言語は等しく多義的か？
－サブワードと分散意味論に基づく形式-意味対応の分析－

中山拓人 (慶應義塾大学大学院生／日本学術振興会)

1. はじめに

「あらゆる言語は等しく複雑である」という,いわゆる言語の等複雑性の検証するために,本研究は (i) 自動処理による計算的アプローチと,(ii) 意味の側面を考慮したアプローチを基盤とした手法を用いての多言語比較を行う. 言語の等複雑性は,20 世紀の間に多くの言語学者によって言及されており,それ以来,言語学者の間で広く信じられてきた. その一方で, この言説は,確かな証拠に基づいて提唱されてきたものではなく,経験的にそう信じられてきたドグマに過ぎないとも指摘されている (Fenk-Oczlon & Fenk, 2014). そんな中, 昨今ではコンピュータ技術の発達により, 膨大なデータを用いた計算的手法による複雑性計測の研究が多く行われ (Ehret & Szmrecsanyi, 2016; Koplenig et al., 2023), 特に議論が盛んになってきていると言える. 一方で,「言語は等複雑であるか？」という問題に関しては, 未だ総意が得られるような解決には至っていない.

等複雑性の研究における障害の一つとして, 複数の言語間で妥当な比較が可能な計測手法が確立されていないことが挙げられる. 単語を例にとると, 分ち書きをしない言語においては, 単語の境界が明確ではない. またその様な言語に関して, 何らかの分割方法で″単語″を得られたとしても, それを基に計測された複雑性の値を比較することは, 自明ではないだろう. そのため本研究は, 計算アルゴリズムによって得られる形式単位を対象とする分析を行う. また言語の複雑性研究において, 意味の側面は形式的側面に比べて研究が盛んではないことも, 課題である. この課題の解決を目指し, 本研究は形式と意味の対応関係の複雑性を, 言語の複雑性の一側面と見做し, その言語間での比較を行う. 本研究の目的は, 計算アルゴリズムによって得られる形式単位と, それらが対応する意味の数を推定することで, 形式が如何に多義的 (または単義的) かの計測を試みることで, 言語の等複雑性にアプローチすることである.

2. 先行研究

2.1. 言語の等複雑性

言語の等複雑性について, 最初期の言及は,Sapir (1921, pp. 268-269) の“When it comes to linguistic form, Plato walks with the Macedonian swineherd, Confucius with the head-hunting savage of Assam.”だとされることが多い. またその後, Hockett (1958, p. 180) の“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”という発言が有名である. 形態論的複雑性と統語論的複雑性のトレードオフ関係は, その後の議論されていく一つのテーマとなる. しかし,Hockett が“impressionistically”ということからも見られるように, ここで言及されている傾向は, 少なくともこの時点では, 単に経験的に知られているものにすぎなかった.

同じ様な, 言語の特定の領域間におけるトレードオフ関係については, その後 Fenk-Oczlon and Fenk (1985) によって言及されてた. 彼女らは, 音素と音節のトレードオフ関係について指摘しており, 音節中の音素数が多ければ多いほど, 文中の音節数が少なく, 反対に文中の音節数が多ければ多いほど, 音節中の音素数が少なくなるというトレードオフを指摘した. このような, 言語のある領域の複雑性が増すと, その釣り合いをとるかのように, 別の領域の複雑性が下がるという傾向, ないし性質のことが, 一般に「言語の等複雑性」と呼ばれている.

2.2. 複雑性の計測手法

形態論的複雑さと統語的複雑さのトレードオフ関係に関しては,Juola (1998, 2008) が, 情報理論的手法として, コルモゴロフ複雑性を取り入れた, 計測手法ついての提案を行なっている. コルモゴロフ複雑性とは端的にいうと, その文章をがどれだけ短いアルゴリズムで再現可能か, という尺度のことである. 即ち, よく繰り返される形式は, 未使用のより短い形式に置換する等の操作により, 本来の文長よりも短い形で記述することが可能であるため, 複雑性が低いという想定である. Ehret and Szmrecsanyi (2016) はこの尺度を用いて, 英語の通時的比較, 及び多言語比較を行なった結果, 形態論的複雑性と統語的複雑性の間にトレードオフ関係が観察された.

また別の尺度として, シャノンの情報エントロピーを用いた研究も多い. Bentz et al. (2017) は, エントロピーを用いて, 単語選択の複雑性を計算し, 言語間でその差が非常に小さいことを示した. Koplenig et al. (2023) は, 各形式の出現頻度から得られるエントロピーを多言語間で比較し, 観察される言語間の差異が複数のコーパスで共通していること, 即ち複雑性が言語間で有意味に異なることを示唆した.

以上のように, 計算的手法により複雑性の計測が発展してきた一方で, その発展は言語の形式的な側面の評価に留まっており, 意味の側面を考慮する計測はほとんど行われていない. これには,Shannon の記述にもある様に, 情報理論では意味の側面は捨象されてきた歴史があるためであると言える.

> 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, 1948, p. 1)

3. 方法論

本研究が行った方法論は，以下の通りである[1].

(1). コーパスデータをサブワードに分割する.
(2). サブワードのトークンごとに，分散表現を得る.
(3). サブワードのタイプごとに，分散表現をクラスタリングする.
(4). 得られたクラスタリングの数，及び頻度分布からシャノンエントロピーを算出する.

第一に，コーパスデータをサブワードと呼ばれる単位に分割する．ここでいうサブワードとは，一定の計算アルゴリズムによって文字を結合して得られる，単語，または単語未満に相当する形式単位のことを指している．本研究では，HuggindFace 社の WordPiece[2]（図 1 参照）を用いた．WordPiece は，計算的なアルゴリズムによって得られる形式単位であり，多くは単語に相当するが，それ未満になる場合もある.

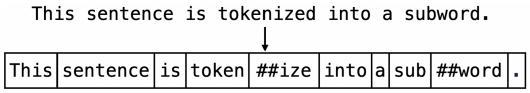

図 1: WordPiece によるサブワード分割

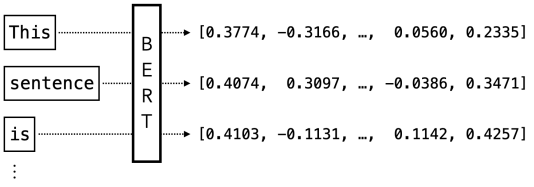

図 2: BERT による分散表現

第二に，サブワードのトークンごとに分散表現を得る．分散表現とは，単語の共起関係に基づいて形式列の意味を推定し，数値の集まりとして表現するものである．膨大なデータをコンピュータに学習させることで，共起環境が似たもの同士ほど近く，共起環境が異なるもの同士ほど遠く，という具合に数値群を割り当てることができる．これを元に，形式が表す意味を計算的な処理にかけることが可能になる．特に本研究では，意味の類似性の算出が可能になることを，応用する．この，分散表現を得る代表的な手法に，Mikolov et al. (2013a, 2013b) の Word2vec がある．この手法は，２層からなる単純なニューラルネットワークを用いて分散表現を得られるという点で演算資源が少なくて済む利点があるが，文脈を考慮せずに，同一の形式を同一の分散表現のみでしか表現できない.

そこで本研究は，ニューラル言語モデルである BERT (Devlin et al., 2019) の最終層を，そのサブワードの分散表現として用いた．BERT モデルは，その形式がどの様な共起関係で出現したのかを，文章中の共起語全てに対して総当たり的に関係性を計算し，それを考慮した数値群として意味を表現することができる（図 2 参照）．本研究の目指す語義数推定には，トークンごとの個別の意味を知る必要があり，BERT モデルは，文脈を考慮した分散表現を得なければならないという条件を満たす.

第三に，サブワードのタイプごとに，得られた分散表現をクラスタリングによって意味の数，及びその頻度分布を推定する．第二段階までの行程を複数の文章に対して行うことで，サブワードのタイプに対する各トークンの意味が，それぞれに文脈を考慮した分散表現として得られた状態になる（図 3 参照[3]）．ここで，タイプで束ねられた分散表現が，その類似性に基づいて分類された時，何個のクラスターに分類できるか，また各クラスターに幾つトークンが所属しているかを調べることで，そのサブワードの語義数を推定する．分散表現に対してクラスターを得る手法は多くあるが（e.g., k-means, etc.），本研究は，DBSCAN と呼ばれる手法を用いる．その他の手法と異なり，DBSCAN は予めクラスター数を想定する必要がなく，その点で本研究が目指す語義推定に適していると考えられる．DBSCAN は，得られた各データが多次元空間に埋め込まれていることと，「半径（epsilon）」と「最低サンプル数（minimum sample）」２つの変数を設定することが必要である．データセット内の各座標について，半径 ε 内の点を数え，最低サンプル数以上の点がある場合，その点を中心に近くの点を結びつけてクラスタを形成する（図 4 参照[4]）．反対に，

[1]本研究で用いたソースコードは,https://github.com/takuto-nakayama/subword-polysemy にて公開済みである.

[2]WordPiece のアルゴリズムは，文字単位で分割した状態から，より共起関係のスコアを計算しの，それが高い要素同士の結合を繰り返すものである．要素のペアとしての頻度を f_{pair}，各々単体での頻度を f_i, f_j とするとき，スコアは $\frac{f_{pair}}{f_i \times f_j}$ から得られる．また，サブワード中の##は，スロットを意味する.

[3]例えば図 3 における三つのデータの内，上と真ん中の例は，下の例に比べて数値が近いため同じクラスターに分類され，同じ語義で使用されていると判断する，という具合である.

[4]図 4 において，一番右にある点は，自身を中心とした半径 ε 内に他の点が無いため，この場合ノイズとして分類される．それ以外の点は自身を中心とする半径 ε の円に他の点が存在しており，これらがクラスターを成す.

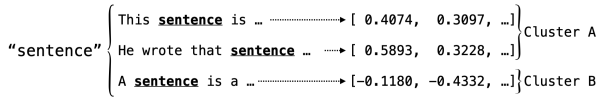

図 3: サブワードタイプのクラスタリング

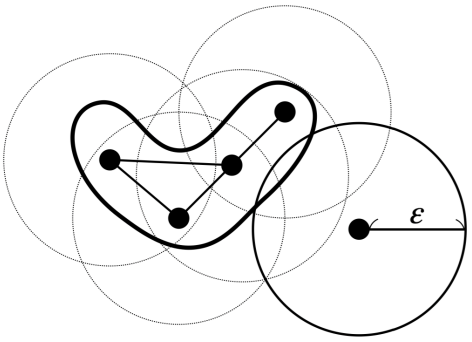

図 4: DBSCAN

どのクラスタに属さない点をノイズとして分類する．即ち，半径を大きくしていくと，ノイズが減りクラスターが形成されていくが，ある点を境にクラスター数が減り，最終的に全てのデータが単一のクラスターに分類されることになる．もちろんこの状態でのノイズは 0 個である．本研究では，最低サンプル数を 2 で固定し，半径はクラスター数が最大になる様，探索的に決定した[5]．

第四に，得られた頻度分布から，サブワードのタイプごとのシャノンエントロピー (Shannon, 1948) を計算する．エントロピーとは，予測不能性を数値化するもので，値が高ければ高いほど，発生するも事象を予測することが困難であることを示している．本研究においては，エントロピーが高いサブワードはそれだけ多くの意味と結合しているが，エントロピーが低いと，そのサブワードは単一の意味と結びついていると言える．第三段階で得られた，各タイプにおける，n 個のクラスターの中で，i 番目のクラスターの意味としてそのサブワードが使用される確率 p_i を算出し，以下の公式で，そのサブワードタイプのエントロピー H を計算する．最後に，全タイプのシャノンエントロピーの平均を求め，得られた各言語の平均エントロピーを言語間で比較する．

$$H = -\frac{1}{n}\sum_{i=1}^{n} p_i \log_2 p_i \tag{1}$$

本研究が使用したデータは，Wikipedia コーパスに存在する言語のうち，mBERT が対応している言語であるの中の，30 言語を対象とした．[6] 各言語の記事からランダムに 1,000 記事を抽出し，上記の手順で分析を行った．特に分散表現を得る際には，段落を与えられた文脈環境として利用し，その中におけるサブワードの分散表現を得た．

4. 結果

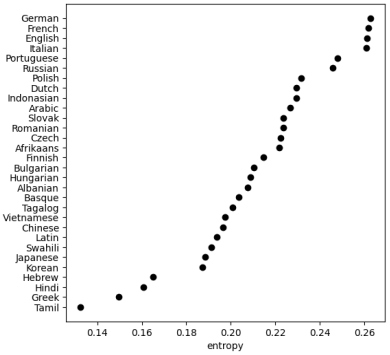

図 5: 各言語のサブワードエントロピー

図 5 は，縦軸に対象とした言語，横軸に計算されたエントロピーの結果が対応しており，下から順に，エントロピーの高い方へと配置されている．グラフを見ると，形式と意味の対応関係に関して考えると，全ての言語が非常に狭い領域に収まっていると言える．エントロピーは底が 2 の対数であるため，最小の値がタミル語の ≈ 0.132，最大の値がドイツ語の ≈ 0.263 であるため，それぞれサブワード当たり平均で 1.11 個と，1.20 個という計算になる．即ち，どの言語もサブワードの多義性という意味では，等しく平均化すれば大きく多義ではないということが示唆される．

[5] 具体的には，ε を徐々に大きくしながらクラスタリングを行い，クラスター数の推移を観察することで，最大クラスター数となる ε を決定した．

[6] 対象とした言語は以下の通りである．Afriaans, Albanian, Arabic, Basque, Bulgarian, Chinese, Czech, Dutch, English, Finnish, French, German, Greek, Hebrew, Hindi, Hungarian, Indonasian, Italian, Japanese, Korean, Latin, Polish, Portugese, Romanian, Russian, Slovak, Swahili, Tamil, Tagalog, Vietnamese.

一方で，その狭い範囲の中で考えると，言語間でその多義性に差異があると言える．値の上位を占めているのはヨーロッパ言語が主であり，アジア圏の言語の多くは小さな値に集中している．このことから，形式と意味の対応関係に関して，ある種の類型論的な差異が存在することも示唆される．

5. 結語

本研究は，言語の等複雑性というトピックに対して，ニューラル言語モデルの BERT とクラスタリング手法である DBSCAN といった自動処理的な手法と，情報理論的なアプローチのもと，形式と意味の対応関係に関する複雑性の比較を行った．結果として，対象とした全ての言語が，非常に狭い多義性の範囲に収まっており，形式と意味の対応関係に関して考えると，形式の多義性は言語を通して多様ではないということ，またその中でも差異が存在することが示唆された．

残る課題としては，対象とする言語数がまだ不足していることが挙げられる．対象とした言語の半数がヨーロッパ言語であるため，今後より多様な背景の言語を対象とすることを目指すべきである．また，DBSCAN におけるノイズの扱いも課題である．本研究ではノイズに分類されたデータは外れ値として除外したが，サブワードの中にはノイズの数が少なくないものもあった．このことから今後の研究では，ノイズとなったデータも考慮する分析が必要であると考えられる．

謝辞

本稿の本研究は JSPS 科研費 JP24KJ1938 の助成を受けたものである．

参考文献

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), 275. https://doi.org/10.3390/e19060275

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), 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). Association for Computational Linguistics. https://api.semanticscholar.org/CorpusID:52967399

Ehret, K., & Szmrecsanyi, B. (2016). An informationtheoretic approach to assess linguistic complexity. In R. Baechler & G. Seiler (Eds.), Complexity, isolation, and variation (pp. 71-94). De Gruyter. https://doi.org/10.1515/9783110348965-004

Fenk-Oczlon, G., & Fenk, A. (1985). The mean length of propositions is 7 plus minus 2 syllables-but the position of languages within this range is not accidental. Cognition, information processing, and motivation, 355-359.

Fenk-Oczlon, G., & Fenk, A. (2014). Complexity trade-offs do not prove the equal complexity hypothesis. Poznan Studies in Contemporary Linguistics, 50(2), 145-155. https://doi.org/10.1515/psicl-2014-0010

Hockett, C. F. (1958). A course in modern linguistics.

Juola, P. (1998). Measuring linguistic complexity: The morphological tier. Journal of Quantitative Linguistics, 5(3), 206-213. https://doi.org/10.1080/09296179808590128

Juola, P. (2008). Assessing linguistic complexity. In M. Miestamo, K. Sinnemäki, & F. Karlsson (Eds.), Language complexity: Typology, contact, change (pp. 89-108). https://doi.org/10.1075/slcs.94.07juo

Koplenig, A., Wolfer, S., & Meyer, P. (2023). A large quantitative analysis of written language challenges the idea that all languages are equally complex. Scientific Reports, 13(1), 15351. https://doi.org/10.1038/s41598-023-42327-3

Sapir, E. (1921). An introduction to the study of speech. Citeseer.

Shannon, C. E. (1948). A mathematical theory of communication. The Bell System Technical Journal, 27(3), 379-423.