述に基づき行った^{iv}。ラテン語の頻度と多義性は、それぞれ、PhiloLogic4^vと Oxford Latin Dictionary[11]から取得した。

3. 実験

ロマンス語同源語ペアの類似度は、コサイン類似度で測った。類似度が低い（高い）ほど、意味変化の度合いは高い（低い）とみなした。表 1 は、類似度が最大・最小となる仏・西ペア各 5 組を示している。

表 1 類似度が最大（上段）・最小（下段）となる仏・西ペア各 5 組

仏	西	類似度
construire ‘to construct’	construir ‘id.’ ^{vi}	0.87
provoquer ‘to provoke’	provocar ‘id.’	0.86
détruire ‘to destroy’	destruir ‘id.’	0.86
assassiner ‘to assassinate’	asesinar ‘id.’	0.85
chanter ‘to sing’	cantar ‘id.’	0.84
recorder ‘to register’	recordar ‘to remember’	0.07
atteindre ‘to reach’	atañer ‘to pertain’	0.06
avoyer ‘to set saw’	aviar ‘to prepare’	0.05
revenir ‘to return’	revenir ‘id.’	0.05
mener ‘to take’	menar ‘to turn (jump rope)’	0.04

本稿では、ロマンス語同源語ペアの類似度を目的変数、下記の 6 つの言語学的変数を説明変数として線形回帰分析を行い、意味変化に影響する変数を特定した。意味変化前の変数は、ラテン語の頻度と多

義性である。この 2 変数の導入は、本稿の新規性の一つである。意味変化後の変数は、ロマンス語同源語ペアの頻度、単語長、ノルムである。ペアの片方[12]ではなく両者の頻度の勘案、単語長とノルムの導入も新規性である。もう一つの新規性は、編集距離の導入である。この変数は、ラテン語の語源とロマンス語との表記上のずれを表す。本来は、音声の比較が望ましいが、簡単のため、表記の比較とした。

- **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)** : [11]に掲載されているラテン語の語源の意味の見出し数。
- **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、Fr. avoir と Sp. 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単語長 (LEN_ROM)** : ロマンス語同源語ペアの各々に含まれる文字数の算術平均。
- **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。
- **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 $|l|$ で正規化したものの平均 (3.1)。

$$\frac{1}{2} \left(\frac{edit(r, l)}{\frac{|r| + |l|}{2}} + \frac{edit(r', l)}{\frac{|r'| + |l|}{2}} \right) \quad (3.1)$$

比較対象は不定詞とした。ロマンス語の母音のアクセント記号等は削除した（例えば、á や à は a に統一した）。子音に関しては、スペイン語の ñ やフランス語の ç などラテン語には存在しないものや、ロマンス語とラテン語で音価の異なるものがあるが (Lat. citare [k] ‘to set in motion’, Sp. citar [θ] ‘to cite’), 単純のためにそのまま扱った。ラテン語の形式受動相動詞[13]の不定詞は、対応する活用 of 能動相で想定される形とし

^{iv} Fr. <https://www.cnrtl.fr>

It. <https://www.etimo.it>

Sp. <https://dle.rae.es>

^v <http://artflsrv02.uchicago.edu/cgi-bin/perseus/LatinFrequency.pl?author=&title=&genre=&editor=&language=NOT+English&displaymorethan=10&displayless>

than=10000000&sortby=decreasingFreq&searchby=searchbylemma (2021 年 1 月 28 日アクセス)。現在、サーバ更新のため、頻度リストの取得ができないようである。

^{vi} Id. (idem) は、意味がフランス語のそれと同じことを表す。

た (Lat. *jocari* ‘to joke’ > *jocare*)。不定詞の語尾が特殊な *esse* ‘to be’, *ferre* ‘to carry’, *velle* ‘to want’ 等は、そのままの形を使用した。

変数は全て平均 0, 分散 1 になるように標準化した。付録の図 1 に、仏・西同源語ペアの類似度 (SIM_FR_SP) と説明変数の散布図と相関係数を示す。FREQ_LAT と POLY_LAT には大きな相関 (0.72) があるが、Variance Inflation Factor が 3 以下なので、そのまま使用した。AIC による変数選択の結果、NORM_ROM を除く、FREQ_LAT, POLY_LAT, FREQ_ROM, LEN_ROM, EDIT の 5 変数を説明変数とするモデルが最良と判断された。表 2 に回帰分析の結果を示す。FREQ_LAT と EDIT 以外の 3 変数が有意水準 0.01 で統計的に有意となった。標準偏回帰係数は、FREQ_ROM (0.66) と LEN_ROM (0.20) が正、POLY_LAT (-0.24) が負となった。回帰係数の絶対値が大きい FREQ_ROM は、ペア類似度の予測に大きく寄与する。自由度調整済決定係数は 0.49 となり、全分散の約半分を説明していることになる。

表 2 仏・西ペアの類似度を目的変数とした回帰分析の結果 (N=270, 自由度調整済決定係数=0.49)。AIC により変数選択を行った。*は有意水準 0.01 の統計的有意を表している。

	回帰 係数	標準 誤差	t	p > t
切片	0.00	0.04	0.00	1.00
FREQ_LAT	0.13	0.07	1.88	0.06
POLY_LAT	-0.24	0.06	-3.82	0.00*
FREQ_ROM	0.66	0.05	14.30	0.00*
LEN_ROM	0.20	0.05	3.85	0.00*
EDIT	-0.11	0.05	-2.28	0.02

4. 考察

回帰分析の結果に基づき、意味変化と各変数との関係について、関連研究と対照しながら考察を行う。

4.1. 頻度

他の条件が一定ならば、ロマンス語における頻度 (FREQ_ROM) が大きいほど、ロマンス語ペアの類似度が高い、つまり、意味変化が生じにくい傾向が確認された。要因として、頻度が高いほど、語と意

味の結びつきが弱化しにくいことが考えられる[14]。この結果は、ロマンス語の同源語に関する先行研究[12]とは反対の結果となった。これは、本稿が動詞のみに注目しているのに対し、[12]は品詞の区別をしていないことに起因している可能性がある。品詞ごとに異なる振る舞いを見せる可能性もあるため、今後、動詞以外の品詞の分析が必要である。一方で、本稿の結果は、英語の通時的意味変化の分析結果と一致する[15]。しかしながら、ロマンス語の頻度は意味変化後のものなので、観察された頻度の大小が意味変化の結果なのか原因なのかは判然としない。

調和平均の代わりに算術平均を用いると、自由度調整済決定係数が 0.49 から 0.14 へ大幅に低下した。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。この性質が、ペアの片方のみで意味変化が生じた結果、類似度が低下し、且つ、その頻度が増加 (減少) した現象を上手く捉えている可能性がある。例えば、仏・西ペアの *pondre-poner* (<Lat. *ponere* ‘to put’) では、Fr. *pondre* ‘for a bird to lay egg’ は意味が特殊化[16]し低頻度語になった一方、Sp. *poner* ‘to put’ は、ラテン語の語源と同様、高頻度語のままである。実際に、このペアの類似度は 0.08 と著しく小さい。このような場合、調和平均の方が算術平均よりも小さくなるため、類似度と頻度の関係を捉えるのにより適していると考えられる。

4.2. 多義性

他の条件が一定ならば、ラテン語の多義性 (POLY_LAT) が大きいほど、ロマンス語ペアの類似度が低い、つまり、意味変化が生じやすい傾向が確認された。これは、多義性と意味変化に正の相関を見出した[12], [15]の知見と一致する。本稿の新規性は、意味変化前のラテン語の多義性を考慮した点である。[12], [15]では、意味変化後の現代語の多義性を使用している。この変数は、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として定義されている。しかし、この係数はデータ内の頻度と大きな相関があるという問題点がある[5]。本稿では、分散表現と独立に、しかも変化前の多義性を使用しており、より健全な分析となっている。

なぜ多義性が意味変化につながるのだろうか？複数の意味を持つ語の場合、仮に意味変化の割合が一定だとしても、言語 1 では意味 A が、言語 2 では別の意味 B が優勢になる可能性が高いことが考えられ

る。例えば、Lat. trahere ‘to drag’ には 22 の意味があるが、各言語で固有の意味変化が起きた結果 (It. trarre ‘to draw’, Fr. traire ‘to milk’, Sp. traer ‘to bring’), ペア間類似度は低くなる: Fr. traire - Sp. traer (0.13), Fr. traire - It. trarre (0.16), It. trarre - Sp. traer (0.21)。

4.3. 単語長

他の条件が一定ならば、単語長が大きいほど、ロマンス語ペアの類似度が高い、つまり意味変化が生じにくい傾向が確認された。単語長が大きい語は、接辞付加や派生により元となる動詞よりも形態的複雑性が増す一方、意味が限定され、多義性が小さくなる傾向にある。そのため、意味変化が生じにくいと考えられる。この現象は既にラテン語の段階で見られ、単語長と多義性の相関係数は-0.36 である。例えば、Lat. praevidere ‘to foresee’ は、videre ‘to see’ に接頭辞 prae- ‘ahead’ が付いた形で、意味が限定されている。実際に、仏・西ペアの voir-ver (< Lat. videre) の類似度が 0.35 に対し、prévoir-prever (< Lat. praevidere) の類似度は 0.72 と高い値を示す。

4.4. 編集距離

編集距離が大きいロマンス語ペアは、ラテン語から直に継承された民衆語である[6]。大きな音声変化を受けたため、語源形との表記上の違いも大きい。一方、編集距離が小さいものは、比較的新しくラテン語から借入された教養語である[6]。最小限の音声変化しか受けていないため、語源形に類似している。編集距離の回帰係数 (-0.11) は負なので、民衆語ほど意味変化しやすい傾向にあるが、統計的に有意ではない。そのため、ロマンス語での使用期間が長い民衆語ほど意味変化を受けやすいとは言えない。

4.5. ノルム

ロマンス語の分散表現のノルムの大きさ (NORM_ROM) は、変数選択により不要と判断された。頻度が高く、特定の文脈で使われる (多義性が低い) ほどノルムは大きくなるが、これらの影響はそれぞれの変数で捉えられていると考えられる。

4.6. ラテン語に関する変数

意味変化前のラテン語の頻度 (FREQ_LAT) と多義性 (POLY_LAT) のみを説明変数として回帰すると、FREQ_LAT の係数は正 (0.26), POLY_LAT は負 (-0.28) になり、ともに有意水準 0.01 で統計的に

有意となった。自由度調整済決定係数は微小 (0.03) なものの、ラテン語の語源が高頻度で多義性が小さいロマンス語同源語ペアほど類似度が高い、つまり、意味変化を受けにくいことを示唆している。

4.7. ロマンズ語への流入時期の違い

辞書[11]に非掲載の動詞は、上述の分析からは除外した。これらの動詞は、俗ラテン語、後期・中世ラテン語、ロマンス語の段階で形成されたもの (Fr. travailler ‘to work’, It. travagliare ‘to suffer’, Sp. trabajar ‘to work’ < Vulgar Lat. tripaliare ‘to torture’) [6], 他言語からの借入語 (Fr. assassiner, It. assassinare, Sp. asesinar < Arabic ḥaššāšīn ‘hashish’; Fr. garder ‘to keep’, It. guardare ‘to look’, Sp. guardar ‘to guard’ < Germanic wardōn ‘to pay attention’) [6], 古典ラテン語では使用が稀なものなどである。つまり、ロマンス語への流入時期が遅い動詞である。

流入時期の違いがロマンス語同源語ペア類似度に影響するか検証するため、語源が[11]に非掲載の動詞も含めて回帰分析を行なった。説明変数は、FREQ_ROM, LEN_ROM, NORM_ROM に加え、辞書掲載の有無を表すダミー変数とした。掲載を 1 (データ数 270), 非掲載を 0 (同 105) とした。流入時期の違いの係数は正 (0.07) だったが、絶対値は極めて小さく、統計的に有意でもなかった。よって、流入時期が古いほど意味変化を起こしやすいとは言えない。

5. おわりに

本稿では、ロマンス語のうち仏伊西の動詞に焦点を絞り、意味変化の大きさと、言語内要因の関係を統計的に分析した。意味変化の大きさは単語の分散表現で測定した。分析の結果、意味変化の大きさには、ロマンス語の頻度、ラテン語の多義性、ロマンス語の単語長が影響することが判明した。

今後の課題は、以下の 3 点である。まず、名詞や形容詞など他の品詞の分析が必要である。動詞以外の品詞では、異なる傾向が見られる可能性がある[8]。第二に、意味変化の種類 (特殊化・一般化・良化・悪化など[16]) を特定することである[3]。第三に、本稿では多義性を辞書掲載の意味の見出し数と定義したが、典型性や上位・下位関係など質的違いを考慮する必要がある。WordNet やポアンカレ埋め込みなどを利用することで、これらの違いを捉えられる可能性がある。

謝辞

本研究はJSPS 科研費 18K11456 の助成を受けたものです。

参考文献

- [1] T. Alkire and C. Rosen, *Romance Languages: A Historical Introduction*. New York: Cambridge University Press, 2010.
- [2] D. Schlechtweg, B. McGillivray, S. Hengchen, H. Dubossarsky, and N. Tahmasebi, “SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection,” in *Proceedings of the 14th International Workshop on Semantic Evaluation*, 2021, pp. 1–23.
- [3] A. Kutuzov, L. Øvrelid, T. Szymanski, and E. Velldal, “Diachronic Word Embeddings and Semantic Shifts: A Survey,” in *Proceedings of the 27th International Conference on Computational Linguistics*, 2018, pp. 1384–1397.
- [4] W. L. Hamilton, J. Leskovec, and D. Jurafsky, “Cultural Shift or Linguistic Drift? Comparing Two Computational Measures of Semantic Change,” in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 2116–2121.
- [5] H. Dubossarsky, E. Grossman, and D. Weinshall, “Outta Control: Laws of Semantic Change and Inherent Biases in Word Representation Models,” in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 1136–1145.
- [6] R. Penny, *A History of the Spanish Language*. Cambridge: Cambridge University Press, 2002.
- [7] H. Takamura, R. Nagata, and Y. Kawasaki, “Analyzing Semantic Changes in Japanese Loanwords,” in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, 2017, vol. 1, pp. 1195–1204.
- [8] H. Dubossarsky, D. Weinshall, and E. Grossman, “Verbs Change More Than Nouns: A Bottom-up Computational Approach to Semantic Change,” *Lingue e Linguaggio*, vol. 15, no. 1, pp. 5–25, 2016.
- [9] R. Řehůřek and P. Sojka, “Software Framework for Topic Modelling with Large Corpora,” in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 2010, pp. 45–50.
- [10] T. Mikolov, Q. V. Le, and I. Sutskever, “Exploiting Similarities among Languages for Machine Translation,” *CoRR abs/1309.4*, 2013.
- [11] P. G. W. Glare, Ed., *Oxford Latin Dictionary*, 2nd ed. Oxford: Oxford University Press, 2012.
- [12] A. S. Uban, A. M. Ciobanu, and L. P. Dinu, “Studying Laws of Semantic Divergence across Languages using Cognate Sets,” in *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, 2019, pp. 161–166.
- [13] R. Oniga, *Latin: A Linguistic Introduction*. New York: Oxford University Press, 2014.
- [14] J. Bybee, *Language Change*. Cambridge: Cambridge University Press, 2015.
- [15] W. L. Hamilton, J. Leskovec, and D. Jurafsky, “Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change,” in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 2016, pp. 1489–1501.
- [16] E. C. Traugott and R. B. Dasher, *Regularity in Semantic Change*. Cambridge: Cambridge University Press, 2005.

付録

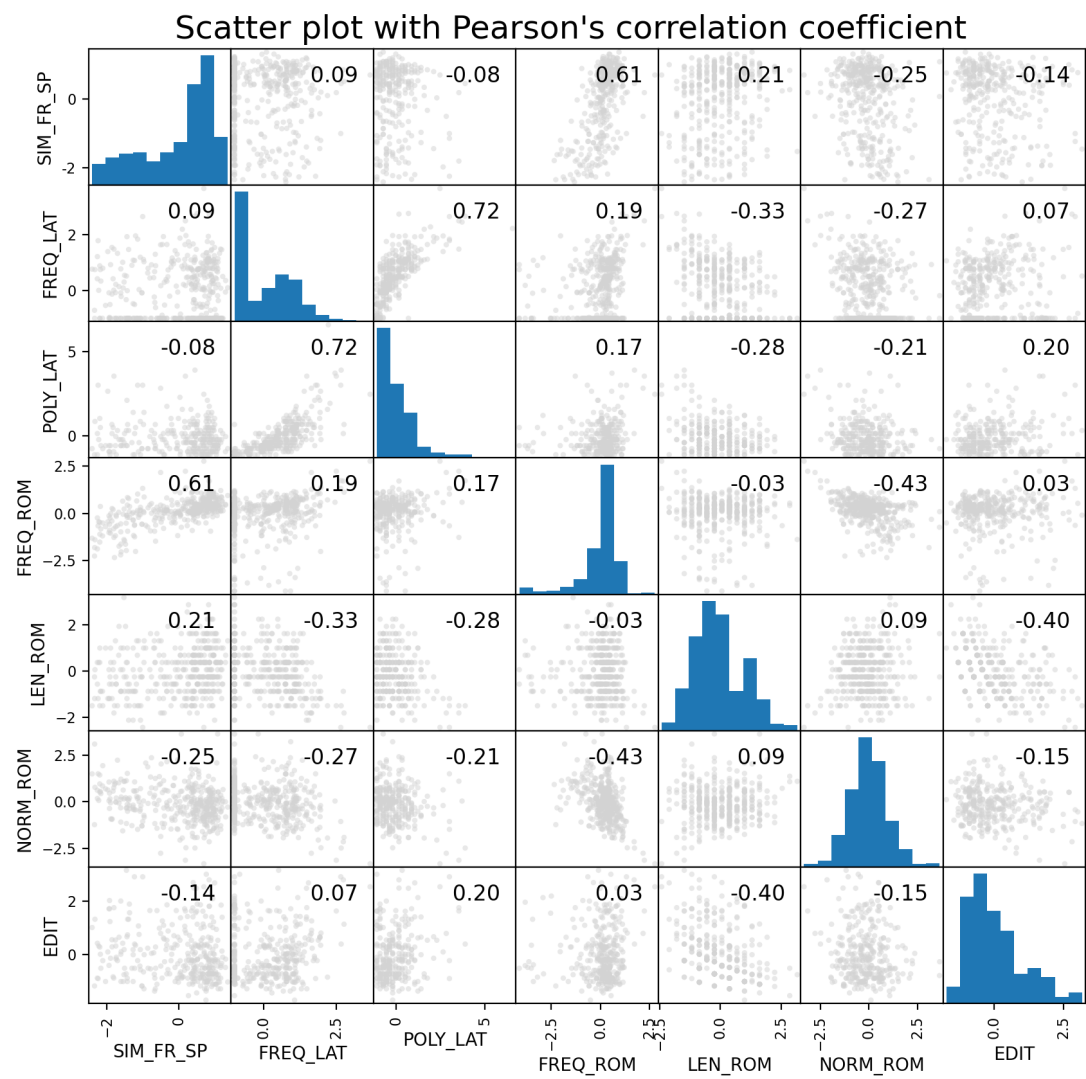


図 1 仏・西同源語ペアの類似度 (SIM_FR_SP) と説明変数の散布図と相関係数