膠着型言語における複雑さのトレードオフ

膠着型言語における複雑さのトレードオフ

李　文超（浙江大学）†

Complexity trade-off in agglutinative languages

Wenchao Li (Zhejiang University)

要旨

本研究は 9 つの言語族にわたる 16 の膠着言語の 25 のツリーバンクを利用し、形態的豊かさと語順の柔軟性との相関関係を探った。形態的豊かさはパラダイムの移動平均形態的豊富さと移動平均サイズで測った。語順の柔軟性はコサイン類似度とエントロピーで計測した。統計分析の結果、形態素が豊富であればあるほど、語順がより柔軟になる。形態と語順の間にはかなり強い相関関係があることが確認され、複雑さのトレードオフ仮説を裏付けている。9 つの言語族の中で、オーストロネシア語、アフロアジア語、ドラヴィダ語は、アルタイ語、インド・アーリア語、ウラル語に比べて形態的にも語順的にも多様性が低いと判明した。トルコ・ウイグル・バスクおよび北サーミ語は、S、V、O の組み合わせの割合が最もバランスが取れていることが観察された。同じ言語族内に形態・語順の分離が窺えた: ウラル語族はフィン語派がフィン・ウゴル語派より柔軟であり ; アルタイ語族はモンゴル語派がチュルク語派より厳格である。形態—統語の一端から言語間距離が如何に外国語習得に左右するかを考察して結果、目標言語の語順自由度の高い言語の場合、母語が形態的に豊富であれば産出した目標言語の語順が自由になると結びついた。

１．はじめに

人間の言語は、複雑かつ動的で、階層的に組織されている規則的なシステムである (Fenk-Oczlon and Pilz 2021)。 その規則性の 1 つは complex trade off である。言語内の 1 つの変数 (音韻論、形態論、構文、意味論など) が洗練されている場合、別の変数が単純化される傾向がある。言語がほぼ同じ程度の複雑さを表現できるように保つ(Menzerath’s law 1954; Shosted 2006; Sinnemäki 2014; Fenk-Oczlon and Fenk 2014)。複雑さのトレードオフ仮説は Sapir (1921) 以来さまざまな言語変数間で確認されていた: 音素-音節-単語 (Coloma 2017)、音韻-形態論 (Shosted 2006)、形態-構文 (Jakobson 1936; McFadden 2003; Koplenig 2017; Yan と Liu 2021、Li、Liu、Xiong 2022)。トレードオフの考え方における重要な問題点は、トレードオフがどの程度成立するか、どの言語変数が多数の変数の中、重要な役割を果たすかという点にある。最初の問題点に関して、ほとんどの研究は、トレードオフは一部の参加者によるものであると考えられている。例えば、単語あたりの音節が多いほど、音節あたりの音素は少なくなり; 音節あたりの音素が多いほど、形態的ケースは少なくなり; 形態的ケースが増えるほど、語順の自由度が高まる(Fenk-Oczlon and Fenk 1999; Shosted 2006; Sinnemäki 2008, 2014; Miestamo 2009; Koplenig et al. 2017)。2 番目の問題点に関しては、Fenk-Oczlon・Pilz (2021) は多言語の翻訳の並列データを調査した結果、音素サイズ、音節サイズ、単語の長さ、文節の長さと人口の間の相関関係が確認できた。多変数の中、音節の複雑さが最も重要であると主張した。

† widelia@zju.edu.cn

本研究はトークン化、見出し語化、POS および形態的特徴のタグ付け、係り受け解析を担う自然言語処理のツールキットを使い、自然言語の形態的多様性と語順の柔軟性との相互作用を洞察する。バスク語、北サーミ語、エストニア語、フィンランド語、ハンガリー語、日本語、マラーティー語、タミル語、テルグ語、トルコ語、ウイグル語、カザフ語、ブリヤート語、ウォロフ語、インドネシア語、コプト語の 16 の膠着語を対象とし、ユニバーサル依存関係からの 25 の注釈付きコーパスを分析する。形態的豊かさは、移動平均形態的豊かさ (MAMR: Čech and Kubát 2018) とパラダイムの移動平均サイズ (MAMSP: Xanthos and Gillis 2010) の 2 つの指標によって測定される。語順の柔軟性の把握にコサイン類似度 (COS) とエントロピー (ENTR: Shannon 1948; Chen et al. 2016、Bentz et al. 2017; Yan and Liu 2021; Li et al. 2022) が使用される。スピアマンの順位相関係数は変数の間の関連性を測定する。

2. 研究方法

2.1 データセット

以下のツリーバンクが採用される: a). 4 つのウラル言語: エストニア語 (2 つのツリーバンク)、フィンランド語 (2 つのツリーバンク)、北サーミ語 (1 つのツリーバンク)、ハンガリー語 (1 つのツリーバンク); b). 4 つのアルタイ語: ブリヤート語 (1 つのツリーバンク)、カザフ語 (1 つのツリーバンク)、トルコ語 (2 つのツリーバンク)、ウイグル語 (1 つのツリーバンク); c). 2 つのドラヴィダ語: タミル語 (2 つのツリーバンク)、テルグ語 (1 つのツリーバンク); d). インド・アーリア語のマラーティー語 (1 つのツリーバンク); e). アフロアジア系エジプト語コプト語 (1 つのツリーバンク); f). ニジェール・コンゴ大西洋言語ウォロフ語 (1 つのツリーバンク); g). 1 つのバスク語 (1 つのツリーバンク); h). オーストロネシア語 1 つ: インドネシア語 (3 つのツリーバンク); 日本語(三つのツリーバンク)。表 1 に、ツリーバンクの詳細を示します。

本研究が扱うツリーバンクはテキストサイズとテキストタイプが異なっている（ブログ、電子メール、フィクション、文法例、法律、ニュース、レビュー、ソーシャル、ウェブ、ウィキ）。そこで、コーパスのサイズが結果に影響を与えるかどうかという疑問が生じた。コーパスサイズの影響を最小限に抑えるため、以下の作業を行った。コーパスの中で最小のコーパスはマラーティー語、3847 単語である故、コーパスを 4 つのサブコーパス: 3000 トークン、9000 トークン、12000 トークン、および完全トークンにサブセットする。次に、各サブコーパスの 5 つのメトリクス値を計算し、平均値を使う。さらに、本稿は主文と、S、O、V からなる副文の両方を調査し、平均値を取ることにする。

表 1. データセット

Treebanks	Text types	Words	Sentences	Treebanks	Text types	Words	Sentences
Basque-BDT	News	121,443	8993	**Marathi-UFAL**	Wiki, fiction	3,847	466
Buryat-BDT	Fiction, grammar-examples, news	10,185	927	**Indonesian-PUD**	News, wiki	19,446	1,000
Japanese-BCCWJ	Fiction, news, blog, conference, nonfiction	1,253,903	57,109	**Indonesian-GSD**	Blog, news	122,019	5,598
Japanese-GSDLUW	blog, news	150,243	8,100	**Wolof-WTB**	bible, wiki	44,258	2,107
Japanese-PUD	news, wiki	28,788	1,000	**Uyghur_UDT**	Fiction	40236	3,456
Tamil-TTB	News	9,581	600	**Indonesian-CSUI**	News, nonfiction	28,263	1,030
Tamil-MWTT	News	2,584	534	**Wolof-WTB**	Bible, wiki	44,258	2,107
Telugu-MTG	Grammar examples	6,465	1,328	**Coptic-Scriptorium**	Bible, fiction, nonfiction	55,858	2,163
Buryat-BDT	Grammar examples, news, fiction	10,185	927	**Estonian-EDT**	Fiction, academic, news, nonfiction	438,245	30,968
Kazakh-KTB	News, fiction, wiki	10,536	1,078	**Finnish-TDT**	Fiction, legal, news, blog, grammar-examples,	202,453	15,136
Turkish-Kenet	News, nonfiction	183,555	16,396	**Finnish-TDT**	Poetry, medical, social, web	19,382	2,122
Turkish-Boun	News, nonfiction	125,212	9,761	**North Sami-Giella**	News, nonfiction	26,845	3,122
Hungarian-Szeged	News	42,032	1,800				

2.2 指標

2.2.1 形態的豊かさ

コーパスサイズの影響を最小限に抑えるために、テキストのサブセットのタイプトークン比 (TTR) インデックスと平均を繰り返し計算する。各言語テキストの形態学的多様性を測定する。Covington と McFall (2010) の移動平均 TTR (MATTR) を使用して、語形と補題語彙の豊富さを計算する。MATTR $(W)_{\text{word form}}$は次の式で求められる。

$$\text{MATTR}\ (W)_{\text{word form}} = \frac{\sum_{I=1}^{N-W+1} F_i}{W\ (N - W + 1)}$$

N はテキストのサイズ、W はランダムに選択されたウィンドウのサイズである。F_i は、特定のウィンドウサイズの拡張形式の数値である。日本語やトルコ語などの膠着語には 2 種類の語形がある。一つは語根の膠着（さびしい→さびしかった）、もう一つは派生（さびしい→さびしみ）。屈折言語に関して、語形には屈折 (kind → kinder) と派生 (kind → kindness) がある。本研究のウィンドウサイズは 500 ワードにする。MATTR $(W)_{\text{word form}}$は共役によって得られる各ウィンドウの TTR である。各ウィンドウの TTR は、同様の式によって取得される。

$$\text{MATTR}\ (W)_{\text{word lemma}} = \frac{\sum_{I=1}^{N-W+1} L_i}{W\ (N - W + 1)}$$

N はテキストのサイズ、W はランダムに選択されたウィンドウのサイズである。L_iは、一定のウィンドウ サイズの拡張形式の数値である。MATTR $(W)_{\text{word lemma}}$ は共役による各ウィンドウの TTR である。MAMR は、語形の多様性と補題の多様性の違いを指す。次の方法で取得される。

$$\text{MAMR}\ (W) = \text{MATTR}\ (W)_{\text{word form}} - \text{MATTR}\ (W)_{\text{word lemma}}$$

MAMR (W) が高いほど、言語の形態が豊かになる。語彙の多様性を捉えるためのもう 1 つの尺度は、パラダイムの平均サイズ(MSP、Xanthos and Gillis 2010)である。次の操作で取得する。まず、さまざまな語形変化の数を補題の数で割る。

$$\text{MSP} = \frac{F}{L}$$

MAMSP を次のように取得する。

$$\text{MAMSP} = \frac{\sum_{I=1}^{N-W+1} \frac{F_i}{L_i}}{W\ (N-W+1)}$$

MAMP と MAMSP の値が高いほど、言語の形態が豊かになる。

2.2.2 語順の柔軟性

語順に関する言語構成要素は約 19 個がある(角田 2009 [1991])。本研究では、以下の理由によって、語順を主語 S、目的語 O、動詞 V の順に絞り込んだ: a). S、V、O 構成要素は自然言語の最も基本的な類型的特徴であり、大部分の文に存在している。したがって、言語間の比較に最適である; b). S、O、V 構成要素は元の注釈スキームに関係なく、すべてのツリーバンクで比較的簡単に識別できる(Kubon et al. 2016)。以上を照らし、考えられる語順パターンとしては、SVO、OVS、VSO、VOS、SOV、OSV の 6 つになる。語順の自由度が最も高い場合は、6 つのパターンが均等に分布する。即ち、各パターンの理想ベクトルは等しい: SVO:

0.166、OVS: 0.166、VSO: 0.166、VOS: 0.166、SOV: 0.166、OSV: 0.166 となる。t が期待値、s が観測値であると仮定すると、COS は次の式を使用して取得される:

$$\text{COS (s, t)} = \frac{\sum_{i=1}^{n} s_i t_i}{\sum_{i=1}^{n} s_i^2 \cdot \sum_{i=1}^{n} t_i^2}$$

COS (s, t) が高くなるほど、サンプルの語順の自由度が高くなる。ENTR は次の方法で取得される。

$$\text{ENTR} = -\sum_{i=1}^{n} t_i \times \ln t_i$$

特定の言語のエントロピーが高くなるほど、語順の自由度が高くなる。

3. 結果

3.1 指標の信頼性

言語変数がどのように相互作用するかという研究課題に取り組む前に、本研究の指標が複数の言語的特徴を示すのに有効であるかどうかを確認しておく。形態的豊かさから始まろう。ツリーバンクの MAMR、MAMSP 平均値を計算し、降順に表 2 にまとめた。

表 2. MAMR, MAMSP の値

Language family	Branch	Languages	Morphology	MAMR	MAMSP
Altaic	Turkic	Uyghur	Agglutinative	0.2418	1.4785
Altaic	Turkic	Kazakh	Agglutinative	0.1942	1.3341
Indo-European	Indo-Aryan	Marathi	Agglutinative	0.1678	1.4344
Uralic	Finnic	Finnish	Agglutinative	0.1673	1.2985
Altaic	Mongolic	Buryat	Agglutinative	0.1668	1.2700
Basque	Basque	Basque	Agglutinative	0.1569	1.2416
Uralic	Finnic	Estonian	Agglutinative	0.1406	1.2503
Indo-European	Iranian	Kurmanji	Agglutinative	0.1381	1.3164
Armenian	Armenian	Armenian	Agglutinative	0.1345	1.2518
Uralic	Finno-Ugric	North Sami	Agglutinative	0.1303	1.2280
Altaic	Turkic	Turkish	Agglutinative	0.1089	1.3600
Dravidian	Dravidian	Tamil	Agglutinative	0.1046	1.2474
Dravidian	Dravidian	Telugu	Agglutinative	0.1043	1.2466
Afroasiatic	Egyptian	Coptic	Agglutinative	0.1001	1.2001
Niger-Congo	Atlantic	Wolof	Agglutinative	0.0977	1.2545
Uralic	Finno-Ugric	Hungarian	Agglutinative	0.0707	1.1094
Austronesian	Malayo-Polynesian	Indonesian	Agglutinative	0.0509	1.0829
Japanese	Japanese	Japanese	Agglutinative	0.0253	1.0488

MAMR と MAMSP の値が高いほど、形態的複雑さがより豊かであることが示されている。形態的豊かさの最も高い値を示す上位 5 言語のうち、三つがアルタイ語族に属す: ウイグル語 (チュルク語派)、カザフ語 (チュルク語派)、ブリヤート語 (モンゴル語派)。そして、ウラル語語族のフィン語派（エストニア語とフィンランド語）は、フィン・ウゴル語派（北

サーミ語とハンガリー語）よりも形態が多様であった。アルタイ語族では、モンゴル語派がチュルク語派よりも厳格であった。インド・アーリア語族において、マラーティー語が唯一の膠着語であり、形態的豊かさは度合いが他のインド・アーリア語、すなわち融合型のウルドゥー語とヒンディー語より高い。インドネシア語と日本語は、より低い値でランク付けされている。日本語に形態的似ていると言われているドラヴィダ語族のタミル語は中程度の形態的多様性を見せた。

形態的豊かさを反映する MAMR と MAMSP の一貫性と信頼性をさらに調べるには、図 3 に回帰直線を使用した散布図をプロットした。

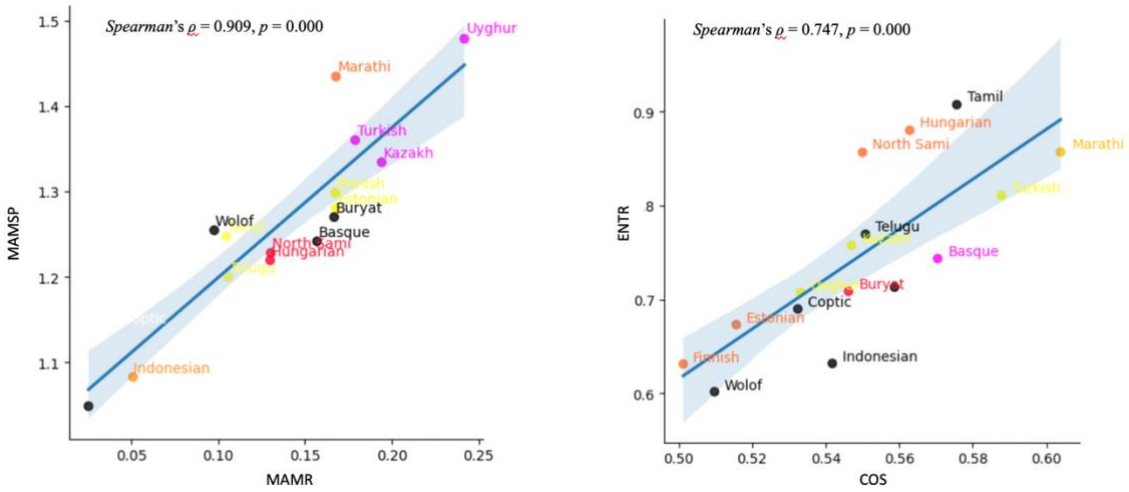

図 1. MAMR、MAMSP; COS、ENTR による散布図と回帰直線

図 1 が示すように、MAMR と MAMSP の関係は回帰直線に一致している。MAMR と MAMSP の間のスピアマンの順位相関係数は正で統計的に有意であった ($\rho = 0.909$ および $p = 0.000$)。これは、2 つの指標が形態的多様性を反映できることが窺えた。

次に、語順に移る。表 3 は、ENTR と COS (語順の柔軟性を表す指標) の平均値を降順に示している。COS と ENTR の値が大きいほど、語順の自由度が高くなる。

表 3. COS と ENTR の値

Language family	Branch	Languages	COS	ENTR
Indo-European	Iranian	Kurmanji	0.7987	1.3791
Armenian	Armenian	Armenian	0.7962	1.0033
Uralic	Finno-Ugric	North Sami	0.7943	1.0021
Uralic	Finnic	Finnish	0.7788	1.3202
Niger-Congo	Atlantic	Wolof	0.7680	1.4246
Afroasiatic	Egyptian	Coptic	0.7289	0.9546
Austronesian	Malayo-Polynesian	Indonesian	0.6728	1.1602
Basque	Basque	Basque	0.6136	1.0293
Altaic	Mongolic	Buryat	0.6070	0.9685
Uralic	Finno-Ugric	Hungarian	0.5989	1.0001
Dravidian	Dravidian	Tamil	0.5963	0.8732

Dravidian	Dravidian	Telugu	0.5507	0.7992
Altaic	Turkic	Kazakh	0.5241	0.7869
Indo-European	Indo-Aryan	Marathi	0.5105	0.7813
Altaic	Turkic	Turkish	0.5081	0.6888
Altaic	Turkic	Uyghur	0.5081	0.7179
Japanese	Japanese	Japanese	0.4877	0.5400
Uralic	Finnic	Estonian	0.4773	0.5821

柔軟性の上位 5 言語は、クルマンジー語、アルメニア語、北サーミ語、フィンランド語とウォロフ語である。ドラヴィダ語、バスク語、アフリカアジア語は中級レベルで見つかった。日本語とエストニア語は語順が最も厳格と観察された。図 2 と表 4 は、各言語の順序の割合を詳しく示している。

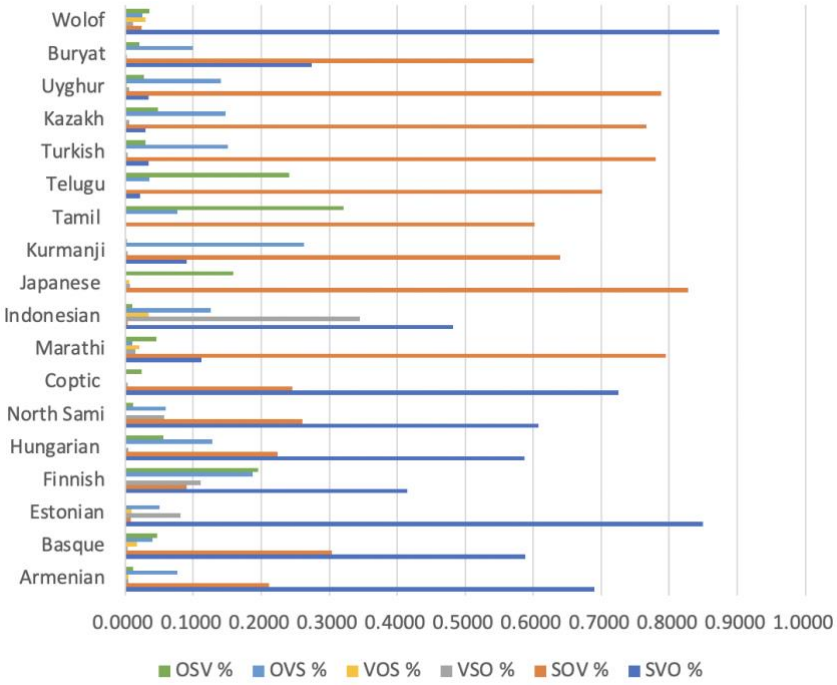

図 2. 各言語の順序の割合

表 4. 語順パターンの分布

Language family	Branch	Language	SVO %	SOV %	VSO %	VOS %	OVS %	OSV %
Indo-European	Armenian	Armenian	0.6895	0.2120	0.0044	0.0049	0.0770	0.0122
Basque	Basque	Basque	0.5884	0.3039	0.0034	0.0170	0.0397	0.0476
Uralic	Finnic	Estonian	0.8497	0.0088	0.0814	0.0091	0.0510	0.0000
Uralic	Finnic	Finnish	0.4147	0.0906	0.1116	0.0000	0.1877	0.1954
Uralic	Finno-Ugric	Hungarian	0.5869	0.2243	0.0047	0.0000	0.1280	0.0561
Uralic	Finno-Ugric	North Sami	0.6075	0.2607	0.0579	0.0020	0.0599	0.0119
Afroasiatic	Egyptian	Coptic	0.7258	0.2460	0.0040	0.0000	0.0000	0.0242
Indo-European	Indo-Aryan	Marathi	0.1123	0.7944	0.0156	0.0211	0.0111	0.0455

Austronesian	Malayo-Polynesian	Indonesian	0.4818	0.0032	0.3451	0.0343	0.1254	0.0102
Japanese	Japanese	Japanese	0.0000	0.8277	0.0076	0.0060	0.0000	0.1587
Indo-European	Iranian	Kurmanji	0.0910	0.6391	0.0033	0.0015	0.2630	0.0021
Dravidian	Dravidian	Tamil	0.0000	0.6020	0.0000	0.0000	0.0772	0.3208
Dravidian	Dravidian	Telugu	0.0220	0.7010	0.0000	0.0000	0.0360	0.2410
Altaic	Turkic	Turkish	0.0343	0.7798	0.0042	0.0010	0.1510	0.0297
Altaic	Turkic	Kazakh	0.0301	0.7661	0.0056	0.0020	0.1477	0.0485
Altaic	Turkic	Uyghur	0.0344	0.7886	0.0058	0.0021	0.1412	0.0279
Altaic	Mongolic	Buryat	0.2745	0.6009	0.0023	0.0017	0.1001	0.0205
Niger-Congo	Atlantic	Wolof	0.8734	0.0240	0.0115	0.0302	0.0255	0.0354

マラーティー語は語順に最も高い柔軟性を示した。イラン語派のクルマンジーは 2 番目の柔軟性を示し、SOV と OVS がかなりの割合が見せた。フィンランド語、ハンガリー語は、S、V、O の組み合わせの最もバランスのとれた比率を示した。ウラル族もすべての語順可能性を示したが、SVO では他の語順を大幅に上回っており、比率のバランスが悪く、値が低くなった。バスク語は本質的に凝集性の形態であり、バランスの取れた順序分布が見出されている。アルタイ語族はあらゆる順序の可能性を示すが、SOV 順序が好まれた。チュルク語派 (トルコ語、カザフ語、ウイグル語) の順序の好みは、SOV (78%) > OVS (15%) > SVO (3%) > OSV になっている。(4%) > VSO (1%) > VOS (0%)。モンゴル系ブリヤート語は、トルコ系と比較して、SVO の割合が著しく高く (27.45%)、OVS の割合が低いと判明した。アフロアジア語族エジプト語派のコプト語に語順の余地が大きい。ドラヴィダ語族のテルグ語とタミル語は、一見中程度に厳格で、OV が好まれていた (SOV>OSV>OVS)。図 1 は COS と ENTR の間の相関関係を視覚化したもので、回帰直線は COS と ENTR の間の関係によく適合している。COS と ENTR の間のスピアマンの順位相関係数は正で統計的に有意であった (ρ = 0.747 および p = 0.000)。これは、語順に関する 2 つの指標は信頼性があり、語順の自由度を把握するのに有効であると示した。

NLP ツールキット、ツリーバンク、計算および数学的分析に基づいて、本研究の指標は、形態的および構文的な多様性を捉える上で信頼できることが確認された。こうして、最初の研究上の疑問が解けた。

3.2 形態と構文の相関関係

前のセクションでは、形態的および構文的な複雑さを捉えるための 4 つの指標 (MAMR、MAMSP、COS、ENTR) の有効性を確認した。本セクションでは、2 番目の研究課題、即ち、形態的に豊かな言語は構文的により自由になる可能性が高いかどうかを探る。スピアマンの順位相関係数分析は、形態と語順の指標の間で実行し、正のかなり強い相関関係が確認されていた。MAMR 対 COS: $\rho = 0.735$ および $p = 0.001$。MAMR 対 ENTR: $\rho = 0.656$ および $p = 0.006$。MAMSP 対 COS: $\rho = 0.624$ および $p = 0.010$。MAMSP 対 ENTR: $\rho = 0.565$ および $p = 0.023$。4 つの指標の散布図行列を図 3 にプロットした。

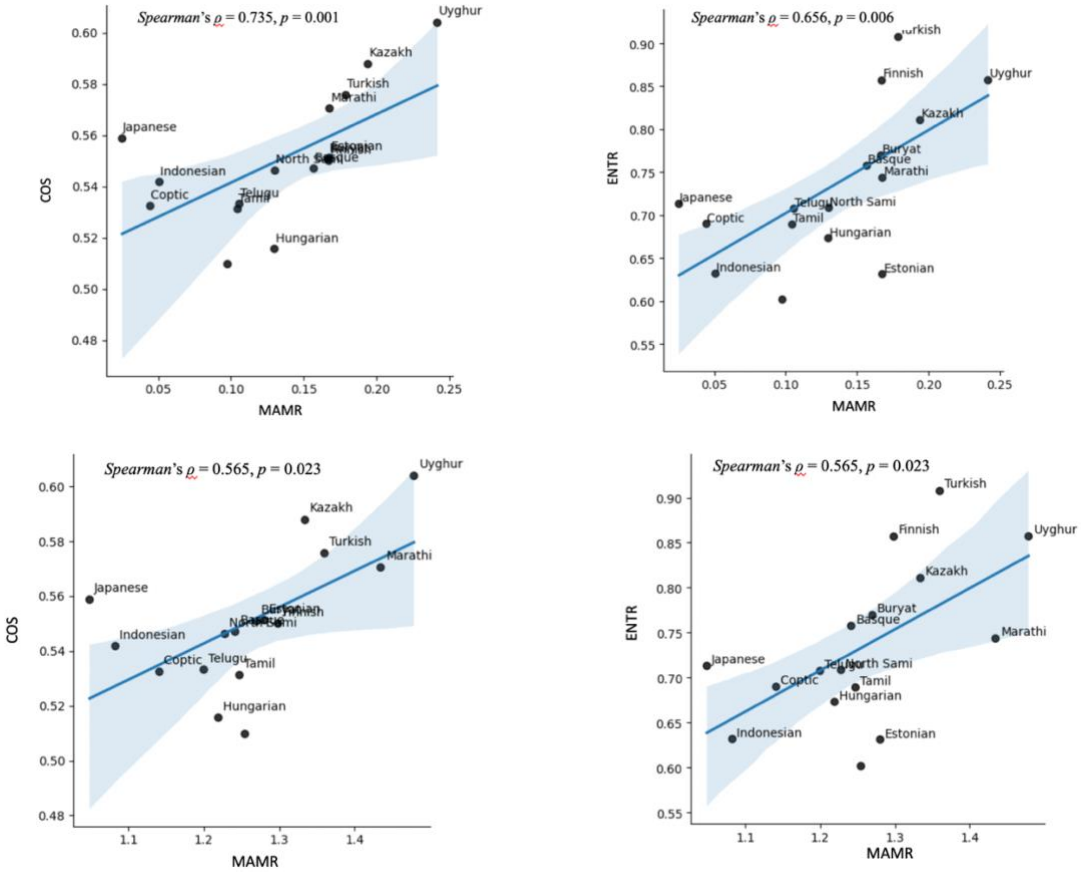

図 3. MAMR、MAMSP、COS、ENTR 間のスピアマンの順位相関係数

図 3 は、4 つの値間の正の相関関係を示している。言語の形態的豊かさが増加すると、語順の柔軟性が高まる。この発見は、スラブ言語 (Yan and Liu 2021) や日本語書き言葉 (李他 2022) と同様に、「複雑さのトレードオフ」仮説と一致している。図 4 と 5 は、形態的および統語的な多様性に基づいた 9 つの言語族の膠着言語のクラスタリングを示している。ウラル語族のフィン語派とフィン・ウゴル語派の分離が窺えた。モンゴル語派とチュルク語派の差が見られた：チュルク語派は、モンゴル語派よりも形態的に豊富で、語順に柔軟であった。日本語は他の 15 の膠着言語に比べて形態学的に豊かではないが、語順の自由度は中程度であると見出された。形態と語順に基づく言語のクラスタリングには、リトアニア語、クルマンジ語、アルメニア語、北サーミ語、フィンランド語、チェコ語、ギリシャ語とバスク語は形態的豊富さと語順の自由度に一致している。

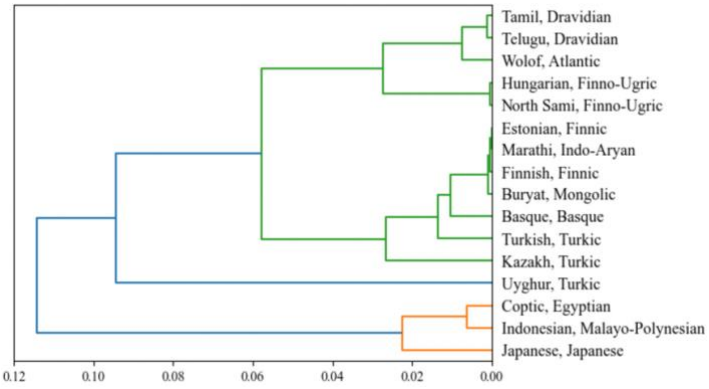

図 4. 形態的豊かさに基づく膠着言語のクラスタリング

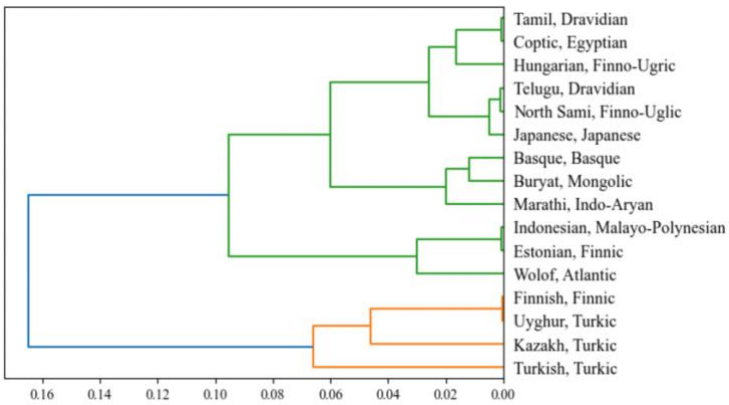

図 5. 語順の柔軟性に基づく膠着言語のクラスタリング

以上を踏まえ、形態—統語の一端から言語間距離が如何に日本語習得に左右するかを考察した。2022年10月時点で、在日外国人労働者数が182万人に達した(厚生労働省)。言語伝達は、L1、L2、Ln、学習環境、モチベーションなどの影響を伴うが(奥野 2019; 宇佐美 2006; 李· 石川· 砂川 2018；迫田· 細井 2020；阿辺川他 2020)、本稿は学習初期段階に絞る。12の母国語（フランス語、ドイツ語、ドイツ語、ドイツ語、ハンガリー語、中国語、インドネシア語、韓国語、ロシア語、スペイン語、タイ語、トルコ語、ベトナム語、英語（イギリス英語、ニュージーランド英語、アメリカ英語）にわたる日本語学習者の作文を3000分横断的に分析した。その結果、孤立型・屈折型言語話者は産出した語順は膠着型言語話者よりはるかに低い。

4. まとめ

本研究では、9 つの言語族にわたる 16 の膠着言語の 25 のツリーバンクにより、形態的豊かさと語順の柔軟性との相関関係を調査した。指標の MAMR、MAMSP、COS と ENTR は、形態的および構文的な多様性を捉えるのに有効であることが検証された。次の三点にたどり着いた。第一に、形態素が豊富であればあるほど、語順がより柔軟になる。形態と構文の間の相関関係がすべての膠着言語で確認されており、複雑さのトレードオフ仮説を裏付けている。9 つの言語族の中で、オーストロネシア語、アフリカアジア語は形態的にも語順的にも固定していることが観察された。形態と統語的に多様な言語は、アルタイ語族のチュル

ク語派、ヨーロッパ語族のインド・アーリア語派、ウラル語族のフィン語派と窺えた。トルコ語、ウイグル語、バスク語および北サーミ語は、S、V、O の組み合わせの割合が最もバランスが取れていた。第二に、言語族内に差が見られた : モンゴル語派はチュルク語派より、フィン語派はフィン・ウゴル語派より語順が柔軟である。第三に、L1 の形態学的豊かさと習得言語の語順の再帰性との間に正の関係が確認された。

謝　辞

本研究は中国国家社会科学基金に基づいて行われたものである ((22BYY186: 李文超)。

文　献

阿辺川武，仁科喜久子，八木豊，ホドシチェック･ボル (2020). 日本語接続表現の計量的分析に基づく指導法の提案『計量国語学 』32(7) 387-402.

Bentz C., Alikaniotis D., Cysouw M., Ferrer-i-Cancho R (2017). The entropy of words. Learnability and expressivity across more than 1000 languages. *Entropy* 19 (6): 1–32.

Čech, R., Kubát, M. (2018). Morphological Richness of Text. In: Fidler, M., Cvrček, V. (eds.) Taming the Corpus. *From Inflection and Lexis to Interpretation. Quantitative Methods in the Humanities and Social Sciences*. Cham: Springer, 63-77.

Chen R., Liu H., Altmann G. (2016). Entropy in different text types. *Digital scholarship in the humanities* 32(3): 528–542.

Coloma, G. (2015). The Menzerath-Altmann law in a cross-linguistic context. *SKY Journal of Linguistics* 28: 139-159.

Coloma, G. (2017). The Existence of Negative Correlation between Linguistic Measures across Languages. *Corpus Linguistics and Linguistic Theory* 13: 1-26.

Coloma, G. (2020). Language complexity trade-offs revisited, *Serie Documentos de Trabajo*, No. 721, Universidad del Centro de Estudios Macroeconómicos de Argentina (UCEMA), Buenos Aires.

Fenk-Oczlon, G., Fenk, A (1999). Cognition, quantitative linguistics, and systemic typology. *Linguistic Typology*, pp. 151-177.

Fenk-Oczlon, G., Fenk, A. (2014). Complexity trade-offs do not prove the equal complexity hypothesis. *Poznan Studies in Contemporary Linguistics*. Volume 50, Issue 2, Pages 145–155.

Fenk-Oczlon G., Pilz J (2021). Linguistic Complexity: Relationships Between Phoneme Inventory Size, Syllable Complexity, Word and Clause Length, and Population Size. *Frontiers in Communication*. 6:626032. doi: 10.3389/fcomm.2021.626032.

Hawkins, J. A. (2004). *Efficiency and Complexity in Grammars*. Oxford UP.

Jakobson, R. (1936). Beitrag zur allgemeinen Kasuslehre, Gesamtbedeutungen der russischen Kasus. In: PLingCP 6: 240-288.

Köhler, R, (1986). *Zur linguistischen Synergetik: Struktur und Dynamik der Lexik*. Bochum: Brockmeyer.

Köhler, R, (2005). Synergetic Linguistics. In: Köhler, Gabriel Altmann & Rajmund G. Piotrowski (eds.). *Quantitative Linguistik. Ein internationales Handbuch. Quantitative Linguistics. An International Handbook*. Berlin, New York: Walter de Gruyter, 760-775.

Koplenig A, Meyer P, Wolfer S, Müller-Spitzer C. (2017). The statistical trade-off between word order and word structure – Large-scale evidence for the principle of least effort. *PLoS ONE* 12(3).

Levshina, N. (2019). Token-based typology and word order entropy: A study based on Universal Dependencies. *Linguistic Typology* 23(3). 533– 572.

李在鎬· 石川慎一郎· 砂川有里子(2018)『新· 日本語教育のためのコーパス調査入門』くろしお出版.

李文平, 刘海涛, 熊子涵 (2022).「日本語における語順の自由度と格標識の豊富さに関する計量的研究」.『計量国語学』33 (5), pp. 325-340.

McFadden, T. (2003). On morphological case and word-order freedom. *Berkeley Linguistics Society* (BLS) 29. 295–306.

Menzerath, P. (1954). *Die Architektonik des Deutschen Wortschatzes*. Hannover; Stuttgart: Dümmler.

Michael A., McFall, D. (2010). Cutting the Gordian Knot: The Moving-Average Type–Token Ratio (MATTR), *Journal of Quantitative Linguistics*, 17:2, pp. 94-100.

Miestamo, M. (2009). *Implicational hierarchies and grammatical complexity*. Oxford University Press.

奥野由紀子· 呉佳穎· 村田裕美子 (2019).「日本語学習者の能動態と受動態の使用傾向にみられる母語による違いー中国語とドイツ語での語りの比較からー」『日本語研究』39, 79-93.

迫田久美子· 細井陽子 (2020) 異なった学習環境における日本語使用の正確さと複雑さ-日本語学習者コーパス（I-JAS）の分析に基づいて- 『計量国語学』32 (7), 403-418.

Sapir, E. (1921). *Language: An Introduction to the Study of Speech*. New York: Harcourt, Brace & World Inc., 33-35.

Shannon C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 1948, 27(4): 623–656.

Shosted, R. K. (2006). Correlating complexity: A typological approach. *Linguistic Typology*, *10*(1), 1-40.

Sinnemäki, K. (2008). Complexity trade-offs in core argument marking. In Miestamo, M., K. Sinnemäki, K. and F. Karlsson, F. (eds.), *Language Complexity: Typology, Contact, Change* (Amsterdam: John Benjamins), 67–88. doi: 10.1075/ slcs.94.06sin.

Sinnemäki, K. (2014). Complexity trade-offs: A case study. In: F.J. Newmeyer and L.B. Preston (eds.), *Measuring Grammatical Complexity*, 179–201. Oxford: Oxford University Press.

角田太作 (1991). 世界の言語と日本語 改訂版—言語類型論から見た日本語. 東京：くろしお出版。

宇佐美洋 (2006). フランス語母語話者の日本語作文における「意図不明表現」の分析−母語訳との対照から見る「分かりにくさ」の理由−『作文対訳データベースの多様な利用のために−「日本語教育のための言語資源及び学習内容に関する調査研究報告書」 81-99.

Xanthos, A., Gillis, S. (2010). Quantifying the development of inflectional diversity. *First Language* 30, 175–198.

Yan, J, Liu, H. (2021). Morphology and word order in Slavic languages: Insights from annotated corpora. *Voprosy Jazykoznanija* 4: 131–159.

Zipf G. K. (1949). *Human behavior and the principle of least effort: An introduction to human*

ecology (Cambridge: Addison-Wesley).