Q8-16

# トポロジー的視点による言語の複雑さの考察

2026/03/12
NLP2026@ライトキューブ宇都宮

中山拓人（大阪大学）
email: nakayama.takuto.hmt@osaka-u.ac.jp

大阪大学
THE UNIVERSITY OF OSAKA

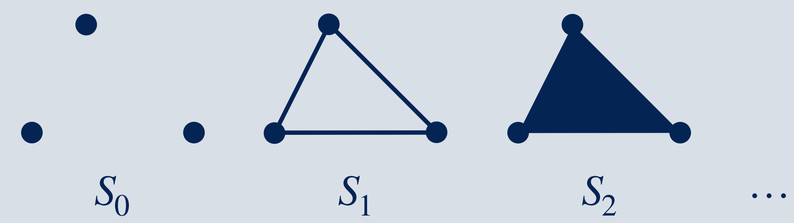
## 1.概要

- 「言語の複雑さ」を言語間で比較したい
  →比較可能性の検討が十分でない[1]
- 特定の言語に依存しない枠組みの一候補として，幾何学的なアプローチを試みる
- トポロジーの概念であるベッチ数を単語n-gram 系列に適用するword manifold[2]を用いた

## 2.データ

- 新約聖書の４つの福音書(Matthew, Mark, Luke, John) を並行コーパスとして使用
- 聖書コーパス内40言語[3]を比較
- 比較に際しては，Stanzaにてトークン化[4]
  * 40言語は，[3]と[4]の対応関係における，共通集合に含まれるもの全て

## 3.Word Manifold [2]




$$\partial[w_0, w_1] = [w_1] - [w_0]$$

$$\partial[w_0, w_1, w_2] = [w_1, w_2] - [w_0, w_2] + [w_0, w_1]$$

$$\vdots$$

$$\mathrm{Betti}_n = dim(\ker(\partial n)) - dim(\mathrm{im}(\partial_{n+1}))$$

**1. 全ての$n$次スケルトン($S_n$)を取得する**

- $n$次スケルトン: 任意の長さの部分列がウィンドウ内に出現するような全てのn-gram列のこと
- $S_0$は単語リスト，$S_{k-1}$はk-gram列群のこと

**2. 境界行列($B_n$)を作成する**

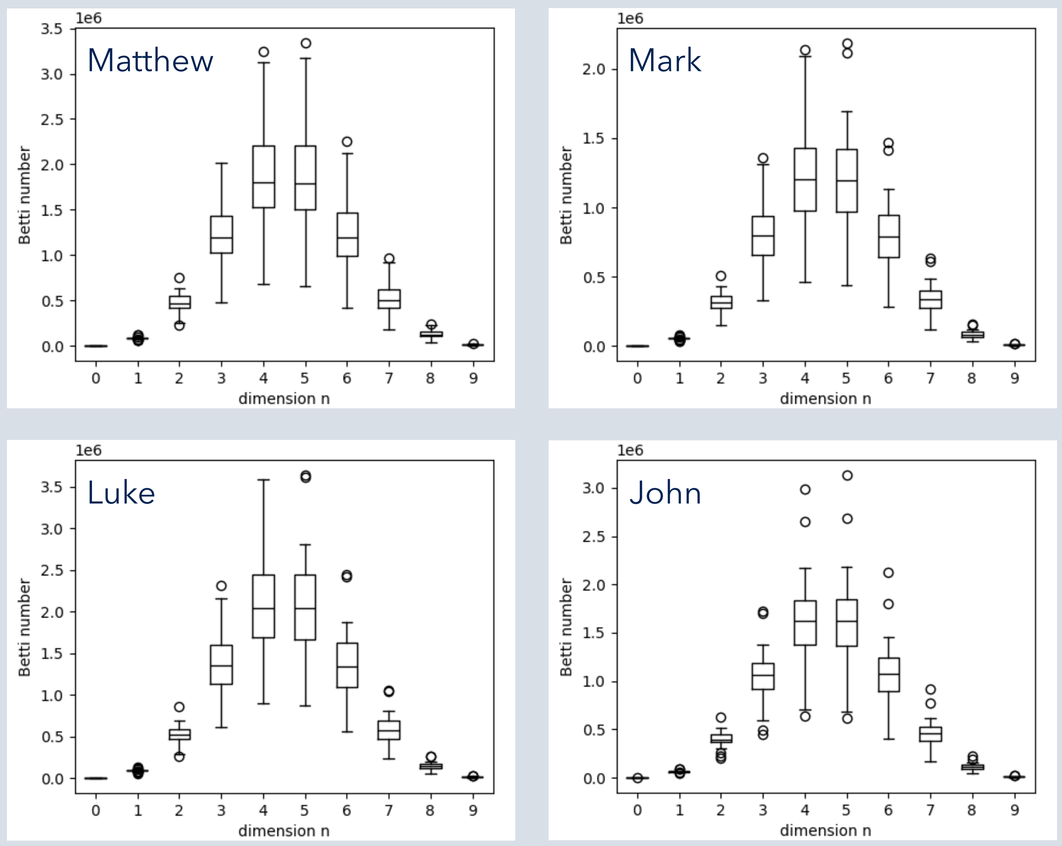
$$\partial[w_0, \ldots, w_n] = \sum (-1)^i \{w_0, \ldots, \hat{w}_i, \ldots, w_n\}$$

- ^: その項は取り除かれている
- 実質的に単体ホモロジーと呼ばれるものに等しい

**3. ベッチ数($\mathrm{Betti}_n$)を計算する**

- $\ker(\partial_n)$: $n$ 次元の核(kernel)
- $\mathrm{im}(\partial_n)$: $n$ +1 次元の像(image)
- n-gram列に含まれる「空洞」に相当する構造の数を数えることと同義

## 4.結果




## 5.結論・今後の課題

- Word Manifoldによる文書の幾何学的分析結果が，何らかの言語差を示唆していることが観察された
- 「単語」という単位を見ていることが言語依存でないと本当に言えるか？
- 幾何学的量が言語について表すものの正体に関しては，まだ明確な答えがない

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] Matti Miestamo. On the feasibility of complexity metrics. In *Finest linguistics. Proceedings of the Annual Finnish and Estonian Conference of Linguistics*, pages 11–26. 2006.
[2] Stephen Fitz. The shape of words—topological structure in natural language data. In *The shape of words–Topological structure in natural language data*, *196*, 116–123. 2022.
[3] 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.
[4] 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.