言語モデルの第二言語獲得

大羽未悠 [1]　栗林樹生 [2,3]　大内啓樹 [1,4]　渡辺太郎 [1]
[1] 奈良先端科学技術大学院大学 [2] 東北大学 [3]Langsmith 株式会社 [4] 理化学研究所
{oba.miyu.ol2,hiroki.ouchi,taro}@is.naist.jp
tatsuki.kuribayashi.e8@tohoku.ac.jp

概要

言語モデルの成功を踏まえ，モデルの第一言語（L1）獲得について，人間の言語獲得を踏まえた分析が行われている．本研究では第二言語（L2）獲得にスコープを広げた調査を行う．単言語の事前学習済みモデルを L1 話者と見立て，L2 コーパスを用いた言語間の転移学習により言語転移をシミュレートし，モデルの文法能力について明らかにする．実験では，L1 やその提示方法の違いによって，第二言語の文法獲得が異なる影響を受けることを観察した．

1　はじめに

近年，言語モデルの言語転移能力に高い関心が寄せられている．例えば超大規模な英語の言語モデルは，学習データに少量しか存在しない英語以外の言語においても，ある程度知的な振る舞いを示しており，英語から他言語へ効率的に言語転移していることを示唆している [1, 2]．このような言語モデルの言語転移能力について，既存研究では，パープレキシティなどの抽象度の高い指標や応用タスクでの性能に基づいた調査，特定の学習済み超多言語モデルを対象とした分析などが行われてきた [3, 4, 5]．一方で，文法知識の獲得・転移や，言語ごとの転移傾向の違いといった言語学的な観点からの統制された分析は限られている．

本研究では，**言語モデルの言語転移について，第一言語での学習が第二言語の文法獲得にどのように影響するか**を言語横断的に調査する．具体的には，類型や学習難易度の異なる 4 つの言語（第一言語; L1）を用いて各言語モデルを事前学習した後に，第二言語（英語; L2）を用いて追加の学習を行う．最後に，モデルの L2 の文法能力に関して文法性判断ベンチマークを用いて分析する．

自然言語処理的な視点からは，言語モデルの言語獲得・転移能力について洞察を深める試みと見ることができ，計算心理言語学的な視点からは，母語干渉についてシミュレーション的な検証をしているとみなせる．後者について，人間を直接観察する方法とは相補的な長所がある．例えばあらゆる言語対について，それらを L1・L2 とする人間を集め，彼らの言語能力について統制的に分析をすることには限界があるが，言語モデルでは言語対を増やす・学習の規模を揃えるといった統制・分析が容易に行える．

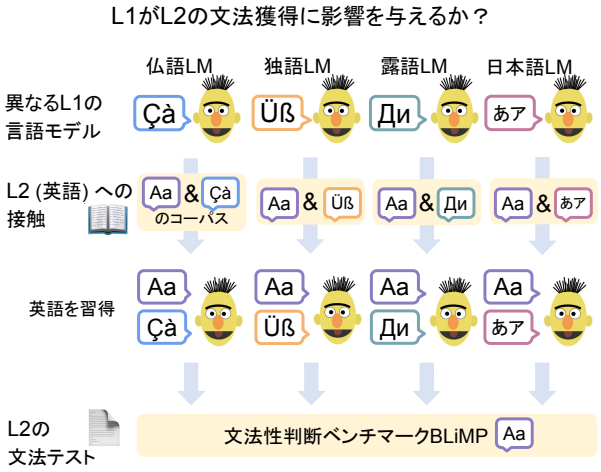

図 1　実験手順．初めに，第一言語（L1）を用いて単言語の穴埋め言語モデリングを学習する（L1 獲得）．次に L1 と L2 を両方用いて既存研究に従った穴埋め言語モデリングを学習する（L2 獲得）[6]．最後に，文法性判断ベンチマーク BLiMP を用いてモデルの文法能力を評価し，L1 が L2 獲得にどのように影響するか調査する．

まず事前実験にて，追学習における L2 の提示方法を比較し，L2 の文法獲得に適した学習設定を調査する．その後，本実験では，L1 を用いた事前学習が L2 の文法獲得に及ぼす影響について，言語の異なりや文法項目の観点から分析する．

実験結果から，**人間の言語学習における英語との乖離が小さい言語での事前学習の方が，英語のモデルの文法獲得に良い影響をもたらす傾向や，特定の言語や文法項目については負の転移を引き起こす場合がある**ことを確認した．例えば，L2 に特有の文

表 1 L1・L2 として実験で用いる言語の性質．FSI は FSI カテゴリを示し，値が大きいほど言語学習において英語との乖離が大きいと想定している．IE はインド・ヨーロッパ語族を，N-IE はそれ以外の語族を指す．

言語	語族	語順	文字	FSI
英語	IE	SVO	アルファベット	-
フランス語	IE	SVO	アルファベット	1
ドイツ語	IE	SOV	アルファベット	2
ロシア語	IE	SVO	キリル文字	3
日本語	N-IE	SOV	かな・漢字	4

法項目については，L1 での事前学習の効果が低く，時に悪影響を与えることが観察された．

また，人間の L2 習得に準えるといくつか直感に反する観察も得られ，必ずしも人間の L2 獲得のアナロジーが通用しないことが示唆された．例えば，言語モデルの L2 の学習時に，L1 と L2 の対訳を提示する設定では，対訳関係を崩したペアを提示する設定よりも，L2 の文法獲得効果が低いことが示された．対訳が提示される場合，語彙的な対応関係を根拠に言語モデリング問題を解けることも多く，この学習負荷の低さが L2 文法獲得効果を下げている可能性がある．一方，人間の言語学習シナリオを踏まえると，母語との対応関係を提示しないほうが L2 文法獲得が促されるというのは直感とは相違する．

2 実験手順

実験手順を図 1 に示す．まず L1 獲得を想定し，言語モデルを L1 の単言語コーパスで事前学習する．次に，L2 獲得を想定し，事前学習済みモデルを L2（英語）を含むコーパスで追学習する．最終的に，言語モデルの文法性判断ベンチマーク（BLiMP）[7] を用いて，L2 におけるモデルの文法能力を評価する．この大枠を基に，L2 学習時の設定の違いや，L1 の異なり，L1 による事前学習の有無などが，L2 の文法獲得に及ぼす影響について実験を行う（3，4 節）．

L2 には英語を使用し，L1 はフランス語，ドイツ語，ロシア語，日本語の 4 言語を調査対象とする（表 1）．これら 4 言語は，アメリカ外交官養成局が報告する英語母語話者にとっての習得難易度（FSI カテゴリ）の観点で異なり，設定の多様性の観点から，各カテゴリにつき 1 言語ずつ採用している．[1] フランス語，ドイツ語，ロシア語，日本語の順で習得が難しくなる．また，既存研究に従い [6]，Transformer ベースの双方向言語モデルを採用し，ハイパーパラメータを定めた（付録参照；表 4）．

2.1 L1 獲得

マスク穴埋め言語モデリング（MLM; Masked Language Modeling）[6, 8] で学習を行う．各言語について，CC-100 からサンプルしたおよそ 100M 語の単言語コーパスを用いた [9, 10]．人間の言語獲得に準え，人間がおよそ 10 歳までに読む単語数（100M 語）と規模を揃えている．

2.2 L2 獲得

多言語を扱うモデルの既存研究 [6] を参考に，L1 と L2 のコーパスを両方用いる設定を想定する．MLM に加え，TLM（Translation Language Model）も使用する．事前実験（3 節）では，MLM や TLM を用いた学習設定の文法獲得への影響を調査する．

コーパスとして，Tatoeba[2] の仏英，独英，露英，日英ペアを使用する．Tatoeba は外国語学習者向けの例文とその翻訳からなる多言語対訳コーパスである．各言語ペアのうち最も文数が少ない言語ペアに合わせて 211,714 ペアを用いた．

2.3 評価

モデルの文法能力を測定するデータセットとして BLiMP [7] を用いる．BLiMP は英語の文法能力を評価対象とし，文法項目ごとに 12 の中分類，67 の小分類からなる．小分類ごとに文法的に容認可能な文と不可能な文のペアが 1000 ペアずつ含まれており，本研究では中分類（以降は文法項目と記す）ごとのスコアとそれらのマクロ平均を報告する．以下は「照応の一致」という項目のペアの例である．(1a) は容認可能な文であるが，(1b) は herself の参照先が存在せず容認不可能な文である．

(1a) Many teenagers were helping themselves.

(1b)* Many teenagers were helping **herself**.

各文の単語を 1 単語ずつマスクして言語モデルに入力し，各単語の確率の相乗平均を求め，容認可能な文と不可能な文のペア全体のうち，前者に高い確率が付与されたペアの割合を計算することで文法性判断スコアを得る．

1) https://www.state.gov/foreign-language-training/ なお，これらの難易度はあくまで英語から該当言語への転移の難しさを示しており，本研究では言語学習難易度に関して転移元・先の対称性を一旦仮定し，該当言語から英語への転移の難しさの議論に持ち出している．

2) https://opus.nlpl.eu/Tatoeba.php

表 2 L2 の学習設定の異なりが文法能力の獲得に及ぼす影響．値は文法性判断スコアを示す．**対訳**の ✓ は，対訳関係のあるコーパスを用いたことを示す．**切替**における ✓ は，L1 側の文を使用するか否かをエポックごとに切り替えて入力したことを表す．

モデル	実験設定		第一言語			
	対訳	切替	仏語	独語	露語	日本語
TLM のみ	✓		51.1	53.6	48.9	51.3
MLM のみ			52.0	57.6	51.2	52.5
MLM+TLM	✓	✓	**58.0**	**61.1**	**52.8**	**56.2**

3 事前実験: L2 の提示方法

初めに，L2 の学習設定による帰納バイアスを調査する．既存研究では，多言語モデルの学習時に対訳を入力しているが，この設定が文法性判断能力の獲得に与える影響を事前に確認する．具体的には，追学習時に，対訳を入力する設定（TLM のみ [6]），対訳関係を崩して入力する設定（MLM のみ [6, 8]），L2 の文に対し，対訳関係のある L1 の文を連結する・しないをエポックごとに切り替えて，TLM と MLM を組み合わせる設定（MLM+TLM）の 3 つを試す．実験設定の詳細は付録 A に記す．

対訳関係の有無の比較（表 2，1 行目と 2 行目の比較）から，TLM のみの設定では MLM のみの設定時よりも L2 の文法獲得の効果が低いことがわかった．TLM の学習では，対訳関係を使用できるため，ソース言語のマスクした単語に紐づくターゲット言語の単語の訳を出力するだけで問題が解ける場合も多い．一方，MLM の学習では，対訳関係を使用できないので，より高負荷な問題を解かせていることに相当し，L2 の文法的理解を促した可能性がある．

また，MLM+TLM 設定においてモデルの文法獲得が最も促された．MLM と TLM が相補的に良い影響を与えた可能性のほか，L2 の文法能力を評価する際は L2 の文を単体で入力しており，学習時にも L2 の文が単体で入力される設定を採用することで，学習と推論の設定の乖離を縮められた可能性がある．以降の実験では，MLM+TLM の設定を採用する．

4 実験: L1 の影響

事前学習で用いた **L1 の違い**がモデルの L2 獲得に与える影響を調査する．各 L1 の単言語コーパスを用いた事前学習を行うか否かで文法能力を比較する．追学習終了時の文法能力を表 3 に示す．

全ての言語に見られる傾向: 表 3 は L1 の事前学習が L2 の文法獲得に与える傾向を示している，各文法項目におけるスコアのマクロ平均 OVERALL の事前学習の有無によるスコア差 Δ から，4 言語全てにおいて，L1 での事前学習により文法能力が向上していることがわかる．言語間で差はあるが，L1 での事前学習は L2 の文法獲得に好影響を促すことが示唆された．

各文法項目におけるスコアの変化に着目すると，IRR. FORM（動詞の不規則活用）については，相対的に事前学習が悪影響を及ぼしていることが分かる．IRR. FORM では，例えば以下の（6a）は容認可能な文で，（6b）は容認不可能な文となる．

（6a） The forgotten newspaper article was bad.

（6b）* The **forgot** newspaper article was bad.

この例では動詞 “forget”の過去分詞形が “forgotten”であることなどが問われている．正しい判断のためには英語固有の動詞の活用を覚える必要があり，英語以外の言語での事前学習が良い影響を与えないという結果は直観的であると言える．なぜドイツ語の設定においてより強い悪影響が観察されたかに関しては，今後の分析課題としたい．

言語の異なりがもたらす傾向: 表 3 の各文法項目におけるスコアのマクロ平均 OVERALL を言語間で比較すると，フランス語が最も高く，ドイツ語が僅差で続き，日本語，ロシア語はこれら 2 言語とは大きく差が開いている．フランス語とドイツ語での事前学習は，日本語とロシア語よりもはるかに英語の文法獲得に効果的であることがわかった．このような結果は，言語間の文字体系（表 1 を参照）の類似と関わる可能性がある．また，表 1 の FSI を参照すると，人間の習得はフランス語，ドイツ語，ロシア語，日本語の順に難しくなることから，FSI の傾向がある程度は言語モデルでも観察された．L1 における L2 習得の難しさは，言語モデルと人間でおおよそ類似していることが示唆される．

文法項目ごとのスコアから，L1 ごとに得意・不得意な文法項目が異なることが分かる．例えば，フィラー・ギャップ依存関係（FILLER-GAP）では，同じ SOV 語順の言語間でも，ドイツ語と日本語モデルで 3.7 ポイントの差が生じており， 言語現象の類似・相違性と一貫している．フィラー・ギャップ依存関係は，例えば以下の（2a）は容認可能な文で（2b）

表 3 L1 の事前学習や言語の異なりが L2 の文法獲得に与える傾向．OVERALL は各文法項目のスコアのマクロ平均を表す．L1 の ✓ は，L1 の単言語コーパスを用いて事前学習したことを示す．Δ は事前学習の有無によるスコア差である．値が大きいほど，事前学習が文法獲得に対してより良い影響を与えたことを示す．

言語	L1	OVERALL	ANAPHORA	ARG. STR.	BINDING	CTRL. RAIS.	D-N AGR.	ELLIPSIS	FILLER-GAP	IRR. FORM	ISLAND	NPI LICENSE	QUANTIFIERS	S-V AGR.
フランス語	✓	58.0	55.8	55.4	51.8	58.6	69.5	67.7	54.6	73.0	52.2	40.5	56.5	60.4
	Δ	5.3	2.3	7.2	1.9	1.5	14.5	17.0	4.5	3.8	−0.8	3.5	−1.2	8.8
ドイツ語	✓	61.1	43.1	53.1	65.2	52.2	68.7	63.5	68.2	69.3	47.7	54.6	80.5	67.0
	Δ	5.2	5.9	4.6	1.4	−2.2	11.1	14.7	4.8	−11.5	4.5	9.8	4.6	14.3
ロシア語	✓	52.8	52.9	47.0	40.7	61.4	58.6	54.2	52.4	72.7	49.3	32.8	56.2	54.9
	Δ	0.7	−3.1	0.5	−0.1	0.3	3.2	5.2	4.1	0.3	−1.9	1.1	−4.5	2.9
日本語	✓	56.2	61.5	52.1	61.0	57.8	65.8	55.3	51.3	70.5	54.0	41.0	50.6	53.0
	Δ	1.5	0.7	2.0	0.8	2.3	5.0	4.3	1.1	−3.3	4.6	2.2	−2.9	1.1

は容認不可能な文といった判断を課す問題設定に対応している．

（2a） Joel discovered the vase that Patricia took.

（2b）* Joel discovered **what** Patricia took **the vase.**

例えば，（3a）から（3b）のように格要素を移動して従属節を作る際，英語ではギャップ（移動前の位置，記号「___」の箇所）がフィラー（移動先，[the student]）の後ろに位置することが多い．

（3a） The teacher advised [the student].

（3b） [The student] the teacher advised ___ were smart.

ドイツ語でも同様の位置関係になるが，日本語では（4a-b）のように逆となる．

（4a） 先生が [学生に] アドバイスした．

（4b） 先生が___アドバイスした [学生は] 賢かった．

このように，日本語-英語間で各要素の移動先に違いがあり，日本語で事前学習したモデルが英語のフィラー・ギャップ依存関係の学習に苦戦する理由のひとつになっていると考えられる．

5 関連研究

「ニューラルモデルはテキストデータのみから人間の言語獲得を模倣できるのか」．この問いに取り組んだ研究は 1980 年台に始まり，先天的な知識なしに言語獲得は可能かという問いや，コネクショニズムの可能性・限界の観点から，議論が繰り広げられてきた [11, 12]．当初は簡易的なニューラルモデルを用いて議論が広げられたが，近年ニューラルモデルを用いた自然言語処理が目覚ましい進展を遂げ [13]，ニューラルネットワーク黎明期に認知科学分野が掲げた問いへ再訪する動きが高まっている [14, 15]．近年盛んに行われているニューラルモデルの言語知識の分析（プロービング）は，そのような一連の議論の延長線上にある [16, 17]．既存研究では L1 の獲得に注目が置かれてきたが，本研究ではニューラル言語モデルの第二言語獲得の傾向を分析しており，多言語モデリングという工学的道具立ての性質の理解と共に，人間の言語転移・第二言語獲得における母語干渉などへの計算機的なアプローチを見据えている．

言語転移については，言語間の転移学習により構文知識や文法誤り知識を転移することで，構文解析 [18] や文法誤り訂正 [19] などの下流タスクに活用する研究がなされている．人工言語を用いた言語転移の研究も行われてきており，楽譜や括弧からなる系列といった言語以外の系列からの転移や [3, 20]，自然言語を規則的に編集することで得られた言語からの転移なども分析されている [4]．本研究は，より人間の学習に条件を近づけた設定で，言語モデルの L1 の L2 への影響やその過程を調査し，言語間の転移能力について分析している．

6 おわりに

本研究では，言語モデルの言語転移について，第二言語における文法獲得への影響という観点から調査を行った．その結果として，L1 は L2 の文法獲得に対し全体的には好影響を与えるが，言語や文法項目に依存し，負の転移を引き起こす場合もあることがわかった．今回得られた結果に対する言語学的観点からの考察の充実化や，より多くの言語を用いた検証を今後の課題としたい．

謝辞

本研究は JSPS 科研費 JP19K20351 の助成を受けたものです.

参考文献

[1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, **Advances in Neural Information Processing Systems**, Vol. 33, pp. 1877–1901. Curran Associates, Inc., 2020.

[2] Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, et al. Language models are multilingual chain-of-thought reasoners. **arXiv preprint arXiv:2210.03057**, 2022.

[3] Isabel Papadimitriou and Dan Jurafsky. Learning Music Helps You Read: Using transfer to study linguistic structure in language models. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 6829–6839, Online, November 2020. Association for Computational Linguistics.

[4] Ameet Deshpande, Partha Talukdar, and Karthik Narasimhan. When is BERT multilingual? isolating crucial ingredients for cross-lingual transfer. In **Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 3610–3623, Seattle, United States, July 2022. Association for Computational Linguistics.

[5] Terra Blevins, Hila Gonen, and Luke Zettlemoyer. Analyzing the mono- and cross-lingual pretraining dynamics of multilingual language models. In **Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing**, pp. 3575–3590, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[6] Alexis Conneau and Guillaume Lample. Cross-lingual language model pretraining. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, **Advances in Neural Information Processing Systems**, Vol. 32. Curran Associates, Inc., 2019.

[7] Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. BLiMP: The benchmark of linguistic minimal pairs for English. **Transactions of the Association for Computational Linguistics**, Vol. 8, pp. 377–392, 2020.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[9] 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 **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 8440–8451, Online, July 2020. Association for Computational Linguistics.

[10] Guillaume Wenzek. Ccnet: Extracting high quality monolingual datasets from web crawl data. In **Proceedings of the Twelfth Language Resources and Evaluation Conference**, pp. 4003–4012, Marseille, France, May 2020. European Language Resources Association.

[11] David E Rumelhart and James L McClelland. On learning the past tenses of english verbs. In **Parallel distributed processing: Explorations in the microstructure of cognition**, pp. 216–271. MIT Press, Cambridge, MA, 1986.

[12] Steven Pinker and Alan Prince. On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. **Cognition**, Vol. 28, No. 1, pp. 73–193, 1988.

[13] Christopher D. Manning. Last words: Computational linguistics and deep learning. **Computational Linguistics**, Vol. 41, No. 4, pp. 701–707, December 2015.

[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. 651–665, 2018.

[15] R Thomas McCoy, Robert Frank, and Tal Linzen. Revisiting the poverty of the stimulus: hierarchical generalization without a hierarchical bias in recurrent neural networks. In **In Proceedings of the 40th Annual Conference of theCognitiveScienceSociety**, pp. 2093–2098, Madison, WI, 2018.

[16] Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. **Transactions of the Association for Computational Linguistics**, Vol. 4, pp. 521–535, 2016.

[17] Alex Warstadt and Samuel R. Bowman. Can neural networks acquire a structural bias from raw linguistic data? In **In Proceedings of the 42nd Annual Meeting of the Cognitive Science Society**, 2020.

[18] Wasi Ahmad, Haoran Li, Kai-Wei Chang, and Yashar Mehdad. Syntax-augmented multilingual BERT for cross-lingual transfer. In **Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)**, pp. 4538–4554, Online, August 2021. Association for Computational Linguistics.

[19] Ikumi Yamashita, Masahiro Kaneko, Masato Mita, Satoru Katsumata, Imankulova Aizhan, and Mamoru Komachi. Grammatical error correction with pre-trained model and multilingual learner corpus for cross-lingual transfer learning. **Natural Language Processing**, Vol. 29, No. 2, pp. 314–343, 2022.

[20] Ryokan Ri and Yoshimasa Tsuruoka. Pretraining with artificial language: Studying transferable knowledge in language models. In **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 7302–7315, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[21] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open source toolkit for statistical machine translation. In **Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions**, pp. 177–180, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[22] Aaron Mueller, Garrett Nicolai, Panayiota Petrou-Zeniou, Natalia Talmina, and Tal Linzen. Cross-linguistic syntactic evaluation of word prediction models. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 5523–5539, Online, July 2020. Association for Computational Linguistics.

表 4　ハイパーパラメータ

dropout, attention_dropout	0.1
accumulate_gradients	4
emb_dim	256
gelu_activation	True
Optimizer	adam_inverse_sqrt
	lr=0.00020, eps=0.000001
	warmup_updates=30000
	beta1=0.9,beta2=0.999
	weight_decay=0.01
epoch	100
n_heads, n_layers	8, 12
amp, fp16	2, True

A　実験設定

L1 獲得　トークナイザとして日本語は kytea[3] を，フランス語とドイツ語とロシア語は mosesdecoder [21] を使用し，fastBPE[4] でサブワード分割を行った．語彙数は 14,000 を設定した．ハイパーパラメータは表 4 に記載している．

L2 獲得　3 節の TLM は対訳コーパスをそのまま用いた対訳関係のある設定，MLM はコーパスの片方の言語をシャッフルして対訳関係を崩した設定，MLM ＋ TLM は L2 側の文へ対訳関係のある L1 側の文を連結する・しないをエポックごとに切り替える設定である．これらのモデルの MLM は，Conneau ら [6] の提案モデルを使用している（図 2 上部），言語埋め込みを使用し，入力文を全て連結し 256 トークンごとに切断した文を入力する．ただし 2 文を用いた設定の場合は文対を入力する．TLM は Conneau らが提案した MLM の一種である（図 2 下部）．MLM とは入力の際に対訳関係のある 2 文を連結する点で異なる．単言語コーパスを用いたモデルの BPE の学習コードと語彙は，単言語コーパスで使用したものに対訳コーパスの英語のものを追加し，重複したトークンや語彙を除く方法で作成した．用いていないモデルでは，対訳コーパスの両方の言語から BPE の学習コードと語彙を作成した．埋め込み層を語彙数方向に増やしたことに伴い，最終層の重み・バイアスも増やしている．サブワード分割のためのトークナイザは，英語は mosesdecoder [21] を使用し，他の言語は事前学習と同じ設定である．ランダムシードを 3 つ用いて，スコアはその平均を計算した．

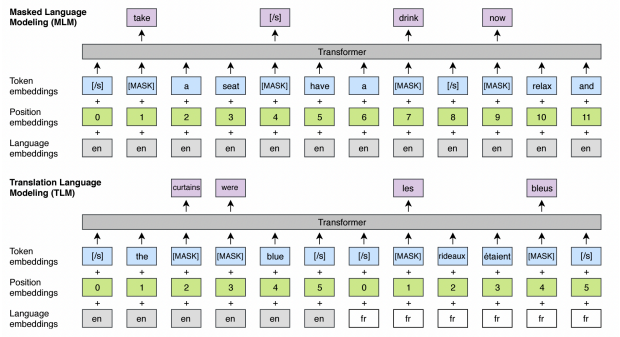

図 2　conneau らの提案モデル

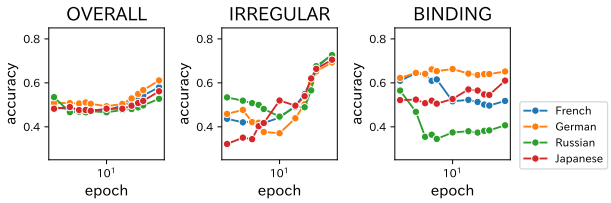

図 3　L2 学習中の各エポックにおける文法能力（抜粋）

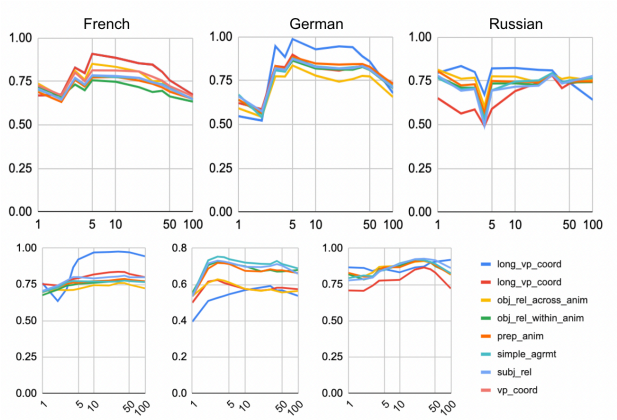

図 4　L2 学習中の各エポックにおける L1 の文法能力

B　学習過程の分析

B.1　L2 の文法獲得過程

L2 学習の過程を分析するため，追学習時の各エポックにおける文法能力を評価（図 3）した．[5] OVERALL から，学習を重ねることによりおおむね文法能力が向上することが示唆される．L1 の異なりの影響について，例えば IRREGULAR（動詞の不規則活用）では，追学習初期段階では L1 ごとに性能が大きく異なるが，L2 の継続的な学習によりそれらの差異は縮まっている．一方で，BINDING（束縛）のように，どの言語もスコアが横ばいまたは低下傾向のある項目も存在する．

B.2　L2 の文法獲得の L1 への影響

L2 文法の獲得時における L1 の文法能力の変化について調査を行う．L2 学習時の B.1 節と同様のエポックにおける L1（フランス語，ドイツ語，ロシア語）の文法能力を評価した（図 4 上部）．文法能力を測るために，CLAMS [22] という多言語の文法能力評価データセットを用いる．L2 を含む追学習をおこなった場合，フランス，ドイツ語では一時的に好影響，ロシア語では悪影響を受けた，どの言語も文法能力が減少する傾向にある．

アブレーションとして，L1 のみで追学習を行う実験も試した（図 4 下部）．全体の傾向として，スコアはある程度向上した後に横ばいとなる．

これらの傾向は，学習の後半では L2 の学習による L1 の忘却の発生を示唆している．正則化項の設計などの忘却を緩和するような方法の考案は今後の課題となる．

3）　http://www.phontron.com/kytea/
4）　https://github.com/glample/fastBPE
5）　エポック数 1，2，3，4，5，10，20，30，40，50，100 の場合をそれぞれ評価した．