トポロジー的視点による言語の複雑さの考察

中山拓人
大阪大学大学院人文学研究科
nakayama.takuto.hmt@osaka-u.ac.jp

概要

本研究は，言語の複雑さに関して，「言語間比較が可能な形」での分析を目指し，特定の言語に依存しない枠組みによるアプローチの提案を目的とする．具体的には，[1] によって提案された，トポロジーの概念であるベッチ数 (Betti number) を単語レベルの n-gram 系列に適用しする word manifold の概念を用い，単語列の持つ構造的差異を言語間で比較する．結果として，ある文書データから得られるベッチ数に対する言語の順位が，他の文書でも観察された．このことから，word manifold が捉える構造的特徴が言語ごとに異なること，すなわち各言語の複雑さには，差異があることが示唆された．

1 はじめに

言語学者は長い間言語の複雑さについて関心を持ってきた．特に，20 世紀初頭の構造主義者の一部が，全ての言語は等しく複雑であることを示唆して以来，「言語の等複雑さ（equi-complexity of language）」として知られるこの考え方は，言語学の複数の下位分野において研究対象となってきた．しかし，その比較的長い研究史にもかかわらず，研究者の間では，言語の全体的な複雑さとは何か，どのように測定すべきか，さらにはそれをどのように定義すべきかについて，いまだに合意には至っていない．これは技術的な問題に加え，とりわけ「言語の全体的複雑さ」という概念自体が曖昧であることに起因している．

実際，技術的問題に関しては，計算機技術が成熟し，言語の複雑さの測定に貢献できるようになったことで，限定的ではあるものの，現在では克服されつつあると言える．21 世紀に入って以降，計算機技術の発展は，言語全体の複雑さに到達するための多様な手法をもたらしてきた．これらの手法でによって，計算規模が計算機技術の存在しなかった時代と比べて，はるかに大きくなっている．しかし，「全ての言語は等しく複雑なのか」という問いは，依然として論争の的であると言える．その主な理由は，現在提案されているアプローチが，言語全体の複雑さに到達するには至らず，あくまでその近似にとどまっているからである．そこで本研究は，言語の複雑さを全体的に捉えるためのアプローチを提案することを目的とし，[1] によって導入された word manifold に基づいて，多言語比較を行う．Word manifold とは，幾何学の一分野であるトポロジーの概念を，単語レベルの n-gram の文脈に置き換えて，その内部構造を考察する枠組みである．

以下では，2 節において，言語の複雑さに関する研究の歴史と，本論文が不十分であると主張する言語全体の複雑さ計測のアプローチについて，先行研究を概観する．3 節では，[1] に基づき，単語レベルの n-gram 系列にトポロジーの概念を適用した word manifold を説明する．4 節では分析結果を報告し，5 節では結論，および今後の展望を示す．

2 先行研究

言語の等複雑性へ言及を行った最初期の研究者の一人は，[2] である．それまでは，コミュニティの文化水準とその言語の複雑さが相関すると考えられていた [3] が，フィールドワークなどを通じて，そのような帝国主義的言語観に対する反例が挙げられていた．

> Both simple and complex types of language of an indefinite number of varieties may be found spoken at any desired level of cultural advance. When it comes to linguistic form, Plato walks with the Macedonian swineherd, Confucius with the head-hunting savage of Assam. (268–269)

加えて，[4] もまた，言語の等複雑性へ言及した主要な文献の 1 つである．

> Impressionistically it would seem that the total grammatical complexity of any language, counting both morphology and syntax, is about the same as that of

any other. (180)

言語の内，一部の側面の複雑さが他言語と比較して高い場合，反対にそれ以外の側面の中に，多言語と比較して複雑さが低くなるものが見られるという，トレードオフの傾向は，経験的に指摘されており，言語の等複雑性も，そのトレードオフ関係によって支持されてきた歴史がある [5].

このような考え方は，20 世紀初頭にはすでに現れていたが，それを実証しようとする動きは，人間が手作業で行うよりもはるかに大規模なデータを扱うことのできる計算機技術の登場を待たなければならなかった．近年の実証的な研究例の一つが [6] によるものである．彼らはワークショップの中で，参加者が各言語のさまざまな側面を評価させ，それらの評価を次元として格納するベクトルを作成し，それらを比較することで，言語全体の複雑さ差の間に有意差があるのかを調査した．結果として，各言語間においては有意差はほぼ認められず，言語の等複雑性の存在が示唆された．. 複数の側面に関して評価し，ベクトルとして扱う手法は，もともと [7] によって導入されたものである．この方法の利点は，言語の複数の下位領域における複雑性を同時に考慮できる点にある．その意味では，この手法は言語の全体的な複雑さを捉えているように見える．しかし，この方法が提供できるのは，その近似にすぎないと言えよう．なぜならば，全体的な言語の複雑さを構成するために必要な側面の数が，どれほどあるのかは誰にも分からず，それが無限である可能性すらあるからである．さらに，単一の側面がどれほどの情報を含んでいるのか，他の側面とどの程度重複しているのか，またそれらが言語間でどの程度変動するのかも，明確ではない．例えば，「単語」という領域に関して勘げてみても，各言語で「単語」と呼ばれる単位が果たす役割が一定ではなく，各下位領域に関して言語固有の条件を考慮しない限り，比較可能な形でベクトルを得ることはできない．したがって，言語の全体的複雑性を達成しようとするのであれば，言語の下位領域ごとの評価を単純に収集し，それらをベクトルとして統合するだけでは不十分であると言える．

この課題を乗り越えるためには，少なくとも 2 つの条件を満たすやり方で言語の複雑さを捉えていなくてはならないと言える．1 つ目は，言わずもがなであるが，ある言語全体の複雑さを考慮できていることである．[**?**] は，少なくともこの点に関して克服を目指していたと位置付けることができる．2 つ目は，言語間で比較可能な形で言語の複雑さを考慮できていることである．特に言語の等複雑性にかに s ての考察を行う場合，言語間比較が必須であるため，それを念頭に置いたやり方で言語の複雑さを考察できなければならない．本研究は，特にこの 2 点目に関して克服を試み，言語を下位領域に分割することなく，特に言語依存的でない分析手法を提案，実践することを目的とする．そのために，単語 n-gram をトポロジー的に捉えることで，その内部構造を分析する word manifold[1] を用いて，言語の複雑さを考察する．この手法は簡潔に言えば，その言語の内部構造に，いくつの「空洞的構造」が存在するのかを示すものであり，その詳細については次節で紹介する．

3 方法論

3.1 Word Manifold [1]

2 節で述べた通り，この枠組みは，言語をトポロジー的な対象として捉え，その構造内に存在する「空洞」の数を明らかにすることを計算，比較することを目的としている．一般に，点は線を形成し，線は平面を形成し，平面は空間を形成し，これは次元の段階的な増加に伴っている．本手法では，このような点・線・平面の関係を，それぞれ単語 n-gram の uni-gram，bi-gram，tri-gram に対応させることで，言語をトポロジー的に捉えようとする．したがって，ここでいう「空洞」とは，n-gram 配列の中でも，特定の構造を指していることになる．

分析は次の三つの段階から成る．(i) スケルトン (skeleton) の取得，(ii) 境界行列 (boundary matrix) の作成，(iii) ベッチ数 (Betti number) の算出である．以下の分析の説明は，全て [1] に基づいている．なお，分析の流れは付録

分析の第一段階は，コーパスから全てのスケルトンを取得することである．n 次元のスケルトン (Sn) とは，以下の条件を満たす全ての $(n+1)$ 語の配列のことで，$[w_0, w_1, \ldots, w_n]$ の集合から成る．満たすべき条件とは，$w_0, w_1, \ldots, w_n$ に含まれる語から構成される任意の長さの部分列得た時，その部分列がコーパス C におけるトークン数 k のウィンドウ内のどこかに必ず出現することである．自然に，S_0 はコーパスの語彙のリストそのものを表し，S_{k-1} はこの設定における最大長の語列，すなわち k-gram

のリストを表す．本分析では，以下の式におけるウィンドウサイズ k を 3 から 10 までとし，スケルトンの次元は 0 から $k-1$ までとした．

第二段階では，これらのスケルトンに基づいて，$\mathfrak{B} := \{B_n | n \in 0, \ldots, k-1\}$ という行列の集合を得る．ここで k はウィンドウサイズであり，この B_n は境界行列と呼ばれる．境界行列 B_n の各要素は，各列について，対応する n 次元スケルトンの要素 $s \in S_n$ を考え，その形式的境界 ∂s を以下の式によって定義する．ここで，ハット記号 $(\hat{w_i})$ は，その下にある項が取り除かれていることを表す．

$$\partial[w_0, w_1, \ldots, w_n] = \sum_i (-1)^i \{w_0, w_1, \ldots, \hat{w_i}, \ldots, w_n\} \tag{1}$$

こうして得られる境界行列 B_n は，その成分が $\{-1, 0, 1\}$ からなる，非常に疎な行列である．

最後に，境界行列の集合 $\mathfrak{B}$ から，n-word 全体に含まれる「空洞」の数を求めることができる．本手法では，その「空洞」の数はベッチ数を計算することによって与えられる．ベッチ数は次の式によって定義される．

$$\text{Betti}_n = dim(\ker(\partial_n)) - dim(\text{im}(\partial_{n+1})) \tag{2}$$

ここで，$\ker(\partial_n)$ は n 次元の核 (kernel) を，$\text{im}(\partial_{n+1})$ は $n+1$ 次元の像 (image) を表す．

以上のように，テキストの n-gram における幾何学的な構造を見出そうとする本手法は，ある言語に依存した理論に依拠しているわけではなく，その意味で，妥当な言語間比較が可能な形で言語を評価していると言える．もちろん，n-gram への分割の差異には，各言語に特化したトークナイザの存在が前提となるが，手続きそのものは，言語固有のものではなく，2 節の終わりに挙げた，２つ目の条件である「言語間で比較可能な形で言語の複雑さを考慮できていること」を満たしていると考えられる．なお，分析アルゴリズムは付録に示されており，GitHub にて公開されている[1)]．

3.2 データセット

本研究では，文書の内容を揃えるため，並行テキストとして四福音書（マタイ，マルコ，ルカ，ヨハネ）を用いる．並行テキストを使用する理由は，文脈を統制することにより，word manifold における言語間の差異が，意味内容の違いによるものではなく，各言語固有の特性に起因するものであることを明確にするためである．すなわち，言語間で結果に差が見られた場合でも，それが「何を意味しているか」の違いではなく，言語そのものの性質に由来することを示すためである．

加えて，利用可能性の観点からも，聖書は都合が良い．これらのテキストはいずれも多言語版がオンラインで公開されている．マタイ，マルコ，ルカ，ヨハネの各福音書は，それぞれ 1071 節，678 節，1151 節，879 節から成っており，分析手続きにおいては各節を超えて n-gram が生成されないように処置を行なった．本研究で用いたデータセットは聖書コーパス [8] に基づく．トークン化には，Stanza[9] を使用した．分析対象とした言語数は，それぞれ 31 言語および 40 言語であり，これらの数は，Stanza で利用可能な言語と，元データに含まれている言語の双方に含まれるものである．

4 結果

下の箱ひげ図は，各言語におけるテキストのベッチ数を示している．図の k のサイズは 10 であり，ゆえに次元数 n は 0 から 9 まで存在する．横軸はスケルトンの次元を表し，縦軸はベッチ数を表す．箱は，各次元における言語間のベッチ数のばらつきの程度を示している．

結果によると，低次元および高次元の双方，すなわち $n = 0, 1$ や $n = 8, 9$ では，ベッチ数は非常に低く，分散も小さい．一方で，中間的な次元である $n = 4, 5, 6$ では，各言語において最大のベッチ数が観察される．[1] によると，$n = 0, 1$ のような低次元の word manifold におけるベッチ数は，文法的な特性を表しているとされ，反対に高次元のベッチ数はより大域的なランダムさを表しているとされている．実際，各次元でのベッチ数の順位順に言語を並べると，どの文書でも低次元時点ではコプト語が，$n = 4, 5$ 程度の高次元になると日本語が最も高い値を示した．また値の低い言語にも目を向けると，どの次元どの文書でも，韓国語が最も低い値を示した．また，それ以外の言語も，概ねその順番を維持していることが，文書間の順位に対する標準偏差 1 からわかる．次元数の増減に伴う各言語の値の順位の移り変わりが，文書を跨いで維持されていることから，言語間で観察された値の差異は，言語の特徴

1） https://github.com/takuto-nakayama/malc

を表していると考えられる．

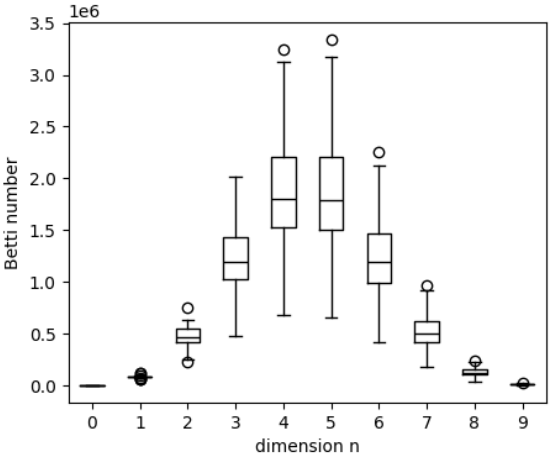

図 1　マタイの福音書

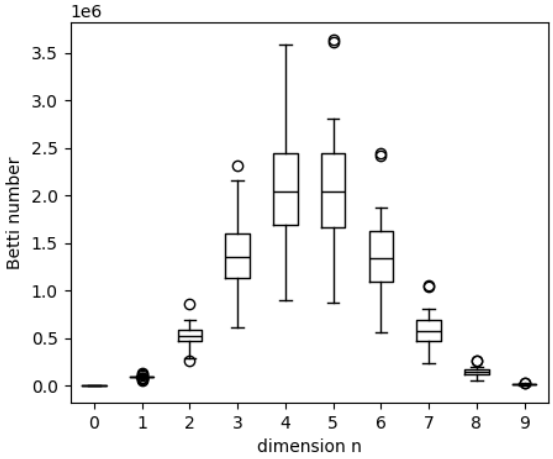

図 3　ルカの福音書

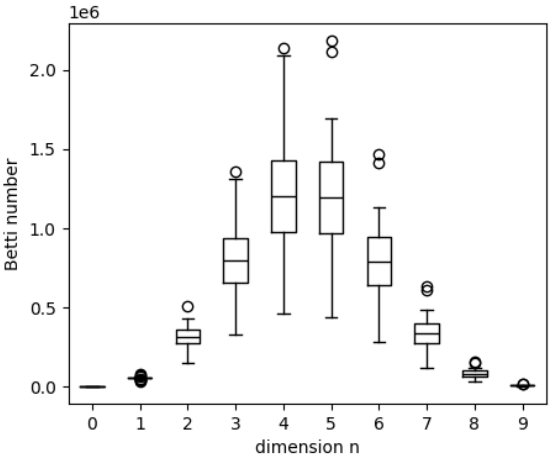

図 2　マルコの福音書

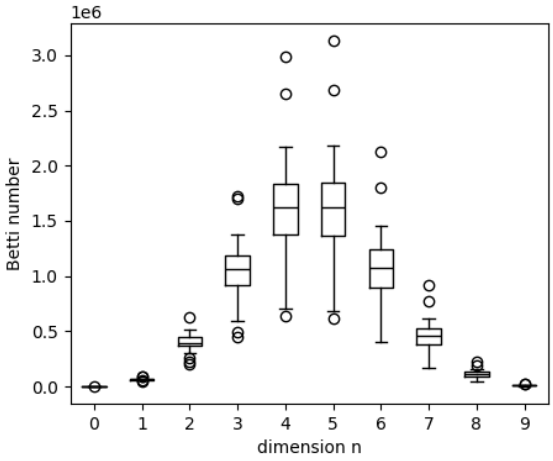

図 4　ヨハネの福音書

5　結語

本論文は，単語 n-gram 系列に対するトポロジー的アプローチである，word manifold[1] を使って，言語の複雑さへの幾何学的アプローチを提案した．本手法は，「言語間の比較を可能にする形で言語の複雑さを考える」という条件に焦点を当て，特定の言語に依拠した枠組みではない枠組みを用いた．文書間における，各言語のベッチ数の順位の標準偏差 1 を見ると，ほぼ全ての言語の順位が，ほとんど変わらないことがわかった．したがって，本研究の結果は，言語の複雑さの関して，言語間で差異があることが示唆された．

本研究で残されている課題として挙げられるのは，特定の言語に依存的でないという側面の精度を上げる必要がある．本研究では，対象とする言語単位に単語を用いているが，それが真の意味で言語依存的でないと言えるかは，明白ではない．その意味で，本手法をさまざまな単位に対して適応することで，言語依存的でない観点を模索する必要があると言える．また，[1] でも指摘されている通り，特に高次元のベッチ数は，その言語の埋め込み空間における幾何学的構造との関連が示唆されるが，その具体的な役割はまだ明らかではない．今後の展望としては，「言語全体の複雑さに言及すること」も目指し，幾何学的なアプローチによる言語の複雑性計測の手法を模索していきたい．

謝辞

本研究は JSPS 科研費 JP24KJ1938 の助成を受けたものです.

参考文献

[1] Stephen Fitz. The shape of words—topological structure in natural language data. 第 196 巻, p. 116–123. PMLR, 2022.

[2] E Sapir. **An introduction to the study of speech**. Citeseer, 1921.

[3] August Schleicher. **Die Sprachen Europas in systematischer Übersicht**. 1850.

[4] Charles F Hockett. **A course in modern linguistics**. Oxford & IBH Publishing Co., 1958.

[5] Gertraud Fenk-Oczlon and August Fenk. Complexity trade-offs do not prove the equal complexity hypothesis. **Poznan Studies in Contemporary Linguistics**, Vol. 50, No. 2, pp. 145–155, 2014.

[6] Christian Bentz, Ximena Gutierrez-Vasques, Olga Sozinova, and Tanja Samardžić. Complexity trade-offs and equi-complexity in natural languages: a meta-analysis. **Linguistics vanguard : multimodal online journal**, Vol. 9, No. s1, pp. 9–25, 2023.

[7] Guy Deutscher. “overall complexity”: a wild goose chase? In Geffrey Sampton, David Gil, and Peter Trudgill, editors, **Language complexity as an evolving variable**, p. 243–251. Oxford University Press, Oxford, 2009.

[8] Christos Christodouloupoulos and Mark Steedman. A massively parallel corpus: the bible in 100 languages. **Language Resources and Evaluation**, Vol. 49, No. 2, pp. 375–395, 2015.

[9] Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. Stanza: A Python natural language processing toolkit for many human languages. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations**, 2020.

付録

分析アルゴリズム

Corpus C, window size k Betti numbers $\{\beta_n\}_{n=0}^{k-1}$

Initialize empty skeleton $S_n \leftarrow \emptyset\ \forall n \in \{0, \ldots, k-1\}$

for each sentence $s \in C$ **do**
    k-gram words $w_s (\in W) \leftarrow$ get_ngrams(s, k)
end for
for $n \leftarrow 0$ to $k-1$ **do**
    for each $(n+1)$-word subsequence $[w_0, \ldots, w_n]$ in w **do**
        if all subsequences of $[w_0, \ldots, w_n]$ appear in C **then**
            Add $[w_0, \ldots, w_n]$ to S_n
        end if
    end for
end for

for $n \leftarrow 0$ to $k-1$ **do**
    $\partial_n [t_1, \ldots, t_n] = \sum_i (-1)^i [t_1, \ldots, \hat{t_i}, \ldots, t_n]$
end for

for $n \leftarrow 0$ to $k-1$ **do**
    $\beta_n \leftarrow \dim(\ker \partial_n) - \dim(\mathrm{im}\, \partial_{n+1})$
end for
return $\{\beta_n\}_{n=0}^{k-1}$

ベッチ数の順位

表 1　文書間の順位に対する標準偏差

$n =$	0	2	4	6	8
アフリカーンス語	0.577	1.414	1.5	0.577	0.5
アラビア語	0.5	0.577	0.5	0.5	0.5
アルメニア語	0.5	0.5	0.577	0.5	0.5
バスク語	0	0.5	1.826	1.826	2.062
ブルガリア語	0.577	1.414	1.291	1.708	1.708
コプト語	4.243	0	0.5	0	0
クロアチア語	1.732	1.633	0.816	0.957	0.957
チェコ語	0.577	0.5	0.816	1.258	1.258
デンマーク語	2.38	0.5	1.732	2	2
オランダ語	0.816	0.957	3.5	4.5	4.5
英語	0.577	0.816	2.062	2.63	2.217
エストニア語	10.145	8.524	7.805	6.856	6.856
フィンランド語	1.258	1.291	1.258	1.258	1.258
フランス語	0.957	0.816	0.577	0	0.5
ドイツ語	0.577	0.957	0.957	0.816	0.816
ギリシャ語	1.258	0.577	1.291	0.816	0.816
ヒンディー語	2.16	1.915	3.109	3.697	3.594
ハンガリー語	0.5	1.258	0.816	0.816	0.816
アイスランド語	0.577	1.708	0.816	1.258	1.258
インドネシア語	1.291	1.291	2.754	2.63	2.63
日本語	2.754	0.577	0.5	0	0
韓国語	2.062	1	1	0.5	0.5
ラテン語	2.217	0.577	0.577	0.5	0.5
ラトビア語	0.816	1.732	1.258	0.816	0.816
リトアニア語	1	0.5	0.577	0.577	0.577
マン島語	0.957	0.957	0	0	0
マラーティー語	3.559	1.915	6.608	4.787	4.435
ノルウェー語	2.062	1	1.5	1.5	1.5
ポーランド語	0.957	2.363	2.63	2.63	2.63
ポルトガル語	1.414	0.816	0.577	0.5	0.5
ルーマニア語	0.957	1.5	1.708	1.732	0
ロシア語	0.957	0.816	0.5	1.291	1.291
セルビア語	1.708	1.291	1.155	1.5	1.155
スロバキア語	1.732	1.893	0.5	0	0
スペイン語	0.957	1.155	0.5	0.5	0.5
スウェーデン語	1.414	0.957	2	1.5	1.414
テルグ語	0	0	0.5	0.5	0.5
トルコ語	0	1.5	0.957	0.957	0.957
ベトナム語	2.16	2.5	2.062	1.826	1.826
ウォロフ語	3.202	2.082	2.944	2.363	1.826