言語ベクトルの類似度に基づく低資源言語への転移学習

李辰仁　大関洋平
東京大学
{ashmore985,oseki}@g.ecc.u-tokyo.ac.jp

概要

近年, 多言語モデルは複数言語にまたがる多くのタスクにおいて高い性能を示している. しかし, その中でも自然言語推論タスクに関しては, 複数言語における転移学習性能の差異が, 言語系統あるいは言語類型によるものなのか必ずしも明確に整理されていない. 本研究では, 言語ベクトルの類似度に基づく定量的な言語間距離により, 自然言語推論タスクを用いた低資源言語への転移学習実験を行うことで, 系統的距離および類型的距離が転移学習性能に与える影響を比較・分析する. 実験の結果, 従来の直感的分類では系統的・類型的類似性が認められてこなかった言語間での転移学習性能が, 類似度が高いとされてきた言語間を上回るケースが見られた. この結果は, 言語間の類似度に関する情報の言語モデルの埋め込みが, 人間による区分とは異なる軸で組織化されていることを示唆する.

1　はじめに

Gerz らによる研究 [1] をはじめ, 多くの多言語言語モデリング（multilingual language modelling, 以下 MLM）に関する研究において, 言語の類型的特徴が言語モデルの性能に影響を与えることが示されている. Choenni ら [2] は類型的に似ている言語は共同して埋め込み表現が作られるとしているし, Bankula ら [3] は, 類型的言語間距離が低資源言語への転移学習の成否と相関するという結果を示している. これらの経験則を活用して, 実際に低資源言語のみでモデリングを試みた例としては, Fan らによる XLM-R の語族ベースのファインチューニング [4], アフリカ言語特化型 BERT モデルとして Ogueji らの「AfriBERTa」[5] と Tonja らの「InkubaLM」[6] があり, いずれも各種ベンチマークにおいて汎用型 LLM より高い性能を報告している.

しかしながら, 言語間転移学習（cross-lingual transfer, 以下 CLT）性能は言語ペアごとに大きく異なり, その差異を生み出す要因については未だ十分に解明されていない. 従来の研究では, 語族や地理的近接性といった言語系統に基づく説明が多く見られるが, これらの尺度はしばしば曖昧に用いられてきた. 例えば, 同一語族に属する言語であっても, 統語構造や形態論的複雑性, 語彙借用の程度は大きく異なる場合があり, 系統関係のみをもって CLT 性能を一義的に説明することは困難である. また, これらの類型的特徴が CLT 性能とどのように関係しているのかについては, 体系的な実証研究が不足している. 既存研究では,「類型的に似ている」といった定性的な表現にとどまる場合が多く, 系統・地理・類型といった異なる尺度が十分に区別されていないことも少なくない.

本研究では, 言語間自然言語推論タスク XNLI を評価に用いて, 複数の言語ペアを二種類の言語距離尺度で分類した上で, CLT 性能を比較・分析する. 言語ペアは, 同一クラスタ内距離, 人間によるラベル付け距離, モデルが学習した埋め込み空間に基づく距離という三つの観点から設計される. 本研究の目的は, CLT 性能に影響を与える要因として, 人間が定義する尺度を改めて精緻化し, また, モデルが獲得した尺度と比較した相対的な重要性を検証し, MLM における言語間距離尺度の課題を明確化することである. これにより, 低資源言語処理におけるモデル設計および評価枠組みに対して, さらに実証的な指針を与えることを目指す.

2　手法

本研究では lang2vec に基づく言語間距離尺度で実験用言語ペアを作成し, XNLI 評価タスクの accuracy で CLT 性能を定量化する.

2.1　言語間距離尺度

言語間距離を定量化する手法として lang2vec[7] に基づくベクトル表現を用いる. lang2vec は, 言語そのものを多次元ベクトルとして表現し, 系統的特徴

や類型的特徴を統一的な枠組みで扱うことを目的としたリソースである. 各言語は, 統語・音韻・地理的分布などの言語学的特徴を符号化した 23 種の特徴ベクトルとして表され, 言語間距離はこれらのベクトル間の距離として定義される. ここでは, その中でも文処理に関係が深い `syntax_wals` と `learned` を対照的に用いる.

`syntax_wals`　lang2vec における `syntax_wals` は, WALS（World Atlas of Language Structures）データベース [8] に基づく構文的特徴を用いた表現である. 具体的には, 語順, 接置詞の位置, 修飾関係の順序などの構文的特徴が, バイナリ, またはカテゴリ変数としてベクトル化される.

各言語 l は, 構文的特徴ベクトル $\mathbf{v}_l^{\text{syn}} \in \{0,1\}^d$ として表され, 言語間距離は通常, コサイン距離またはユークリッド距離によって計算される.

コサイン距離を用いる場合, 言語 l_1,l_2 間の距離は次のように式 1 で定義される：

$$D_{\text{syn}}(l_1, l_2) = 1 - \frac{\mathbf{v}_{l_1}^{\text{syn}} \cdot \mathbf{v}_{l_2}^{\text{syn}}}{\|\mathbf{v}_{l_1}^{\text{syn}}\| \|\mathbf{v}_{l_2}^{\text{syn}}\|} \tag{1}$$

この尺度は, 人手で定義された言語類型的知識に基づく距離であり, 言語学的解釈が比較的容易であるという利点を持つ.

`learned`　こちらは一方, Malaviya らの研究 [9] で言語間の関係を言語モデルがデータ駆動的に学習したベクトル表現で, ニューラルモデルにおける多言語学習過程から得られた言語埋め込みを基に構成されており, 明示的な言語学的特徴を前提としない.

各言語 l は実数値ベクトル $\mathbf{v}_l^{\text{learned}} \in \mathbb{R}^k$ で表し, 同じくコサイン距離を表す式 2 で定義される：

$$D_{\text{learned}}(l_1, l_2) = 1 - \frac{\mathbf{v}_{l_1}^{\text{learned}} \cdot \mathbf{v}_{l_2}^{\text{learned}}}{\|\mathbf{v}_{l_1}^{\text{learned}}\| \|\mathbf{v}_{l_2}^{\text{learned}}\|} \tag{2}$$

この表現は, 語彙的類似性, 構文的傾向, 分布的特徴などが混在した形で反映されている可能性があり, 言語モデルの内部表現との対応関係を考察する上で有用である. なお, この尺度は後述する一部の実験対象言語（en, ru, ar）の値が欠損しているため, これらを `learned` 距離の計算からは除外した.

以上のように, `lang2vec` は, 人手定義によるトップダウン型の類型的距離と, モデル予測による複合的距離を定量的に, 同一の距離計算枠組みで扱うことを可能にする, という長所がある.

2.2　評価タスク

本研究では, CLT 性能を評価するタスクとして XNLI（Cross-lingual Natural Language Inference）[10] を用いる. XNLI は, 自然言語推論（Natural Language Inference, 以下 NLI）タスクを多言語（15 言語）に拡張したデータセットであり, 文レベルの意味理解能力を評価する目的で設計されている. 具体的な言語とその例文の一部については付録 A に記す.

NLI では, 前提文（premise）と仮説文（hypothesis）, 正解ラベルで一組のデータが与えられ, それらの関係を 0: entailment（含意）, 1: neutral（中立）, 2: contradiction（矛盾）の三値分類問題として扱う.

XNLI は, 英語の NLI データセットを原文とし, クラウドソーシングによって複数言語に人手翻訳されたものから構成されている. これにより, ラベル体系と意味関係を保ったまま, 多言語間で比較可能な評価が新たに可能となっている. 評価指標としては, accuracy（分類正解率）が一般的に用いられる.

NLI は語彙的理解だけでなく, 構文構造や意味役割の把握を必要とし, 言語間の構造的類似性が性能に影響しやすいタスクであると考えられる. そのため, CLT における言語距離の影響を分析する上で, 広く用いられてきた標準的なベンチマークである.

3　実験設計

言語モデル　本研究では XLM-RoBERTa-base（XLM-R）[11] を用いる. XLM-R は, 100 以上の言語で事前学習された Transformer ベースの多言語モデルであり, CLT 性能に優れることが知られている. 本研究では, モデル構造や事前学習設定を変更せず, 下流タスクに対するファインチューニングのみを行う.

データ　XNLI には, 各言語につき約 393,700 組の訓練データ, 約 5,000 組のテストデータ, 約 2,500 組の評価データが含まれる. 本研究で対象とする言語は, 全 15 言語の中から, 以下の 11 言語に選定した：英語（en; eng）, ロシア語（ru; rus）, ブルガリア語（bg; bul）, フランス語（fr; fra）, スペイン語（es; spa）, 中国語（zh; zho）, ヒンディー語（hi; hin）, ウルドゥー語（ur; urd）, アラビア語（ar; ara）, タイ語（th; tha）, ベトナム語（vi; vie）

これらの言語の選定は, 付録 B に示す言語間距

離を指標として行なった. 特に注目したのは, 同語族で距離も近い言語, 語族は異なるが距離が近い言語, 語族も距離も遠い言語である. 興味深いことに, syntax_wals と learned の二つでは各言語の類型的距離がかなり異なる.

無条件指標として英語を起点言語としたペアも含めつつ, 以下の三つの観点から転移学習を行なう言語ペア群を構成し, 目標言語には比較的低資源な言語を設定して, CLT 性能を比較する：

A 群: ICD（Intra-cluster Distance） 同一語族の中でさらに系統的に近い言語ペア（スラヴ語派, ロマンス諸語, インド・イラン語派）と, 英語起点の言語ペアを比較する実験群で, 先行研究で主張されてきた類似度の効果を検証する：
ru–bg vs. en–bg | fr–es vs. en-es | hi–ur vs. en-ur

B 群: LTD（Labeled Typology Distance） 付録 B の (a) と (b) で観測されたウルドゥー語と各言語の syntax_wals 距離に着目し, 系統的には遠いが, 類型的には近い・遠いとされる言語ペア間での転移を比較し, 人手定義による言語間距離の効果を検証する：
zh–ur vs. en–ur vs. ar–ur vs. hi–ur

C 群: ETD（Embedded Typology Distance） 付録 B の (c) と (d) で観測されたベトナム語と各言語の learned 距離に着目し, 言語モデルが学習した埋め込み空間上の類型的距離に基づく言語ペアを比較し, モデル予測による言語間距離の効果を検証する：
fr–vi vs. th–vi vs. es–vi vs. en-vi

これらに加えて, 低資源言語単体でのモデルにおける学習性能を評価するため, 各群に対してベースラインとして同一言語ペア（bg–bg, es–es, ur–ur, vi–vi）を適用する.

4 結果

表 1, 2, 3 は各言語ペアの CLT 性能を示したものである. 系統的に近い言語ペアや類型的に近い言語ペアが, 無関係な英語からの転移と比較して, 一貫して高い accuracy を示した. ただし, その差は言語ペアによって異なり, ウルドゥー語のベースラインが CLT ペアを下回る場合も見られた.

また, 系統的に遠い zh-ur が系統的に近い hi-ur を上回ったことは特筆に値する. 埋め込み空間上で近接する言語ペア（th–vi, fr–vi など）が, 系統的関係や syntax_wals で示された類型的距離とは無関係に, 独自の相関関係を示す傾向が確認された.

さらに, 前述の相関関係を精査し, 人手定義尺度とモデル予測尺度のどちらがより強い効果を示すか比較するため, Pearson 相関係数を測定した. 結果, 図 1 で示されるように, 全実験ペアで syntax_wals 距離との中程度の相関関係（ρ=-0.47）が見られた.

一方, learned が欠損しているものを除いた言語ペアに限定すると, 図 2, 3 で見られるように, learned 距離との相関がより強いこと（ρ=-0.50 < ρ=-0.39）が確認された.

表 1: A 群（ICD）の XNLI タスクにおける CLT 性能

Train	Test	Accuracy
ブルガリア語 (bg)	ブルガリア語 (bg)	0.7904
ロシア語 (ru)	**ブルガリア語 (bg)**	**0.7840**
英語 (en)	ブルガリア語 (bg)	0.7695
スペイン語 (es)	スペイン語 (es)	0.7936
フランス語 (fr)	**スペイン語 (es)**	**0.7900**
英語 (en)	スペイン語 (es)	0.7778
ウルドゥー語 (ur)	ウルドゥー語 (ur)	0.6369
ヒンディー語 (hi)	**ウルドゥー語 (ur)**	**0.6745**
英語 (en)	ウルドゥー語 (ur)	0.6583

表 2: B 群（LTD）の XNLI タスクにおける CLT 性能

Train	Test	Accuracy
ウルドゥー語 (ur)	ウルドゥー語 (ur)	0.6369
中国語 (zh)	**ウルドゥー語 (ur)**	**0.6868**
アラビア語 (ar)	ウルドゥー語 (ur)	0.6780
ヒンディー語 (hi)	ウルドゥー語 (ur)	0.6745
英語 (en)	ウルドゥー語 (ur)	0.6583

表 3: C 群（ETD）の XNLI タスクにおける CLT 性能

Train	Test	Accuracy
ベトナム語 (vi)	ベトナム語 (vi)	0.7667
タイ語 (th)	**ベトナム語 (vi)**	**0.7645**
スペイン語 (es)	ベトナム語 (vi)	0.7583
フランス語 (fr)	ベトナム語 (vi)	0.7521
英語 (en)	ベトナム語 (vi)	0.7481

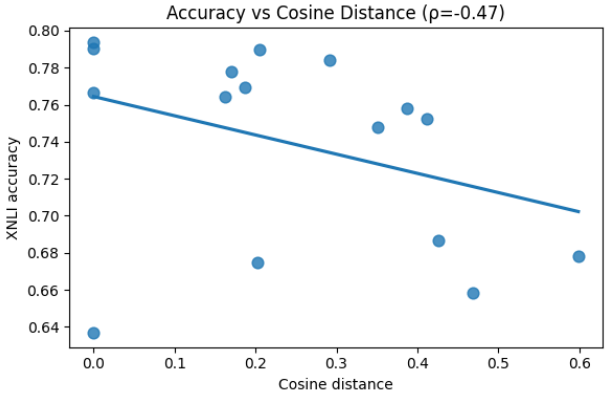


図 1: 全ペアの accuracy と syntax_wals の相関

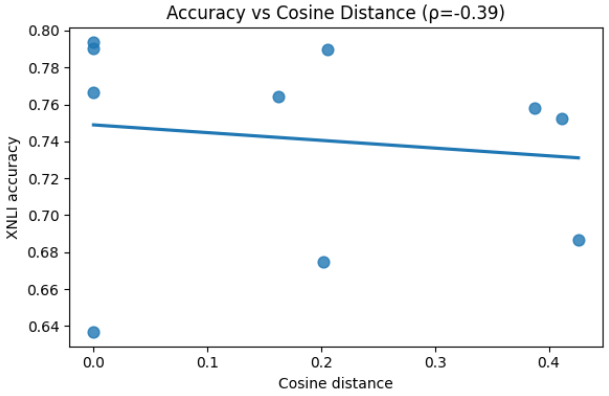


図 2: 一部ペアの accuracy と syntax_wals の相関

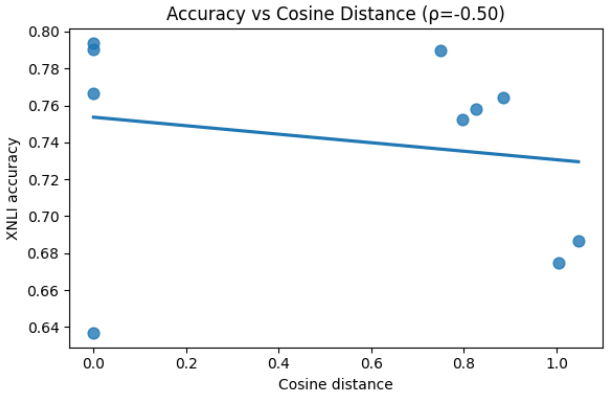


図 3: 一部ペアの accuracy と learned の相関

5 議論

本研究の結果において特に注目すべき点は, learned と accuracy の間に, 比較的高い相関が観測されたことである. 欠損している言語があることは留意が必要だが, この事実は, 言語学的知識に基づく人手定義尺度ではなく, モデル予測的に得られた言語情報の埋め込みが, より高い説明能力を持つ可能性があるという, 従来の研究では考慮されてこなかった重要な示唆を与える.

多くの先行研究では, 語族や語順, 形態論的特徴といった人手定義の指標が, CLT 性能を説明する主要因として, なかば直感的に扱われてきた. しかし, 本研究の結果は, 人手定義の指標が言語学的解釈の容易さという利点を持つ一方で, 言語モデルが内部で利用している表現構造とは乖離している可能性を示している.

learned がより高い説明能力を示した背景として, 言語モデルの埋め込み空間には, 語彙的類似性, 構文的傾向, 共起分布などが複合的に反映されている可能性が挙げられる. これらの要素は, 従来の系統関係や類型論では十分に表現できない. すなわち, モデル内部で形成される「言語類似性」は, 人手定義の区分とは異なる軸で組織化されている可能性がある.

この結果は, 人手定義の言語的距離を完全に否定するものではないが, CLT 性能の絶対的な説明変数として用いることに対する, 問いかけと位置づけることができる. 少なくとも, モデル挙動の分析においては, モデル学習による指標を併用する必要性が示されたといえる.

6 まとめと今後の展望

本研究では, 自然言語推論タスクを用いた転移学習実験を通じて, 言語間距離と転移学習性能の関係を検証した. 特に, 従来の直感的分類に基づく言語間距離と比べて, 言語ベクトルの類似度に基づく言語間距離を導入することで, 転移学習性能をより適切に説明できる可能性を示した.

今後の展望としては, 本研究で示した言語レベルだけでなく, モデル内部における意味表現の構造を分析するために, 特定の文に関して言語モデルから直接得られる埋め込み表現を用いた, トークンレベルでの距離分析が挙げられる. 本研究では言語レベルの距離に焦点を当てたが, 文レベル表現を用いることで, より細かに言語間距離を捉えることが可能になると考えられる.

また, 本研究は, 低資源言語への応用にも直結する. 低資源言語の言語ベクトルとの類似度を計算することで, 当該低資源言語に対して最も転移学習性能が高い起点言語を選択する指針が得られる可能性がある.

謝辞

本研究は, JSPS 科研費 JP24H00087, JST さきがけ JPMJPR21C2, JST CREST JPMJCR2565, JST BOOST JPMJBY24B2 の支援を受けたものです.

参考文献

[1] Daniela Gerz, Ivan Vulić, Edoardo Maria Ponti, Roi Reichart, and Anna Korhonen. On the relation between linguistic typology and (limitations of) multilingual language modeling. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing**, pp. 316–327, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[2] Rochelle Choenni and Ekaterina Shutova. Cross-neutralising: Probing for joint encoding of linguistic information in multilingual models. **arXiv preprint arXiv:2010.12825**, 2021.

[3] Ajitesh Bankula and Praney Bankula. Cross-linguistic transfer in multilingual nlp: The role of language families and morphology. **arXiv preprint arXiv:2505.13908**, 05 2025.

[4] Yimin Fan, Yaobo Liang, Alexandre Muzio, Hany Hassan, Houqiang Li, Ming Zhou, and Nan Duan. Discovering representation sprachbund for multilingual pre-training. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, **Findings of the Association for Computational Linguistics: EMNLP 2021**, pp. 881–894, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[5] Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. Small data? no problem! exploring the viability of pretrained multilingual language models for low-resourced languages. In Duygu Ataman, Alexandra Birch, Alexis Conneau, Orhan Firat, Sebastian Ruder, and Gozde Gul Sahin, editors, **Proceedings of the 1st Workshop on Multilingual Representation Learning**, pp. 116–126, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[6] Atnafu Lambebo Tonja, Bonaventure F. P. Dossou, Jessica Ojo, Jenalea Rajab, Fadel Thior, Eric Peter Wairagala, Anuoluwapo Aremu, Pelonomi Moiloa, Jade Abbott, Vukosi Marivate, and Benjamin Rosman. Inkubalm: A small language model for low-resource african languages. **arXiv preprint arXiv:2408.17024**, 2024.

[7] Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In Mirella Lapata, Phil Blunsom, and Alexander Koller, editors, **Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers**, pp. 8–14, Valencia, Spain, April 2017. Association for Computational Linguistics.

[8] Matthew S. Dryer and Martin Haspelmath, editors. **The World Atlas of Language Structures Online**. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013. https://wals.info.

[9] Chaitanya Malaviya, Graham Neubig, and Patrick Littell. Learning language representations for typology prediction. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel, editors, **Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing**, pp. 2529–2535, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

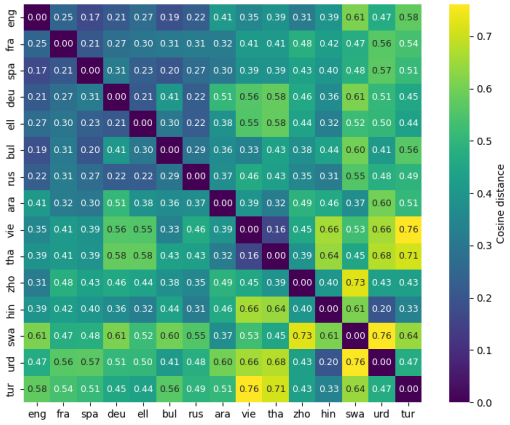
[10] Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. XNLI: Evaluating cross-lingual sentence representations. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing**, pp. 2475–2485, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[11] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 8440–8451, Online, July 2020. Association for Computational Linguistics.

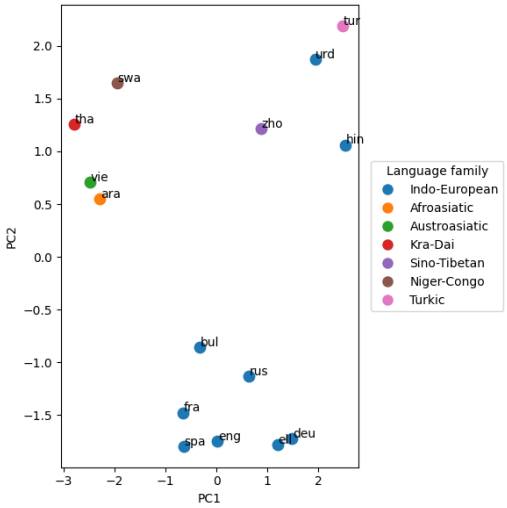
A　XNLI タスクの言語と例文の抜粋（ID=2026）

Language	Premise	Hypothesis	Label
English	What does it matter that you have the whole truth if you can't understand it?	Would you rather not try to understand the truth?	1
Bulgarian	какво значение има, че имаш цялата истина, ако не можеш да я разбереш?	искаш ли да не се опитваш да разбереш истинината?	1
Spanish	¿Qué importa que tengas toda la verdad si no puedes entenderlo?	¿Prefieres no tratar de entender la verdad?	1
Vietnamese	Điều gì quan trọng là bạn có toàn bộ sự thật nếu bạn không thể hiểu được điều đó?	Anh không muốn cố gắng hiểu sự thật sao?	1
Russian	Какое это имеет значение, что у тебя есть вся правда, если ты не можешь понять это?	Ты бы предпочел не пытаться понять правду?	1
Turkish	Ne fark eder ki, eğer bütün gerçeği var?	Gerçeği anlamaya değil mı tercih edersin?	1
French	Qu' est-ce que ça peut faire que tu aies toute la vérité si tu ne comprends pas?	Préféreriez-vous ne pas essayer de comprendre la vérité?	1
Swahili	Ina maana gani una ukweli mzima kama huwezi kuelewa?	Basi, je, wewe utakuwa ni wakweli?	1
German	Was spielt es für eine Rolle, dass du die ganze Wahrheit hast, wenn du es nicht verstehen kannst?	Würdest du lieber nicht versuchen, die Wahrheit zu verstehen?	1
Greek	Τι σημασία έχει που έχεις όλη την αλήθεια αν δεν μπορείς να την καταλάβεις;	Θα προτιμούσες να μην προσπαθήσεις να καταλάβεις την αλήθεια;	1

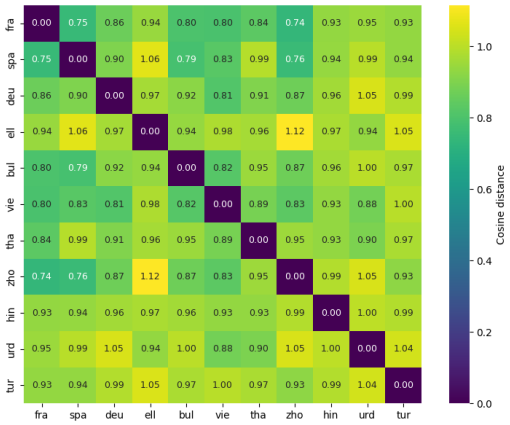
B　実験設定のコサイン距離と PCA



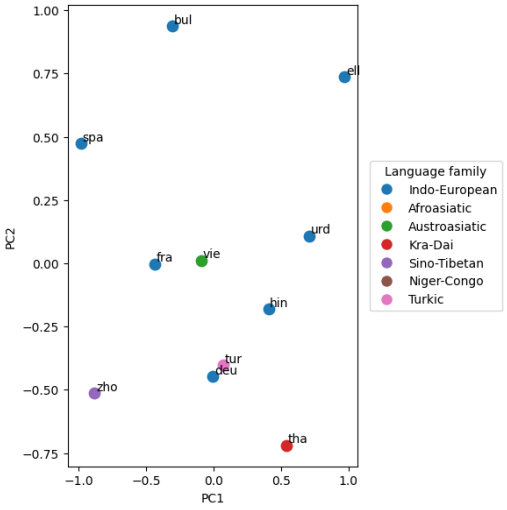
(a) syntax_wals に基づくコサイン距離



(b) syntax_wals に基づく PCA



(c) learned に基づくコサイン距離



(d) learned に基づく PCA