Vol.2022-NL-254 No.22
Vol.2022-SLP-144 No.22
2022/12/1

言語モデルの第二言語獲得効率

大羽 未悠[1,a)]　栗林 樹生[2,3]　大内 啓樹[1,4]　渡辺 太郎[1]

概要：言語モデルの成功を踏まえ, モデルの言語獲得過程や有する言語知識について, ヒトの言語獲得と対照した分析が行われている. 既存研究ではヒトと言語モデルの第一言語獲得に焦点が当てられていたが, 本研究では第二言語獲得（習得）の過程・効率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第一言語話者と見立て, 第二言語となる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, その効率や獲得された知識について明らかにする. 第一言語として語順や文字体系などの言語学的特徴が異なる言語を用いて言語転移を比較する.

Efficiency of language models to acquire second languages

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. 例えば超大規模英語言語モデルは, 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, 英語から他言語への効率的な言語転移が示唆されている [3], [19]. 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, パープレキシティといった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, 特定の訓練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 [2], [6], [14]. 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, ある言語での訓練が第二言語の文法獲得効率にどのように干渉するか**を, 言語横断的に調査する. 具体的には, 類型的に異なる 4 言語（第一言語）それぞれで言語モデルを事前訓練した後, 第二言語として英語で追加の言語訓練を行う. このとき, 第一言語や 2 言語を用いた訓練の設定が, 英語 (第二言語) の文法獲得にどのような影響を与えるかについて, 文法性判断ベンチマークデータを用いて分析する.

自然言語処理的な視点からは, 「人間の言語転移で示唆されている傾向が言語モデルでも観察されるのか」という問いを通して, 言語モデルの言語獲得・転移能力について洞察を深める. 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. この見方については, 人間を直接観察する方法論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, それらを第一・第二言語とする人間を集め, 彼らの言語能力について統制のとれた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, 学習データの規模を揃えるといった統制・分析が容易に行える.

実験では初めに, 第一言語の異なりが, 第二言語（英語）の獲得・習得に異なる学習バイアスを与えることを確認する. 次に, 第二言語を用いた追学習について, 学習設定（例えば第二言語の文と共に第一言語の対訳を入力するか）の帰納バイアスを調査する. 最後に, 大量の第一言語データを用いた事前学習を実施する場合としない場合を比較し, 事前学習が及ぼす影響を詳細に分析する.

実験結果から, **第一言語での学習は, 第二言語の学習に影響を与え, その傾向は著しく言語依存である**ことが確認された. 例えば今回の実験設定では, 英語と類型的に大きく異なる日本語での事前学習は, その後の英語の文法性獲得に対して, 比較的小さな影響しかもたらさないことが示唆された. 特定の文法項目や言語については, 第一言語での学習が**負の転移**を引き起こすことも確認され, 必ずしも第一言語での学習が効率的な第二言語学習を促すとは限らないことも確認できた.

また, 人間の第二言語習得に準えると, いくつか直観に反する観察も得られた. 例えば, 言語モデルに第二言語を学

[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
[a)] miyu.oba.ol2@is.naist.jp

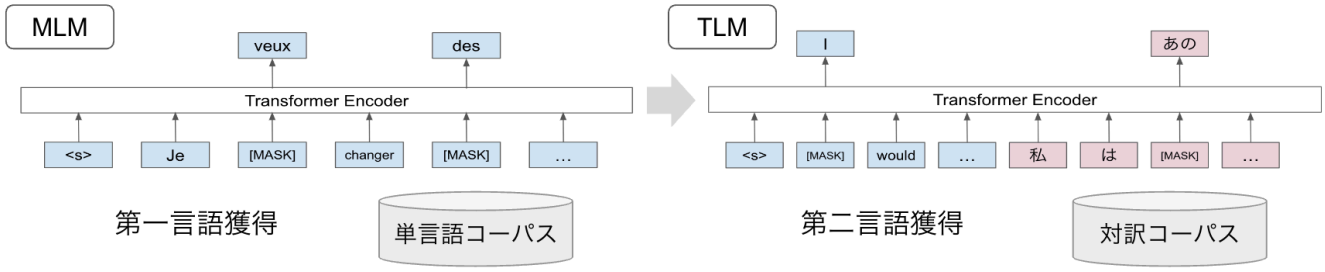

図 1: 言語モデル学習の手順． 左が第一言語獲得時の学習設定（MLM; Masked Language Modeling）であり， 通常の穴埋め言語モデリングと同様, 入力の一部をマスクし, 元々存在したシンボルを出力できるよう訓練を行う． 右が第二言語獲得時の学習設定（TLM; Translation Language Modeling）であり, 既存研究に従い二言語の対訳を入力し, MLM 同様マスク穴埋めができるよう訓練を行う [5].

習させる際, 第一言語の対訳の提示は第二言語の文法獲得を妨げることが示された. 対訳の存在により言語モデリングの難易度が下がるためであると考えられるが, 人間の言語学習シナリオを踏まえると, 母語との対応情報が第二言語習得を阻害するというのは非直観的である. このような知見は, 言語モデルの言語転移について, 必ずしも人間の第二言語獲得のアナロジーが通用しないことを示唆している.

2. 実験手順

実験手順の全体像を**図 1** に示す. まず第一言語獲得を想定し, 言語モデルを特定言語の単一言語コーパスで訓練する. 次に, 第二言語獲得を想定し, 学習済みモデルを第二言語（英語）を含むコーパスで追学習する. 最終的に, 言語モデルの文法能力測定データ（BLiMP） [21] で, 第二言語（英語）におけるモデルの文法能力を評価する. この大枠のもと, 第一言語の違いや, 第二言語学習時の設定の違い, 第一言語による学習の有無が第二言語獲得に及ぼす影響について, 実験を行う（3 節）.

第一言語については, 具体的に, フランス語, ドイツ語, ロシア語, 日本語の 4 言語を調査対象とし, 第二言語には英語を使用する（**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, 各カテゴリから 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, 日本語の順で習熟が難しくなる. また, 多言語言語モデリングを行う既存研究（XLM）に従い [5], 穴埋め方式の言語モデリング目的関数と Transformer ベースの双方向言語モデルを用いる.

[*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特定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, ある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.

表 1: 第一言語として実験で用いる 4 言語と英語の性質. 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. 語族について IE はインド・ヨーロッパ語族を, N-IE は非インド・ヨーロッパ語族を指す.

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	N-IE	SVO	キリル文字	3
日本語	N-IE	SOV	かな・漢字	4

2.1 第一言語獲得

図 1 左部のようにマスク穴埋め言語モデリング (MLM; Masked Language Modeling) 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 [4], [22]. 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 単語）と規模を揃えている. マスクの確率や， モデルのハイパーパラメータは， 多言語言語モデルを訓練する既存研究に従った [5] （付録参照；**表 A·1**）.

2.2 第二言語獲得

既存研究 (XLM) に従い, 対訳データを用いて第二言語獲得を行う設定を一旦想定する [5]. 対訳データを用いることの効果については, 実験 (3 節) で調査する. 図 1 右部のように, 対訳データを連結して入力し, MLM マスク穴埋め型の言語モデリングを行う (TLM; Translation Language Modeling). 対訳コーパスとして, Tatoeba[*2]の日英, 仏英, 独英, 露英ペアを使用する. 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

[*2] https://opus.nlpl.eu/Tatoeba.php

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るための評価データセットとして, BLiMP [21] を用いる. BLiMP は英語におけるモデルの文法能力を評価するデータであり, 文法項目について, 12 の中分類, 67 の小分類からなる. 本研究では, 中分類ごとの性能とそれらのマクロ平均を報告する. 各文法項目には, 文法的・意味的に容認できる文とそうでない文のミニマルペア（分析したい観点が分析できる最小の差をもつペア）が 1000 ペア存在する. 67 項目の詳細と例は, Warstadt らの論文を参考にされたい [21]. 例えば, 以下は照応の一致という項目に含まれるペアの例であり, 例（1a）は容認可能な文であるが, 例（1b）は herself の参照先が存在せず容認できない文である.

（1a） Many teenagers were helping themselves.

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

それぞれの文について, 1 単語ずつをマスクして言語モデルに入力して, マスクした各単語の確率を得る. 各文について構成単語の確率の相乗平均を求め, どちらに高い確率が付与されたかを比較する [10]. 全文ペアのうち, 容認性の高い文に高い確率が付与されたペアの割合を計算する.

3. 実験

3.1 実験 1: 第一言語の違いと文法能力獲得過程

初めに, 事前学習で用いた**第一言語の違い**により, 言語モデルの英語学習にどのような影響が生じるかを調査する. 第一言語で単言語学習をしたのち, 既存研究に従い, 対訳データを用いて第二言語の学習を行った [5]. 追学習終了時の文法性判断能力を表 2 に示す.

どの第一言語が英語の文法獲得を有利にするか？平均スコアでは, ドイツ語が最も高く, 次いでフランス語, ロシア語, 日本語の順で小さくなるものの, ドイツ語以外の 3 言語間については, 大きな差は得られなかった. また, データ効率の観点から, 追学習開始 5 エポックにおけるエポック毎平均スコア（表 2 内, 効率）も調査したが, 平均スコアと同様に第一言語としてのドイツ語が良い影響を与えており, 他 3 言語の傾向に大きな差はなかった. ただし, 前述の通り, 項目ごとの得意不得意については言語間で異なる特徴が得られている. FSI カテゴリの観点では, 英語に最も近いフランス語で高いスコアが得られ, ロシア語や日本語では低いスコアが出ることを予想していたが, そのような傾向は得られず, **言語ごとの第二言語獲得の難易度は, 言語モデルと人間では異なる**ことが示唆される.

文法項目ごとの第一言語の影響: 表 2 から, 第一言語ごとに得意・不得意な文法項目が異なることが分かる. また第一言語の影響を受けやすい文法項目とそうでない項目があることも示唆された（表 2 内, Max−Min 行）. 例えば, NPI（否定極性項目）, FILLER-GAP（フィラー・ギャップ依存関係）, QUANTIFIERS（量化）などでは, 第一言語の違いによる影響が大きく出ている.

フィラー・ギャップ依存関係では, 同じ SOV 語順の言語間でも, ドイツ語事前学習と日本語モデルで 27.6 ポイントの開きが生じている. ここでフィラー・ギャップ依存関係は, 例えば以下の例（2a）は文法的に正しい文であるが, 例（2b）は非文である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.**

このような判断において, ドイツ語で獲得が有利になり日本語で獲得が不利になる観察は, 言語現象の類似/相違性とは一貫している. 例えば, 例（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, ___）がフィラー（移動先, [the student]）の後ろに存在する位置関係になることが多い.

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

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

例（4a-b）のようにドイツ語でも同様の語順になるが, 日本語では例（5a-b）のようにフィラーとギャップの位置関係が逆になる.

（4a） Der Lehrer riet [denm Schüler].

（4b） [Die Schüler] die der Lehrer riet ___ waren klug.

（5a） 先生が [学生に] アドバイスした.

（5b） 先生が___アドバイスした [学生は] 賢かった.

日本語言語モデルが英語におけるフィラー・ギャップ依存関係の学習に苦戦する結果について, 一つの仮説としては, 日本語-英語間のこのような語順の乖離に紐づく可能性が挙げられる. その他, ロシア語言語モデルが NPI（否定極性項目）の汎化に弱いといった他の目立った結果についても, 言語的な性質と紐づけた解釈や仮説の提示をすることは今後の課題としたい.

文法能力獲得過程: また, 第二言語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（表 2）した.[*3] Overall からは, 学習を重ねることによりお

[*3] 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.

表 2: 異なる第一言語を用いた場合の文法獲得傾向を示す．平均は全文法項目に関する正解率のマクロ平均を，効率は追学習開始時 5 エポックにおける平均スコアのエポック毎平均を示す．Max−Min は，各文法項目ごとの最も高い/低い正解率間の差であり，第一言語の違いによる影響の大きさの目安として示す．

第一言語	平均	効率	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	51.1	48.3	53.7	52.7	46.9	52.3	55.6	65.5	42.5	53.9	52.8	42.9	45.0	49.6
ドイツ語	55.7	50.3	36.3	54.7	61.1	47.7	61.4	58.1	64.6	60.9	44.3	53.9	67.4	57.5
ロシア語	50.1	47.9	57.0	49.3	39.4	55.6	51.6	60.5	51.1	59.6	47.7	25.5	53.0	51.3
日本語	50.6	48.0	50.0	52.4	55.7	53.1	55.2	62.1	37.0	55.7	44.2	44.8	44.7	52.2
Max−Min	5.6	2.4	**20.7**	5.4	**21.7**	7.9	9.84	7.4	**27.6**	7.0	8.6	**28.4**	**22.7**	7.9

表 3: 第二言語学習設定の異なりが文法能力獲得に及ぼす影響．値は BLiMP における平均スコアを示す．対訳における ✓ は，対訳関係を崩さずに対訳コーパスを用いたことを示す． Drop における ✓ は，第一言語側の文を確率的に削除してモデルに入力したことを表す．

実験設定		第一言語			
対訳	Drop	フランス語	ドイツ語	ロシア語	日本語
		54.6	57.1	50.6	51.9
✓		51.1	55.7	50.1	50.6
✓	✓	**56.9**	**60.8**	**54.4**	**56.1**

おむね文法能力が向上することが示唆される．第一言語の異なりの影響については，例えば IRR. FORM（動詞の不規則活用）では，追学習初期段階では第一言語ごとに性能が大きく異なるが，第二言語の継続的な学習によりそれらの差異は縮まっている．また，ANAPHORA（照応の一致）では，ある第一言語の場合は数エポック目で大きくスコアが減少した後に徐々に改善し，ある言語では改善することなく徐々に悪化していくというような現象も確認された．全ての言語現象の詳細は図 A·1 に記載している．

3.2 実験 2: 第二言語の提示方法の影響

次に，第二言語学習設定による帰納バイアスを調査する．既存研究では，多言語訓練時に対訳を入力しているが，文法性判断能力の獲得の観点から，この設定が適切であるかを調査する．具体的には，（i）対訳データをそのまま用いる設定の他に，（ii）対訳関係を崩して入力する場合と，（iii）対訳データのうち，第一言語側の文全体を適当な確率で入力しない設定（drop）を試した[*4]．対訳関係の有無の比較（表 3, 1 行目と 2 行目の比較）から，対訳を常に与える設定では，第二言語における文法獲得が相対的に妨げられることがわかった．対訳が常に入力される場合，穴埋め言語モデリングの問題は，両言語に出現する単語の紐付けを行い，片方の言語に存在しない単語を翻訳したものを出力するといった，語彙的な翻訳知識に基づく解法である程度解けてしまう可能性がある．単一言語での言語モデリングを課すことは，ある種高負荷な問題を解かせていることに相当し，第二言語におけるより高度な言語理解が促される可能性がある．

また，入力対訳データのうち第一言語側を適当な確率で削除する設定において，モデルの文法獲得が最も促されることがわかった．BLiMP で文法性を問う際には，第二言語の文のみを入力して確率を計算している．Drop 以外の設定では，学習時に必ず 2 言語のペアが入力されており，確率計算時の設定との乖離が悪影響を生んでいる可能性がある．

3.3 実験 3: 第一言語の観察量と文法能力の変化

最後に，各第一言語について，単一コーパスを用いた事前訓練を行った場合と行わなかった場合で文法能力を比較し，第一言語の大規模事前学習による文法能力の変化を調査する．実験 2 の結果を踏まえ，第二言語学習では対訳を用い，適当な確率で第一言語側の文を削除する．すなわち，事前訓練を行わない設定では，追学習時の対訳データ中にのみ第一言語が出現する．

平均スコアに注目すると， 今回検証した 4 言語全てにおいて， 第一言語での事前学習により文法能力の向上が見られている．このことから， 言語間で違いはあるものの，第二言語以外の言語での事前学習はその後の多言語での文法獲得に良い影響を促すことが示唆された．英語以外の言語を第二言語とした場合の検証は今後の課題とする．

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， どの言語でも事前学習が悪影響を及ぼしていることが分かる．この項目では，例えば以下の例（6a）は文法的に正しく，例 (6b) は非文であるといった，不規則活用に関する知識が問われている．

[*4] 具体的には，エポックごとに第一言語の対訳を入力する/しない設定を交互に適用した

表 4: 第一言語で事前学習をすることの英文法獲得への影響. L1 列における ✓ は, 第一言語単言語コーパスでの事前学習の実施を表す. Δ は, 事前学習を行った時と行わなかった時の BLiMP における正解率の差を示し, この値が大きいほど事前学習が文法獲得を有利にする影響を与えたことを示す.

第一言語	L1	平均	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	✓	56.9	56.4	54.2	50.4	57.8	67.7	69.0	53.2	73.0	51.7	37.6	55.0	56.2
		53.5	56.1	50.5	50.3	57.3	57.5	47.1	49.6	75.0	51.9	36.3	57.5	52.5
	Δ	3.4	0.3	3.7	0.1	0.5	10.2	21.9	3.6	−2.0	−0.2	1.3	−2.5	3.7
ドイツ語	✓	60.8	40.9	52.8	64.2	52.9	69.2	61.3	67.9	70.2	48.3	56.4	80.0	65.3
		56.9	36.6	48.7	67.0	54.8	58.1	48.8	61.1	84.2	44.4	45.5	79.8	53.6
	Δ	3.9	4.3	4.1	−2.8	−1.9	11.1	12.5	6.8	−14.0	3.9	10.9	0.2	11.7
ロシア語	✓	54.4	51.4	47.9	42.6	61.7	60.8	59.8	54.4	72.5	53.5	37.9	54.4	56.0
		51.7	57.8	46.0	36.7	62.0	54.3	46.8	49.0	74.5	48.9	32.2	60.9	51.2
	Δ	2.7	−6.4	1.9	5.9	−0.3	6.5	13.0	5.4	−2.0	4.6	5.7	−6.5	4.8
日本語	✓	56.1	64.6	52.4	58.7	58.4	66.6	54.3	53.4	67.6	55.0	39.3	49.9	53.1
		54.4	62.1	50.4	60.3	56.4	61.3	52.5	50.1	74.4	49.3	32.9	51.7	51.2
	Δ	1.7	2.5	2.0	−1.6	2.0	5.3	1.8	3.3	−6.8	5.7	6.4	−1.8	1.9

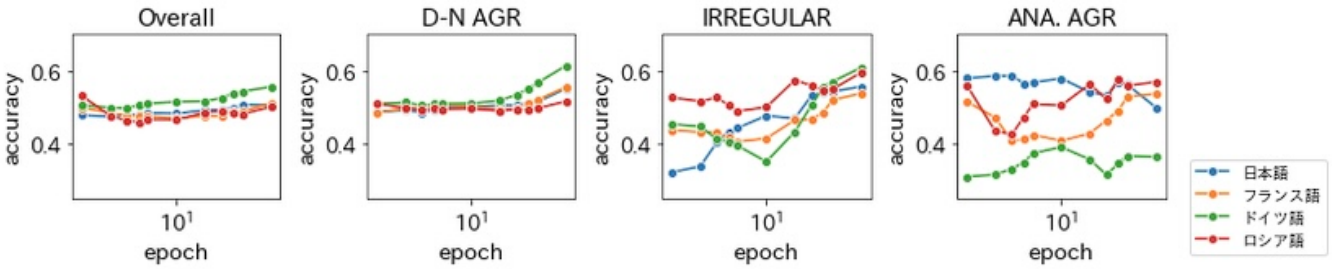

図 2: 第二言語学習中の各エポックにおける文法能力の評価結果（抜粋）

（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（かつ forgotten/forgot の位置の動詞は過去分詞形であるべきこと）が問われている．この項目については, 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, ある種直観的である．なぜドイツ語の設定において特に著しい悪影響が観察されたのかといった点への言語学的な考察は今後の課題としたい.

また各言語における事前学習の効果について, ある言語や文法項目における結果は言語的な性質に即しているが, ある部分では即していないことも多い．例えば, ANAPHORA（照応の一致, 例 1）について, ドイツ語は {him, her, it, your, them}self などを同じ単語で表し, 英語でのこれらの単語の使い分けに苦戦することが予測されるが, 4 言語の中ではドイツ語において最も性能向上が得られている．どのような言語的要因が第一言語学習のバイアスに紐づくのかといった解釈は, 今後の課題としたい．特定の現象が英語のものと似ているかといった観点のみならず, その現象が第一言語でどの程度生じやすいかといった頻度なども影響を与えると予想される.

4. 関連研究

ニューラルモデルがデータのみから人間の言語獲得を模倣できるのかといった調査は 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 [15], [17]. 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 [12], 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 [8], [13]. 近年盛んに行われているニューラル言語処理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 [11], [20]. 既存研究では母語（単一言語）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 [1] や文法誤り訂正 [7] 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 [14], [16], 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 [6]. また, 第二言語話者の文法誤りの産出を直接予測するような問題設定も提案されてきた [18]. 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.

5. 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の獲得への影響という観点から調査を行った. 第一言語の異なりが文法獲得傾向に違いを及ぼすこと, 学習設定次第では第二言語での文法獲得が妨げられること, 第一言語での事前学習の影響が特定の文法項目に紐づくことなどが観察された. 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や, より多くの言語を用いた検証, 英語を第一言語とした場合の調査などは今後の課題としたい.

参考文献

[1] Ahmad, W. U., Li, H., Chang, K. and Mehdad, Y.: Syntax-augmented Multilingual BERT for Cross-lingual Transfer, *CoRR*, Vol. abs/2106.02134 (online), available from ⟨https://arxiv.org/abs/2106.02134⟩ (2021).

[2] Blevins, T., Gonen, H. and Zettlemoyer, L.: Analyzing the Mono- and Cross-Lingual Pretraining Dynamics of Multilingual Language Models (2022).

[3] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I. and Amodei, D.: Language Models are Few-Shot Learners, *Proceedings of NeurIPS* (Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.-F. and Lin, H.-T., eds.) (2020).

[4] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L. and Stoyanov, V.: Unsupervised Cross-lingual Representation Learning at Scale, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, Association for Computational Linguistics, pp. 8440–8451 (online), DOI: 10.18653/v1/2020.acl-main.747 (2020).

[5] CONNEAU, A. and Lample, G.: Cross-lingual Language Model Pretraining, *Advances in Neural Information Processing Systems* (Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. and Garnett, R., eds.), Curran Associates, Inc.

[6] Deshpande, A., Talukdar, P. and Narasimhan, K.: When is BERT Multilingual? Isolating Crucial Ingredients for Cross-lingual Transfer, pp. 3610–3623 (2022).

[7] 山下郁海，金子正弘，三田雅人，勝又 智，Imankulova, A.，小町 守：言語間での転移学習のための事前学習モデルと多言語の学習者データを用いた文法誤り訂正，自然言語処理， Vol. 29, No. 2, pp. 314–343（オンライン），DOI: 10.5715/jnlp.29.314 (2022).

[8] Kirov, C. and Cotterell, R.: 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).

[9] Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A. and Herbst, E.: Moses: Open Source Toolkit for Statistical Machine Translation, *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, Prague, Czech Republic, Association for Computational Linguistics, pp. 177–180 (online), available from ⟨https://aclanthology.org/P07-2045⟩ (2007).

[10] Lau, J. H., Armendariz, C., Lappin, S., Purver, M. and Shu, C.: How Furiously Can Colorless Green Ideas Sleep? Sentence Acceptability in Context, *Transactions of the Association for Computational Linguistics*, Vol. 8, pp. 296–310 (2020).

[11] Linzen, T., Dupoux, E. and Goldberg, Y.: Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies, *CoRR*, Vol. abs/1611.01368 (online), available from ⟨http://arxiv.org/abs/1611.01368⟩ (2016).

[12] Manning, C. D.: Last Words: Computational Linguistics and Deep Learning, *Comput. Linguist.*, Vol. 41, No. 4, pp. 701–707 (2015).

[13] McCoy, R. T., Frank, R. and Linzen, T.: Revisiting the poverty of the stimulus: hierarchical generalization without a hierarchical bias in recurrent neural networks, *cogsci.mindmodeling.org*.

[14] Papadimitriou, I. and Jurafsky, D.: Learning Music Helps You Read: Using Transfer to Study Linguistic Structure in Language Models, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online, Association for Computational Linguistics, pp. 6829–6839 (2020).

[15] Pinker, S. and Prince, A.: On language and connectionism: analysis of a parallel distributed processing model of language acquisition, *Cognition*, Vol. 28, No. 1-2, pp. 73–193 (1988).

[16] Ri, R. and Tsuruoka, Y.: Pretraining with Artificial Language: Studying Transferable Knowledge in Language Models, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Dublin, Ireland, Association for Computational Linguistics, pp. 7302–7315 (online), DOI: 10.18653/v1/2022.acl-long.504 (2022).

[17] Rumelhart, D. E. and McClelland, J. L.: On learning the past tenses of English verbs, Technical report, CALIFORNIA UNIV SAN DIEGO LA JOLLA INST FOR COGNITIVE SCIENCE (1985).

[18] Settles, B., Brust, C., Gustafson, E., Hagiwara, M. and Madnani, N.: Second Language Acquisition Modeling, *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, New Orleans, Louisiana, Association for Computational

Linguistics, pp. 56–65 (online), DOI: 10.18653/v1/W18-0506 (2018).
[19] Shi, F., Suzgun, M., Freitag, M., Wang, X., Srivats, S., Vosoughi, S., Chung, H. W., Tay, Y., Ruder, S., Zhou, D., Das, D. and Wei, J.: Language Models are Multilingual Chain-of-Thought Reasoners (2022).
[20] Warstadt, A. and Bowman, S. R.: Can neural networks acquire a structural bias from raw linguistic data?, *CoRR*, Vol. abs/2007.06761 (online), available from ⟨https://arxiv.org/abs/2007.06761⟩ (2020).
[21] Warstadt, A., Parrish, A., Liu, H., Mohananey, A., Peng, W., Wang, S.-F. and Bowman, S. R.: BLiMP: The Benchmark of Linguistic Minimal Pairs for English, *Transactions of the Association for Computational Linguistics*, Vol. 8, pp. 377–392 (online), DOI: 10.1162/tacl_a_00321 (2020).
[22] Wenzek, G.: CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Marseille, France, European Language Resources Association, pp. 4003–4012 (online), available from ⟨https://aclanthology.org/2020.lrec-1.494⟩ (2020).

付　　録

A.1 実験設定

第一言語獲得: トークナイザとして日本語は kytea[*5]を, フランス語とドイツ語とロシア語は mosesdecoder [9] を用いた．その後，fastBPE[*6]でサブワード分割を行った．語彙数はどの言語も 14,000 を設定した．ハイパーパラメータは 表 A·1 に記載している．

第二言語獲得: 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については, 英語は mosesdecoder [9] を用いており, 他の言語は事前学習と同じ設定である．

表 A·1: 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
	warmup_updates=30000
	beta1=0.9,beta2=0.999
	weight_decay=0.01
	eps=0.000001
epoch	100
n_heads	8
n_layers	12
clip_grad_norm	1.0
amp	2
fp16	True

A.2 実験結果の詳細

3.1 節の実験について, 文法項目ごとの習得過程を図 A·1 に示す.

[*5] http://www.phontron.com/kytea/
[*6] https://github.com/glample/fastBPE

Vol.2022-NL-254 No.22
Vol.2022-SLP-144 No.22
2022/12/1

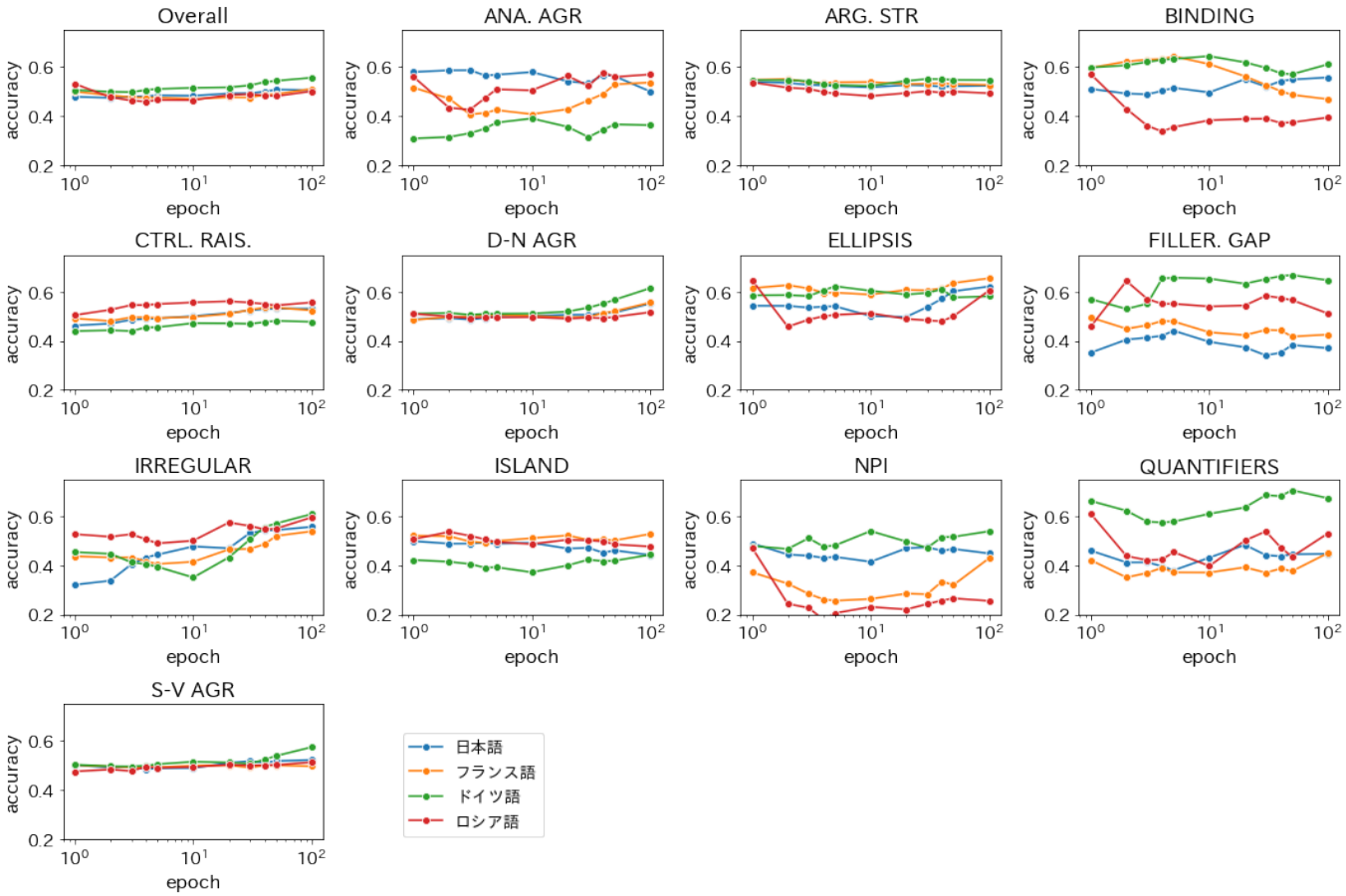

図 A·1: 第二言語学習中の各エポックにおける文法能力評価