深層学習モデルにおける言語特徴分布に関する研究

平野 颯　上垣外 英剛　渡辺 太郎
奈良先端科学技術大学院大学
{hirano.hayate.hc2,kamigaito.h,taro}@is.naist.jp

概要

Transformer ベースの言語モデルは、自然言語処理の幅広いタスクで活用されるようになった。モデルが行った判断の解釈のため、複数の分析方法が提案されてきた。以前の研究により、複数言語のコーパスで学習した言語モデルにおいて、言語の系統的な分布がモデル内部で学習されることが知られているが、事前学習の目的関数にそのような制約は含まれていないため特筆すべきである。本研究では、そのような分布が得られることが、言語特徴を学習できているためであると結論づけられるかどうかを、モデル表現の分布と比較することで分析する。

1　はじめに

Transformer [1] は、機械翻訳を目的に開発された Encoder-Decoder モデルだが、その応用範囲は広がりつづけている。これは、学習時に入力系列を並列に学習でき、また以前の手法と比較して高いタスク解決能力をもつためである。入力系列を並列に学習できる特性は、より多くのデータをモデルに学習させることを可能にした。

機械翻訳では Transformer の登場以前より、複数言語の翻訳を単一のモデルで学習する方法が成功を収めていた [2] が、Transformer 登場以後は、さまざまなタスクを複数の言語で解く試みがより活発に行われるようになった。そのような複数言語で学習されたモデルの利点のひとつは、リソースが比較的取得しづらい言語のデータで、単一の言語でのみ学習されたモデルよりもよい性能を発揮することである。これは、複数言語で学習されたモデルは、リソースの多い言語の学習で得た知識をリソースの少ない言語に対して用いることができるためと転移学習の文脈から説明されてきた。mBERT [3] は、多言語で学習された事前学習済み言語モデルだが、言語を明示するラベルは与えられておらず、機械翻訳を解くときのような言語どうしの変換も学習しない。このような場合でも、タスクを解く際に言語間での転移の効果が見られることは特筆すべきである。この能力について、入力する言語の特性の影響としては言語同士の近さと言語間の語彙の重複に焦点が当てられてきた。語彙の重複度が高いことが、モデルの言語間転移能力と相関するとした結果 [4] が報告されている。しかしながら、表記体系が異なる言語間での知識の転移 [5] も報告されており、語彙の重複のみから議論することはできない。

複数言語におけるマルチタスク学習は、語彙をはじめとした表面的な違いがあるにも拘らず、言語どうしには学習に利用できる共通した性質があることを仮定している。そのような言語がもつ普遍的性質の探求は言語学で進められており、複数の言語で定義可能な言語特徴が、人手により経験的に整理されてきた [6]。複数の言語を同時に学習した深層学習モデルにおいて、データから学習された言語共通の知識は、前述の言語同士の近さや語彙の重複の他にもこの言語特徴からも分析することが可能であると考えられる。

本研究は、複数の言語で学習された言語モデルがどのような言語特徴を獲得しているかを調査する。具体的にはまず、事前学習済みマスク言語モデルである mBERT [3] および、機械翻訳を学習した条件付き言語モデル m2m100 [7] から文表現を獲得する。獲得した文表現から、それぞれの言語特徴に対応する言語特徴表現を導出し、分析手法 SVCCA [8] を用いて言語特徴間およびモデル間の分析を行った。また、導出した言語特徴表現が空間上でどのように配置されているかを低次元空間上に射影することで可視化し、言語特徴ごとにどのような違いがあるかを確認した。実験結果は、mBERT と m2m100 ともに語順を示す言語特徴については文表現の空間上での配置と関連があり、先行研究の結果を支持した。一方で、動詞範疇などそれ以外の性質についての言語特徴に対しても文表現との関連があることがわかった。

2 言語特徴

本研究の分析に用いる言語特徴は、言語横断的なデータベース WALS [9] から獲得する。WALS は、各言語の文法記述など世界中の言語に対するデータをもとに、言語および言語特徴の地理的分布をまとめたものである。言語学の一分野である言語類型論は、このような記述された言語のデータから言語を分類することで、言語の多様性および共有する性質を研究する。たとえば、以下の 2 文のように

$$\underbrace{\text{桃太郎は}}_{\text{主語}}\ \underbrace{\text{鬼を}}_{\text{目的語}}\ \underbrace{\text{退治しました}}_{\text{述語}}。$$

$$\underbrace{\text{Momotaro}}_{\text{主語}}\ \underbrace{\text{beats}}_{\text{述語}}\ \underbrace{\text{the demons}}_{\text{目的語}}.$$

日本語と英語では語順、ここでは主語・述語・目的語が並べられる順序が異なることがわかる。言語類型論では、通言語的に比較できるものを言語特徴として扱い、WALS には 144 種類の言語特徴が収録されている。各言語特徴には、それが取りうる実現値の一覧とそれぞれの言語がどの値を取るかが設定されており、主語・述語・目的語の順序であれば 7 種類[1)]の値をもち、日本語は SOV、英語は SVO という実現値が設定されている。

3 言語特徴単位での表現分析

3.1 言語モデルからの文表現の獲得

言語モデルから特定の入力文表現 $\boldsymbol{s}$ を獲得する方法を説明する。n 単語からなる文 $\{x_1, x_2, \ldots, x_n\}$ が与えられたとき、これをモデルに入力して各単語位置に対する層 L の中間出力 $\{\boldsymbol{h}_1^L, \boldsymbol{h}_2^L, \ldots, \boldsymbol{h}_n^L\}$ を得る。このとき、層 L における入力文表現 $\boldsymbol{s}^L$ を位置 i に対するトークンの分散表現 $\boldsymbol{h}_i$ を系列内で平均して計算する。ここで、下流タスクに用いるために挿入されている記号位置に対応する中間出力は平均の計算に含めない。分析に用いるモデルにおいては、mBERT における [CLS] および [SEP] トークン位置、および m2m100 における言語タグ < xx > 位置の出力を無視する。

1) 主語・述語・目的語の並び順 6 種類と、基本となる語順が定まらないことを意味する”No dominant Order”。後者は主節では主語-述語-目的語の順だが、従属節では主語-目的語-述語を原則とするドイツ語のような、基本の語順が明確でない場合があるために用意されている。

3.2 言語特徴分析のための表現獲得

いま分析の対象とする言語特徴 F が n 種類の実現値 $\{F_1, F_2, \ldots, F_n\}$ をもつとする。ある言語 L に対する m 文の組 $\{\boldsymbol{x}_1^L, \boldsymbol{x}_2^L, \ldots, \boldsymbol{x}_m^L\}$ に対して、文表現の組 $\{\boldsymbol{s}_1^L, \boldsymbol{s}_2^L, \ldots, \boldsymbol{s}_m^L\}$ を得る。分析の対象とする言語のうち、同じ言語特徴 F の実現値 F_r をもつ言語内で同一の意味をもつ文表現を平均する。これを実現値 F_r における文 $\boldsymbol{s}_i$ の言語特徴表現 $f_{s_i}^{F_r}$ とする。

$$f_{s_i}^{F_r} = \mathrm{mean}(\{\boldsymbol{s}_i^l\}_{l \in L_{F_r}}) \tag{1}$$

言語特徴は品詞のような他の言語情報と異なり、文ごとではなく言語ごとに定義される。そのため、分析の対象とする m 文すべての $f_{s_i}^{F_r}$ の組 $\boldsymbol{f}^{F_r} = \{f_{s_1}^{F_r}, f_{s_2}^{F_r}, \ldots, f_{s_m}^{F_r}\}$ が分析の対象となる言語特徴表現である。

3.3 SVCCA による中間表現の比較

本研究では、モデル間の表現分析に SVCCA [8] を用いる。これは、深層学習モデルで扱われるベクトル表現どうしの比較を高速に、またアフィン変換に不変な形で行うことのできる手法として提案された。具体的には、n 個の入力データに対して、対象とする深層学習モデルのある中間層が出力する中間出力の組を $l \in \mathbb{R}^{n \times d}$ とする。比較したい 2 つの層について計算した中間出力を $l_1 \in \mathbb{R}^{n \times d_1}, l_2 \in \mathbb{R}^{n \times d_2}$, とする。ここで、対象とする層は異なるモデルに属していてもよい。l_1 および l_2 を入力に受け取り、SVCCA は以下の手続きを行う。

1. l_1 および l_2 に対して特異値分解を行い、$l_1' \subset l_1$, $l_2' \subset l_2$ を得る。本研究では、元のデータの分散が 99% 保持されるように軸の数を決定する。
2. l_1' および l_2' に対して正準相関分析 [10] を行う。それぞれに線形変換を施し $\tilde{l}_1' = W_1 l_1'$, $\tilde{l}_2' = W_2 l_2'$ を得る。この線形変換後の空間を張る各データの相関係数が最大になるようにする。

本研究では Raghu ら [8] にならい、相関係数の平均で分析を行う。正準相関分析は同じデータに対して複数種類の観測結果が得られる際に、共通して含まれる情報を抽出する分析手法である。そのため本研究では、Kudugunta ら [11] と同様に複数言語で同一の意味をもつ文を収録する多言語平行コーパスを利用する。

表 1 言語特徴中の任意の 2 つの実現値間で平均した SVCCA での射影先におけるデータの相関係数。各言語特徴の分類に対応する言語特徴のリストは付録 A に示した。

	音韻	形態	名詞範疇	名詞修飾・構文	動詞範疇	語順	一致	複文	語彙
mBERT	74.2	73.3	76.2	78.3	78.8	76.5	76.1	71.8	81.6
m2m100_418M	79.4	80.0	80.5	80.3	83.3	80.6	80.4	76.1	85.4
m2m100_1.2B	76.4	74.9	75.5	78.9	80.3	77.7	78.4	71.7	84.1

4 実験

4.1 実験設定

本実験では、Transformers [12] 上に実装された mBERT[2] および、m2m100[3] を用いる。m2m100 はパラメータ数の異なるモデルが提供されており、実験では 418M パラメータ (m2m100_418M) と 1.2B パラメータのモデル (m2m100_1.2B) を用いる。各モデルからの表現の獲得のために、Wikipedia から抽出された文を人手で 101 言語に翻訳し構築された FLORES-101 [13] コーパスを利用する。これは、本実験で利用するモデル mBERT および m2m100 の学習に用いられておらず、また対象とする任意の 2 つの言語に対して意味的に等価な文を用意することができるためである。本研究の目的は、言語がもつ機能である言語特徴に関する分析を行うことであり、言語間で意味が等価である平行コーパスを用いる必要がある。モデルおよびコーパスに共通の対象言語は 64 言語であり、評価用サブセット devtest から言語ごとに 1012 文を取得する。モデル表現は、m2m100 はエンコーダから獲得する。WALS に収録されている 144 の言語特徴のうち、対象の 64 言語のうち 1 つの言語のみが属する実現値をもつものは分析対象から除外し、75 の言語特徴を分析に用いる。

4.2 中間表現に反映される言語特徴の違い

分析の対象とした 75 種類の言語特徴を、音韻や形態をはじめとした、WALS 上で示されているグループに分類する。各グループはそれぞれが語や文のもつ機能に対応している。この分類のもとで、各言語特徴の任意の 2 つの異なる実現値どうしの関連性を SVCCA により分析することで、同種の言語特徴の共通性をモデルが捉えられているかを判断することができる。特異値分解の結果、mBERT・m2m100_418M・m2m100_1.2B それぞれのモデルから獲得した言語特徴表現は、平均で 169 次元・218 次元・114 次元に次元圧縮された。表 1 に各モデルにおける分析結果を示す。最も相関の低い m2m100_1.2B の複文の言語特徴においても 71.7 と、いずれも高い相関係数を示している。これは平行コーパスである FLORES-101 から計算した文表現をもとにしていることから、文表現に捉えられている文どうしの意味の同一性も影響していると考えられる。言語特徴ごとに見ると、名詞・動詞範疇、語順が他のグループの言語特徴と比べ高い数値を獲得している。とくに実現値どうしの相関が低いものは複文に関する言語特徴であり、構文解析などの言語処理タスクでもこれまで解決が難しかった文の性質と対応する。またモデルの違いについては、m2m100 の間での比較を行うとパラメータ数の大きい m2m_1.2B が相関係数が低くなっている。この点は、モデルのサイズが及ぼす影響についてのさらなる分析が必要である。また、特定の語彙の情報はすべての文に含まれるわけではなく弱い情報であるのに対して、語彙の言語特徴は高い相関を示した。これについては、低次元空間での可視化を用いて考察する。

4.3 表現空間での言語特徴の分布

図 1 および図 2 は、mBERT 言語特徴表現を t-SNE [14] で 2 次元空間上に射影し、各言語特徴の実現値ごとに色分けをしたものである。いずれの図中においても、各点はある文・実現値における言語特徴表現である。モデルがもつ層の数は異なるものの、m2m100 においても可視化結果は同様の傾向を示し結論は変わらないため、実験結果は mBERT によるもののみを示す。図 1 は言語特徴 81A、主語・述語・目的語の順序について、図 2 は言語特徴 138A、茶を意味する単語の歴史的な分類についての可視化結果である。いずれも下位の層からの出力結果から計算した表現は、言語特徴の実現値が異なる場合同じ文を表していても空間上では異なった位置に配置されている。しかし、より上位の層からの

2) https://huggingface.co/bert-base-multilingual-cased
3) https://huggingface.co/docs/transformers/model_doc/m2m_100

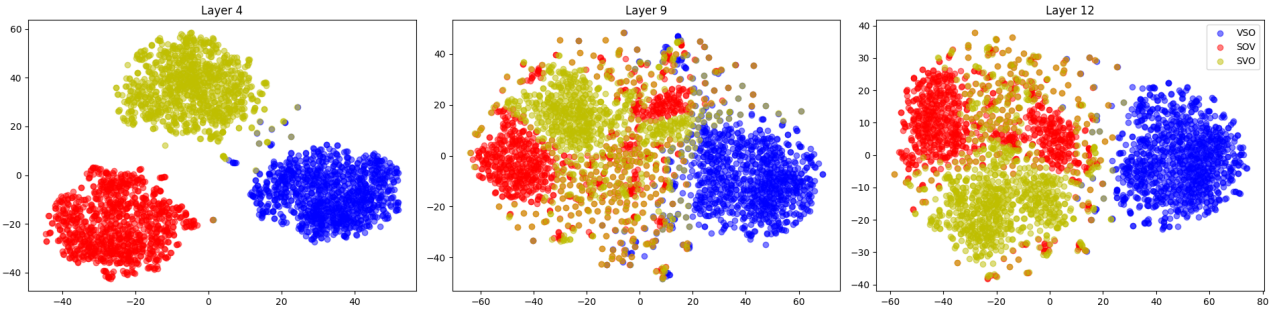

図 1　81A 特徴の特徴空間での分布

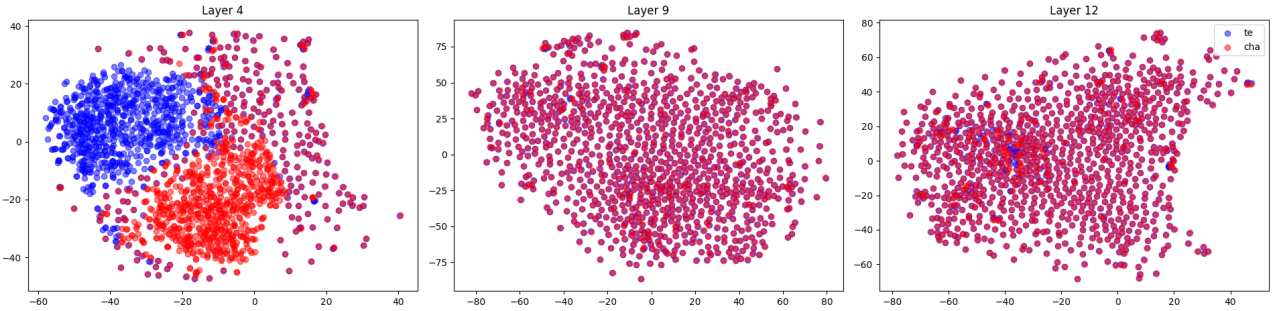

図 2　138A 特徴の特徴空間での分布

中間表現から獲得した言語特徴表現では、言語特徴 138A の 2 つの異なる実現値に対する表現が分離されていない。138A のような語彙の言語特徴は、実現値に関係なく空間上で非常に近い位置に配置されているため、SVCCA によるデータの相関の計算結果で高い数値を獲得したと考えられる。一方で言語特徴 81A は、上位の層においても異なる実現値の言語特徴表現が分離されつつ、表 1 に示すように同じ言語特徴どうしの関連性も捉えられている。

5　関連研究

Probing [15, 16] は、品詞タグなどの特定の言語処理タスクのラベル分類を行うための分類器を、分析の対象とする学習済みの深層学習モデルの内部表現から構成し、分類がうまくできるほど分析対象の言語知識をモデルが獲得しているとする分析方法である。ここで、学習する分類器に何を選択すべきかは自明ではない [17]。また、言語情報へのクラスタ割り当てを行い分類する、分類器を新たに学習しない分析方法 [18] が提案されている。本研究は、言語固有の特性を分析に用いる点で、Probing と類似している。しかしながら、本研究で対象とする言語特徴は言語ごとに定義され、Probing における分析に用いられるものとは異なる。

複数言語で学習された機械翻訳モデルに対して、平行コーパスを用いることで SVCCA を異なる言語の文の組に適応可能にした研究 [11] は、異なる言語の Encoder 表現は言語の類似性を捉えることを示した。本研究の分析の用いる言語特徴は、言語自体の機能に基づく分析が可能であり、その点で同じく分析に言語特徴を用いた [19] と同様の分析を行った。しかしながら [19] は、言語特徴に基づく Probing 分類器を学習し分析を行っており、分類器の選択が適当かどうかの判断が困難である [17] 課題が残る点で本研究の分析手法とは異なる。

6　おわりに

本研究は、言語横断的に学習された Transformer ベースの言語モデルが、言語の系統的な分布を獲得することに関連して、同様に言語がもつ機能的な特徴についても獲得できているかを分析した。言語特徴に注目することで、言語の系統関係に基づいた類似性を分析に用いる場合では行えない、言語がもつ機能に基づく分析を可能にした。本研究の分析方法および結果は、事前学習済み言語モデルがデータのみから、明示的な制約なしでどれほど言語がもつ要素を学習できるかを考えることに役に立つ。これにより今後のモデル開発・応用タスクへの精度向上に必要な要素の特定に活用できる。

参考文献

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In **Advances in Neural Information Processing Systems**, Vol. 30, 2017.

[2] Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's multilingual neural machine translation system: Enabling zero-shot translation. **Transactions of the Association for Computational Linguistics**, Vol. 5, pp. 339–351, 2017.

[3] 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, 2019.

[4] Shijie Wu and Mark Dredze. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**, pp. 833–844, 2019.

[5] Wietse de Vries, Martijn Wieling, and Malvina Nissim. Make the best of cross-lingual transfer: Evidence from POS tagging with over 100 languages. In **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 7676–7685, 2022.

[6] William Croft. **Typology and universals**. Cambridge University Press, 2002.

[7] Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Michael Auli, and Armand Joulin. Beyond english-centric multilingual machine translation. **Journal of Machine Learning Research**, Vol. 22, No. 107, pp. 1–48, 2021.

[8] Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In **Advances in Neural Information Processing Systems**, Vol. 30, 2017.

[9] Matthew S. Dryer and Martin Haspelmath, editors. **WALS Online (v2020.3)**. Zenodo, 2013.

[10] David R. Hardoon, Sandor Szedmak, and John Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. **Neural Computation**, Vol. 16, No. 12, pp. 2639–2664, 2004.

[11] Sneha Kudugunta, Ankur Bapna, Isaac Caswell, and Orhan Firat. Investigating multilingual NMT representations at scale. In **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**, pp. 1565–1575, 2019.

[12] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations**, pp. 38–45, 2020.

[13] Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. **Transactions of the Association for Computational Linguistics**, Vol. 10, pp. 522–538, 2022.

[14] Geoffrey E Hinton and Sam Roweis. Stochastic neighbor embedding. In S. Becker, S. Thrun, and K. Obermayer, editors, **Advances in Neural Information Processing Systems**, Vol. 15, 2002.

[15] Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 2126–2136, 2018.

[16] Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Sam Bowman, Dipanjan Das, and Ellie Pavlick. What do you learn from context? probing for sentence structure in contextualized word representations. In **International Conference on Learning Representations**, 2019.

[17] John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**, pp. 2733–2743, 2019.

[18] Yichu Zhou and Vivek Srikumar. DirectProbe: Studying representations without classifiers. In **Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 5070–5083, 2021.

[19] Rochelle Choenni and Ekaterina Shutova. Investigating language relationships in multilingual sentence encoders through the lens of linguistic typology. **Computational Linguistics**, Vol. 48, No. 3, pp. 635–672, 2022.

A　分析に用いた WALS 言語特徴

表 2 に分析に用いた WALS 言語特徴 75 種類の番号と、各言語特徴がどのような言語情報のグループに属するかの情報を示す。

表 2　言語特徴が属するグループと、対応する言語特徴

グループ名	WALS での言語特徴番号
音韻	1A 2A 4A 9A 10A 12A 13A 16A
形態	21A 24A 27A 28A 29A
名詞範疇	32A 34A 36A 37A 38A 39A 41A 42A 43A 44A 45A 46A 47A 48A 52A 53A 54A 55A 56A 57A
名詞修飾・構文	63A 64A
動詞範疇	65A 66A 67A 68A 69A 71A 73A 74A 75A 76A 77A 79A
語順	81A 82A 83A 86A 87A 89A 90A 92A 94A 95A
一致	100A 103A 107A 110A 111A 113A 115A 116A 118A 119A 120A 121A
複文	125A 126A 127A
語彙	129A 136A 138A