言語は等しく複雑か？:
多義語埋め込み表現による形式–意味対応の複雑性

中山拓人
慶應義塾大学
tnakayama.a5ling@gmail.com

概要

「あらゆる言語は，等しく複雑である」という言語の性質は，言語学において長く信じられてきた．実際に，この性質を支持するような事例が報告されている一方で，反証的な研究結果も同時に挙げられており，現在に至るまで，その真偽については未解決である．本研究は，言語全体の複雑性を決定する主要な要因の 1 つとして，形式 − 意味の対応関係に注目し，情報理論に基づく複雑性計測手法の提案，及び多言語間の複雑性比較を行った．結果として，極端に複雑な対応関係を持つ言語は無いが，他に比べて，より単純な対応関係を持つ言語が存在すること，及び言語の等複雑性はそれほど強い性質ではない，ということが示唆された．

1 はじめに

言語学において，「あらゆる言語は，等しく複雑である」という性質が，広く信じられてきた．一般に言語の等複雑性 (linguistic equi-complexity) と呼ばれるこの性質は，70 年近く前から示唆されているものである．例えば，Hockett は，“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” [1, p. 180] と述べている．実際に，Everett は Pirahã 語について，“[n]o one should draw the conclusion from the paper that the Pirahã language is in any way ‘primitives’. It has the most complex verbal morphology I am aware of. And a strikingly complex prosodic system” [2, p. 62] と主張しており，等複雑性を支持する事例を報告している．この様に，言語間において，個別の側面ごとの複雑性に差異が見られても，全体を均した複雑性には差がないという性質が，暗黙の裡に支持されてきた．しかし，この性質の真偽については，後述の通り，未解決と言わざるを得ない状況にある．この主たる理由として挙げられるのは，言語全体の複雑性を計測する方法について，統一的な見解が無いことである．

そこで本研究は，言語全体の複雑性を決定する主要な要因の 1 つが，形式 − 意味の対応関係であると仮定し，その尺度を用いた複雑性計測手法の提案，及び多言語間の複雑性比較を実践する．以下では，2 節で，これまでの複雑性研究の概観と，意味の側面を考慮した計測をするための，埋め込み表現について概観する．その後，3 節で方法論の提案を行い，4 節で多言語比較の結果と，その考察を示す．

2 先行研究

2.1 言語の等複雑性

コンピュータ技術の向上により，2000 年代以降から，膨大な計算力を用いた複雑性研究が行われてきたが，現在に至るまでも，言語の等複雑性に対する支持／不支持のいずれの立場も見られる．例えば，Bentz らの研究では，言語全体の複雑性を，様々な側面の情報を格納したベクトルとして表現する手法を用いた [3]．それを用いて，各次元間における相関の有無によって，各側面同士のトレードオフ関係に，ベクトル間の有意差の有無によって，言語全体の複雑性にアプローチした．結果として，トレードオフ関係にはほとんど有意な相関は見られなかったが，その一方で，言語全体の複雑性については有意な差が見られず，言語の等複雑性を示唆する結果が示されている．また，Shannon の情報理論 [4] に基づいて，単語についての uni-gram エントロピーとエントロピーレートを，多言語間で比較した研究もある [5]．結果として，2 つの尺度とも極めて狭い範囲に収まっており，言語の等複雑性が示唆される結果が示されている．

一方で，否定的な立場を示す研究もある．Koplenig らの研究では，「それ以前の系列が与えられた条件下で，次に出てくる要素の予測しづらさ」を言語の複雑性と定義し，情報エントロピーとして算出したその値を，コーパス間で比較した [6]．結果として，あるコーパスの中でエントロピーが高い／低い言語は，別のコーパス内でも，同様に高い／低いエントロピーを持つという，言語の等複雑性に対して否定的な示唆がなされている．また，人口の多いコミュニティで使用される言語程，エントロピーが高い，即ち複雑性が高い傾向があることを指摘しており，実際に言語が使用される環境が，言語の複雑性を決める要素の 1 つであることが示唆されている研究もある [7]．この様に，現在に至るまで，言語の等複雑性についての論争は，未解決である．

これまでの複雑性研究は，形式的側面の分析が主な関心の対象であるが，これは多くの研究で活用されている，Shannon の情報理論 [4, p. 1] で明示されている態度である．

> Frequently the messages have meaning; that is they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem. The significant aspect is that the actual message is one selected from a set of possible messages. The system must be designed to operate for each possible selection, not just the one which will actually be chosen since this is unknown at the time of design.

Shannon は，上記の様に言及しており，この態度は現在の研究にも大きく継承されているといえる．しかし，言語の複雑性研究において，意味的側面を考慮に入れるべきであるという手法も，昨今取り上げられている (e.g., [8])．本研究は，形式的側面だけでなく，意味的側面も考慮した上で，言語の等複雑性にアプローチすることを目的とする．

2.2 多義性と埋め込み表現

構文文法などで言われている，形式と意味のペアリングを，言語の基本的な単位という立場に立てば，形式 − 意味の対応関係の複雑性が，言語全体の複雑性を決定する要因の 1 つであると見做せる．そして，形式 − 意味の対応関係の複雑性のあり方として，「ある形式を見た時に，その形式がどんな意味を表しているか，予測しづらい程，複雑であり，一対一対応に近い程，単純である」という尺度が考えられる．故に，この尺度を計測するためには，ある形式が，いくつの語義と結びついているのかを，自動で推定する手法が必要となる．

これを解決するために，本研究では文脈を考慮した埋め込み表現を用いる．埋め込み表現としてよく知られているのは，word2vec [9] [10] である．しかし，word2vec は文脈を考慮していないため，それから得られる埋め込み表現からは，形式が持つ複数の語義を推定することはできない．各トークンにつき得られた埋め込み表現であれば，それらを利用して語義巣を推定することができるだけでなく，各トークンがどの語義で使われているかも，得ることができる．

文脈を考慮した埋め込み表現を得られるモデルは，よく知られているものが複数ある．Devlin らが提案した BERT (Bidirectional Encoder Representation from Transformers) [11] は，双方向の文脈を学習することで，埋め込み表現を得るが，事前学習だけでなく，タスクに合わせた微調整を組み合わせて行うことが，特徴である．Peter らが提案した ELMo (Embeddings from Language Model)[12] は，双方向 LSTM を用いて，各トークンに対し入力文脈全体を加味した埋め込み表現を算出する点が特徴である．この特徴から，多義語をモデル化することが可能になっている．また ELMo には，多言語用の学習済みモデルが公開されており [13] [14]，本研究ではこれを理由に，以下で取り組む実験では，ELMo を使用した．

3 方法論

3.1 提案する手法

本研究が提案する手法では，「ある形式から，それが表している意味をどれだけ予測しづらいか」を言語全体の複雑性を決定する主要な要因と仮定し，Shannon の情報理論に基づいて，これを計算する．

手法のアルゴリズムとしては，まずデータとなるテクストを，任意の単位に分割し，それで得られたトークンごとの埋め込み表現を得る．ここで想定される単位とは主に単語であるが，場合によってはより大きな，複数単語の系列や，より小さい n-gram も考えられる．その場合，文脈に相当する単位を事前に設定することが想定される．次に得られたトークンごとの埋め込み表現を，タイプごとにクラスタリ

ングを行う．ここで得られたクラスター数が，そのタイプの語義の推定数となる．各クラスターに属するトークン数から出現確率の分布 p_k を算出し，それを基に各タイプのエントロピー H を，以下の公式で求める．

$$H = -\sum_k p_k \log_2 p_k$$

エントロピーに関しては，各タイプのエントロピーの単純な総和では，タイプ数に依存して増加するため，各タイプ自体の出現確率を掛けることで，平均化する．以下の Algorithm 1 は，疑似コードによる手順の記述である．

Algorithm 1: 形式 − 意味の対応関係についてのエントロピー計算

Data: Input: テクストを T，文脈の窓幅を w
Result: Output: エントロピー H

for each $t_i^* \in set(T)$ **do**
  for each $t_j \in T$ **do**
    if $t_i^* = t_j$ **then**
      $c_l \leftarrow T[j-w, j]$
      $c_r \leftarrow T[j+1, w+1]$
      $c_j \leftarrow c_l + t_j + c_r$
      $E_i \ni e_j \leftarrow \text{ContextEmbed}(c_j)$
      $Freq_i \leftarrow \text{count}(t_i^*)/\text{len}(T)$
  $E \ni E_i$
  $F \ni Freq_i$
for eacf $E_i \in E$ **do**
  $ClstDist_i \leftarrow \text{Cluster}(E_i)$
  for each $clst_k \in set(ClstDist_i)$ **do**
    $P_i \ni p_k \leftarrow \text{count}(clst_k)/\text{len}(ClstDist_i)$
  $P \ni P_i$
$H = -\sum_i Freq_i \sum_k^{\text{len}(P_i)} p_k \log_2 p_k$
return H

3.2 実験

本実験では，単語を対象として，10 言語 (英語，ギリシャ語，スペイン語，中国語，チェコ語，ドイツ語，日本語，フランス語，ヘブライ語，ロシア語) の比較を行った．文脈を考慮した埋め込み表現を各単語トークンに対して得るために，学習済みの ELMo モデルを使用した．多言語比較するにあたって，学習済みモデルの ELMoForManyLangs [13][14] を使用した．これは公開されている 44 言語の学習済みモデルであり，以下では 10 言語のモデルを使い，比較した．使用するデータは，コーパスから無作為にダウンロードした各言語 10,000 文を使用した．各コーパスと抽出したデータの情報を，それぞれ表 1，表 2 にまとめた．

表 1 コーパスの概要

	コーパス名	総語数
英語	enTenTen21	52,268,286,493
ギリシャ語	elTenTen19	2,342,091,029
スペイン語	esTenTen18	16,951,839,897
中国語	enTenTen17	13,531,331,169
チェコ語	csTenTen12+17+19	11,722,066,502
ドイツ語	deTenTen20	17,512,733,172
日本語	jaTenTen11	8,432,294,787
フランス語	frTenTen23	23,874,070,858
ヘブライ語	heTenTen21	2,775,686,699
ロシア語	ruTenTen17	9,034,837,939

表 2 使用データの概要

	抽出単語数 (/10,000 文)
英語	1,377,225
ギリシャ語	1,646,915
スペイン語	1,697,552
中国語	462,993
チェコ語	1,164,488
ドイツ語	1,395,315
日本語	731,999
フランス語	1,571,260
ヘブライ語	202,686
ロシア語	1,219,047

クラスタリング手法としては，BDSCAN を使用した．クラスター数を未知の語義数の推定値として利用するため，クラスター数を予め設定する必要が無い手法を用いた．そうして得られた，クラスターに対するトークンの分布から，エントロピーを算出した．

4 結果

算出された各言語についてのエントロピーは，図 1 の通りであった．全体として，0.5 未満の小さな値を示しており，最小値がチェコ語 (≈ 0.0938) で，最大値がフランス語 (≈ 0.4496) であった．ここで計算したエントロピーは，0 に近い程一対一対応に近く，1 に近い程 1 つの形式が平均して 2 つの語義に対応していることを意味している．その意味では，どの言語も 1 つの形式に対しては，1 つ以上 2 つ未満の語義のみが結びついており，極端に一対多数の様な，複雑な対応関係を示す言語は見られな

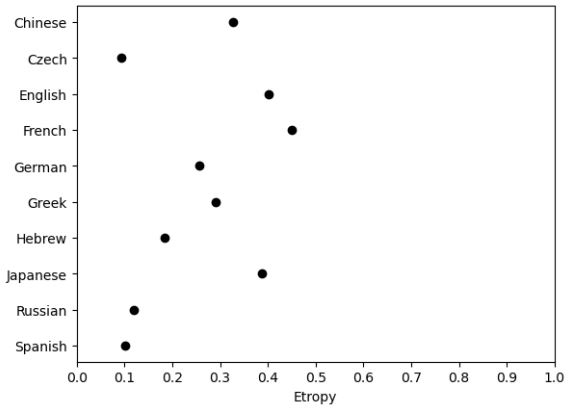

図 1　エントロピー比較

かった．

一方で，チェコ語とロシア語については，全体の中でも特に低い，0.1 未満の値を示している．このことから，形式と語義の対応関係が，極めて一対一対応に近い言語があることが示唆される．故に，形式－意味の対応関係における言語の複雑性は，極端に高い値を示す言語は無いが，他に比べて，より一対一対応に近い単純な対応関係を持つ言語が存在すること，及び，言語の等複雑性はそれほど強い性質ではない，ということが示唆される．

5　結語

本研究は，言語全体の複雑性を決定する主要な要因の 1 つとして，形式－意味の対応関係に注目し，情報理論に基づく複雑性計測手法の提案，及び多言語間の複雑性比較を行った．結果として，形式－意味の対応関係における言語の複雑性は，極端に高い値を示す言語は無いが，他に比べて，より一対一対応に近い単純な対応関係を持つ言語が存在すること，及び，言語の等複雑性はそれほど強い性質ではない，ということが示唆された．今後の課題として，本研究では，10 言語のみの分析に留まったため，学習済みモデルが公開されている 44 言語全ての分析が必要である．また，本研究が対象とした単位は単語であったが，単語以外の単位にも拡張が必要である．これは，単語という単位が明確に定義できない言語 (e.g., 日本語，中国語) においては，意味と結びついている形式系列の単位として，単語が妥当であるか否かは，明確ではないためである．最後に，意味を離散的にしか扱えていない点が，課題として挙げられる．本研究では，語義数をクラスター数によって推定したが，実際には，語義は連続値的な振る舞いをしていると考えられる．今後は，多次元空間に埋め込まれた分散表現を利用し，形式と連続値としての意味との対応関係を対象としていく必要がある．

謝辞

本研究は，JST 次世代研究者挑戦的研究プログラム JPMJSP2123 の支援を受けたものです．

参考文献

[1] Charles . Hockett. **A course in modern linguistics**. Macmillan, 1958.

[2] Daniel L. Everett. Cultural constraints on grammar and cognition in piraha . **Current Anthropology**, Vol. 46, pp. 621–646, 2005.

[3] Christian Bentz, Ximena Gutierrez-Vasques, Olga Sozinova, and Tanja Samardžić. Complexity trade-offs and equi-complexity in natural languages: a meta-analysis. **Linguistics Vanguard**, Vol. 9, pp. 9–25, 2022.

[4] Claude E. Shannon. A mathematical theory of communication. **The Bell System Technical Journal**, Vol. 27, pp. 623–656, 1948.

[5] Christian Bentz, Dimitrios Alikaniotis, Michael Cysouw, and Ramon Rerrer i Cancho. The entropy of words— learnability and expressivity across more than 1000 languages. **Entropy**, Vol. 19, p. 275, 2017.

[6] Alexander Koplenig, Sascha Wolfer, and Peter Meyer. A large quantitative analysis of written language challenges the idea that all languages are equally complex. **Scientific Reports**, Vol. 13, , 2023.

[7] Alexander Koplenig, Sascha Wolfer1, and Peter Meyer1. Human languages trade off complexity against efficiency. **Scientific Reports**, in press.

[8] Christian Bentz. Beyond words: Lower and upper bounds on the entropy of subword units in diverse languages. In **The 16th International Cognitive Linguistics Conference**, Düsseldorf, Germany, 2023.

[9] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeffery Dean. Distributed representations of words and phrases and their compositionality. In **Advances in Neural Information Processing Systems**, Vol. 26, pp. 3111–3119. Curran Associates, Inc., 2013.

[10] Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In **International Conference on Learning Representations**, 2013.

[11] 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, Minneapolis, Minnesota, 2019. Association for Computational Linguistics.

[12] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. New Orleans, Louisiana, 2012. Association for Computational Linguistics.

[13] Wanxiang Che, Yijia Liu, Yuxuan Wang, Bo Zheng, and Ting Liu. Towards better UD parsing: Deep contextualized word embeddings, ensemble, and treebank concatenation. In **Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies**, pp. 55–64, Brussels, Belgium, 2018. Association for Computational Linguistics.

[14] Murhaf Fare, Andrey Kutuzov, Stephan Oepen, and Erik Velldal. Word vectors, reuse, and replicability: Towards a community repository of large-text resources. In **Proceedings of the 21st Nordic Conference on Computational Linguistics**, pp. 271–276, Gothenburg, Sweden, 2017. Association for Computational Linguistics.