2025/02/28

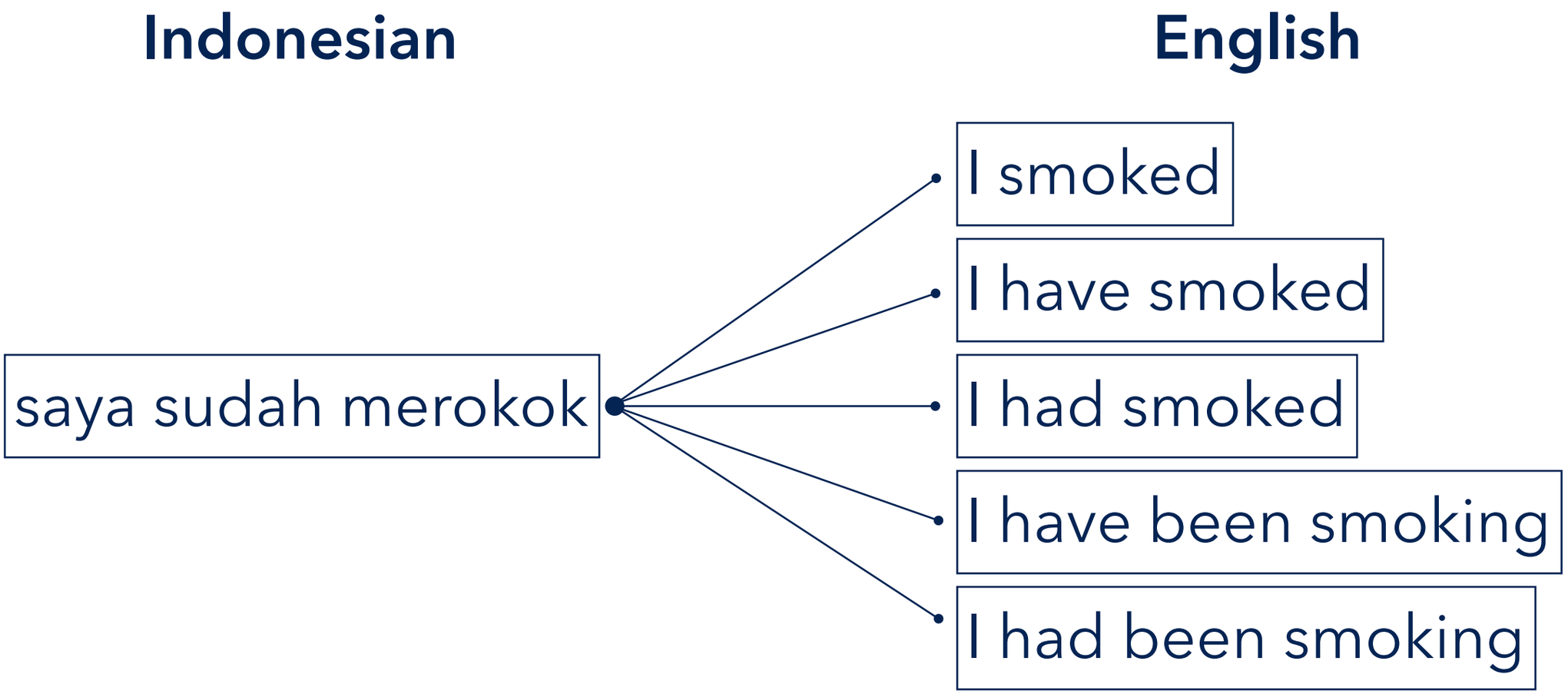
社会言語科学会第49回大会@慶應義塾大学三田キャンパス

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 / 日本学術振興会）

はじめに



- 言語の多義性(単義性) → 言語の複雑性と関係
- 言語の多義性の計測 → 意味の "計算" 技術が必要

発表のアウトライン

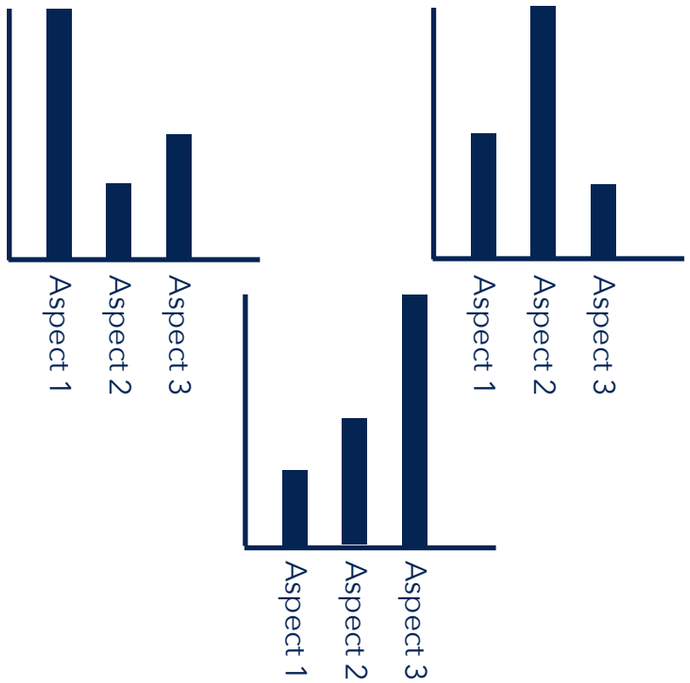
1. 研究の背景と先行研究
2. 方法論とデータ
3. 結果・考察
4. 結論と今後の課題

Slide

1.研究の背景と先行研究

研究の背景

All languages are equally complex!



- 屈折が複雑な言語→語順規則がない
- 語彙が多い言語→屈折規則が少ない

etc.

→個別の複雑さは異なっても、全体的にはそれぞれが釣り合いを取って、等しい値を取る様に見える

→「言語の等複雑性」、「等複雑性の言説」

研究の背景

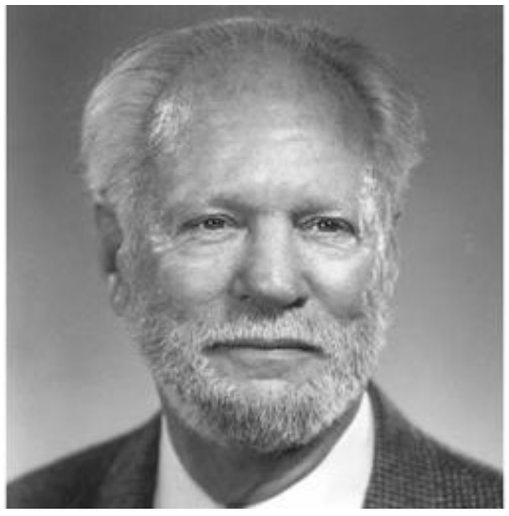
Die Sprache ist gleichsam die äusserliche Erscheinung des Geistes der Völker; ihre Sprache ist ihr Geist und ihr Geist ihre Sprache […] (Humboldt, 1836, 53)
(The language is the external appearance of the spirit of the people; their language is their spirit and their spirit is their language.)

Wilhelm von Humboldt

研究の背景

"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." (Sapir, 1921, 268–269)

Edward Sapir



Charles Hockett

"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."(Hockett, 1958, 180)

研究の背景

One of the most striking differences between H and L in the defining languages is in the grammatical structure: H has grammatical categories not present in L and has an inflectional system of nouns and verbs which is much reduced or totally absent in L (Ferguson, 1959, 333).

Charles Ferguson

コルモゴロフ複雑性による計測

Ehret and Szmrecsanyí (2016)

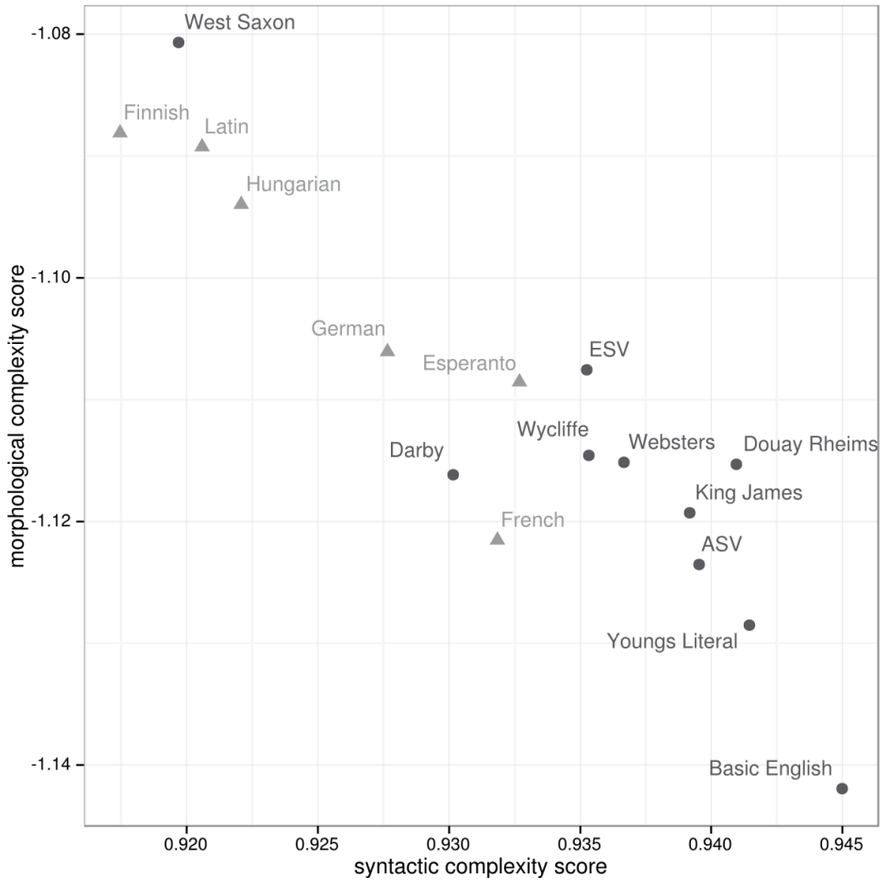
- 複雑さの定義:

  文書の最短アルゴリズム長(=コルモゴロフ複雑性)

  1. cdcdcdcdcd (10文字)→5*cd (4文字)
  2. cdgh39aby7 (10文字)→cdgh39aby7 (10文字)

  「より長いアルゴリズムが必要なら，より複雑でだ」という想定

- 方法論: 元データとzipファイルのバイト数差の大小比較
- データ: パラレル＆大規模な非パラレルデータ
- 結果: 形態素と統語間にトレードオフ関係



Gospel of Mark (79)

シャノンエントロピーによる計測

Bentz et al., 2017

- 学習のしやすさと表現の豊かさの関係を観察するため，単語の情報量(=エントロピー)を計測
- 語のエントロピーは一定の範囲に収まる
- 全体の語彙数と表現力のバランスを取る圧力が働いていることを示唆

Koplenig et al., 2023

- コーパス内で、言語ごとの複雑さを算出・比較
- 複数のコーパスで、同様の処理を行う
- あるコーパスにおける複雑性の差は、別のコーパスにも見られる

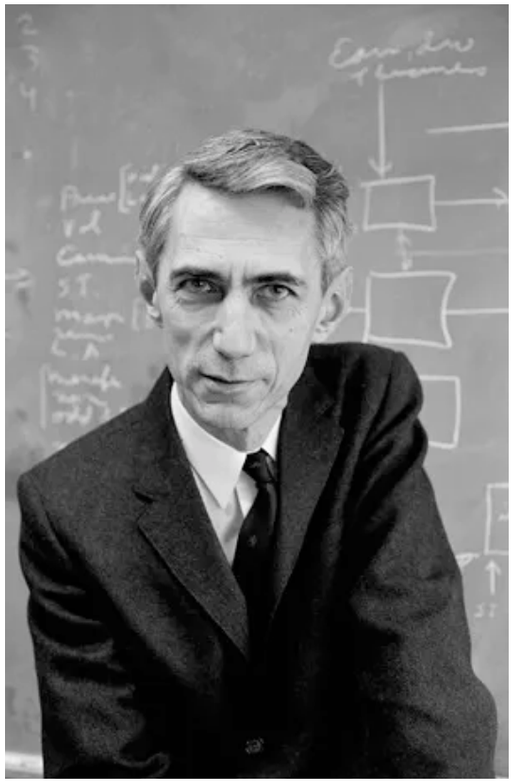
  =等複雑性を支持しない

先行研究における課題

意味の側面に焦点を当てた研究が無い

- 言語全体の複雑さを計測したいはずなのに，形式的な側面ばかりに焦点が当てられる.
- 意味を扱いながら，大規模なデータを処理することが難し(かった).

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, 1)



Claude Shannon

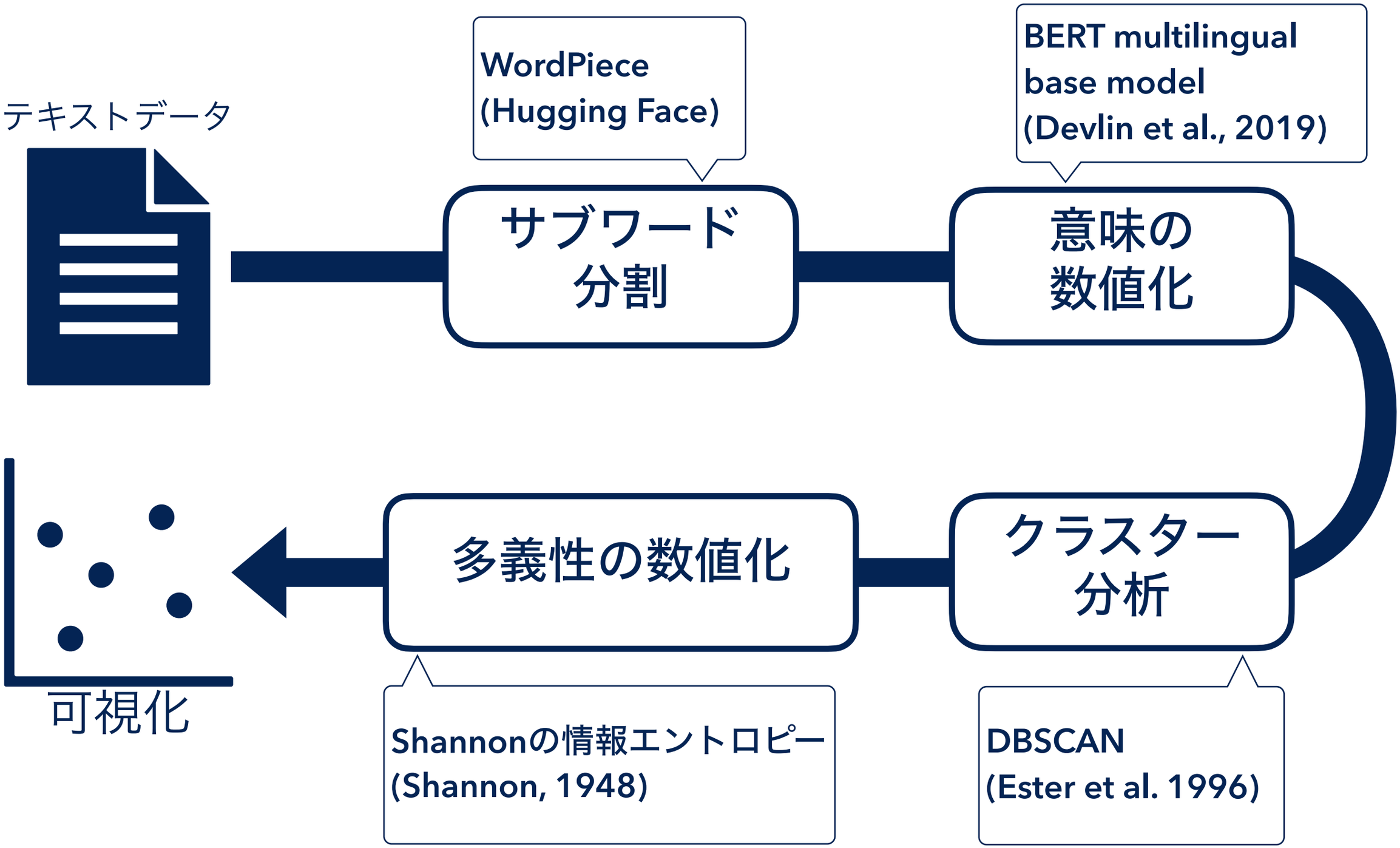
研究の目的

1. 形式–意味対応の多義性 (または単義性)を，言語の複雑さの一側面と見做し，それを言語間で比較する

2. ある形式に対応する語義数を自動処理によって推定する手法を提案する

2.方法論とデータ

プロセスの概観


テキストデータ
WordPiece
(Hugging Face)
サブワード
分割
BERT multilingual
base model
(Devlin et al., 2019)
意味の
数値化
クラスター
分析
多義性の数値化
可視化
Shannonの情報エントロピー
(Shannon, 1948)
DBSCAN
(Ester et al. 1996)

①サブワードを得る

“the two thin hounds found good foods on the big log” の分割 (HuggingFace, n.d. を参照)

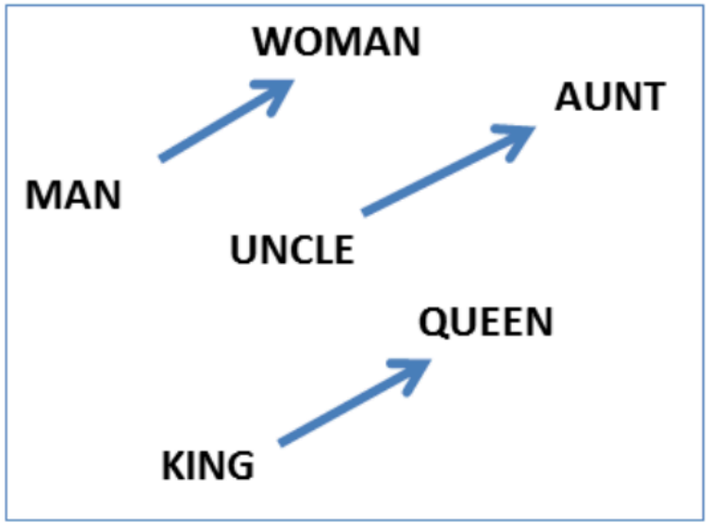
Merge	Corpus	Vocabulary
0	(t, ##h, ##e:2), (t, ##w, ##o:1), (t, ##h, ##i, ##n:1), (h, ##o, ##u, ##n, ##d, ##s:1), (f, ##o, ##u, ##n, ##d:1), (g, ##o, ##o, ##d:1), (f, ##o, ##o, ##d, ##s:1), (o, ##n:1), (b, ##i, ##g:1), (l, ##o, ##g:1)	##d, ##e, ##g, ##h, ##i, ##n, ##o, ##s, ##u, ##w, b, f, g, h, l, o, t
1	(t, ##h, ##e:2), (t, ##w, ##o:1), (t, ##h, ##i, ##n:1), (h, ##o, ##u, ##n, ##d, ##s:1), (f, ##o, ##u, ##n, ##d:1), (g, ##o, ##o, ##d:1), (f, ##o, ##o, ##d, ##s:1), (o, ##n:1), (**bi**, ##g:1), (l, ##o, ##g:1)	##d, ##e, ##g, ##h, ##i, ##n, ##o, ##s, ##u, ##w, b, **bi**, f, g, h, l, o, t
2	(t, **##he**:2), (t, ##w, ##o:1), (t, ##h, ##i, ##n:1), (h, ##o, ##u, ##n, ##d, ##s:1), (f, ##o, ##u, ##n, ##d:1), (g, ##o, ##o, ##d:1), (f, ##o, ##o, ##d, ##s:1), (o, ##n:1), (**bi**, ##g:1), (l, ##o, ##g:1)	##d, ##e, ##g, ##h, **##he**, ##i, ##n, ##o, ##s, ##u, ##w, b, **bi**, f, g, h, l, o, t

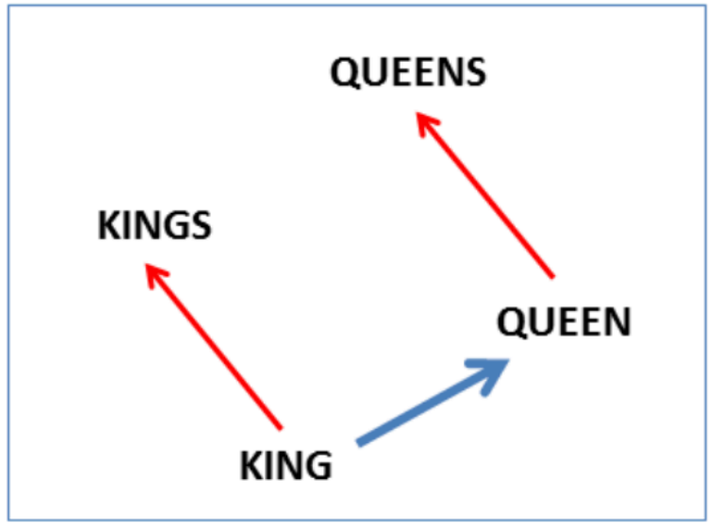
- WordPieceと呼ばれる単位に分割
- 1文字ずつに分割し，語中の文字にはそれを示す接頭辞(##)を追加する
- 以下のスコアを計算し，最も高いペアの結合を繰り返す

$$\text{Score} = \frac{\text{frequency}_{ij}}{\text{frequency}_i \times \text{frequency}_j}$$

- このアルゴリズムで得られた単位を，以降サブワードと呼ぶ

②埋め込み表現を得る





Mikolov et al. (2013, 749)

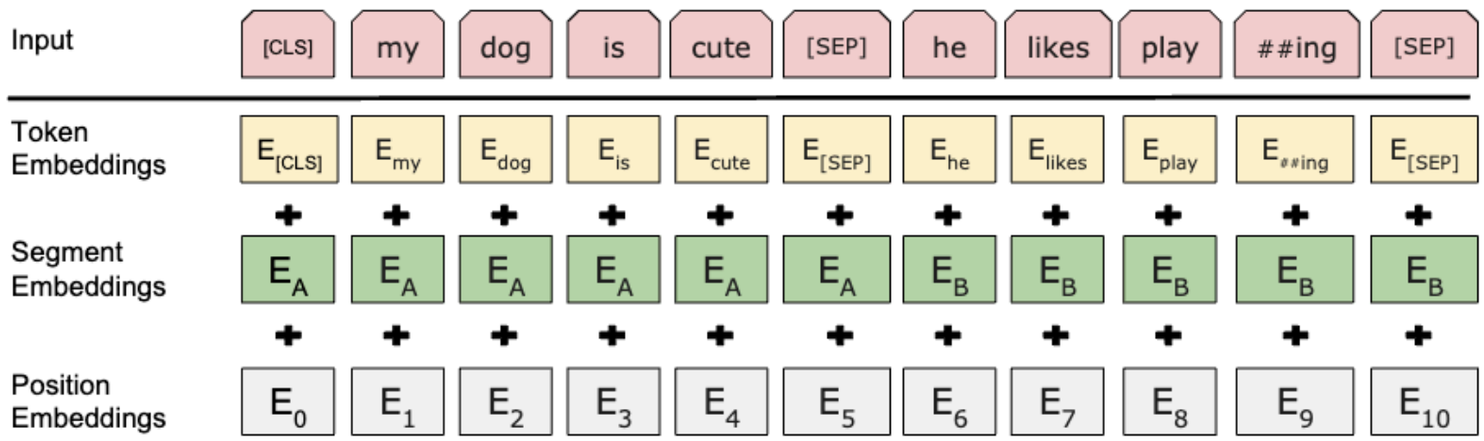
【左図】

- 埋め込み表現
  ＝意味を**方向**と**長さ**の情報に変換したもの
- (例) [男性→女性]という意味
- 青い矢印は，同じ長さと方向を持つため，同じ意味を表している

【右図】

- 各語彙間で，意味の演算が成立するように，長さと方向の情報へ変換を行う
- (例) 赤い矢印は，[単数→複数]という意味
- KING + [男性→女性] + [単数→複数] = QUEENS

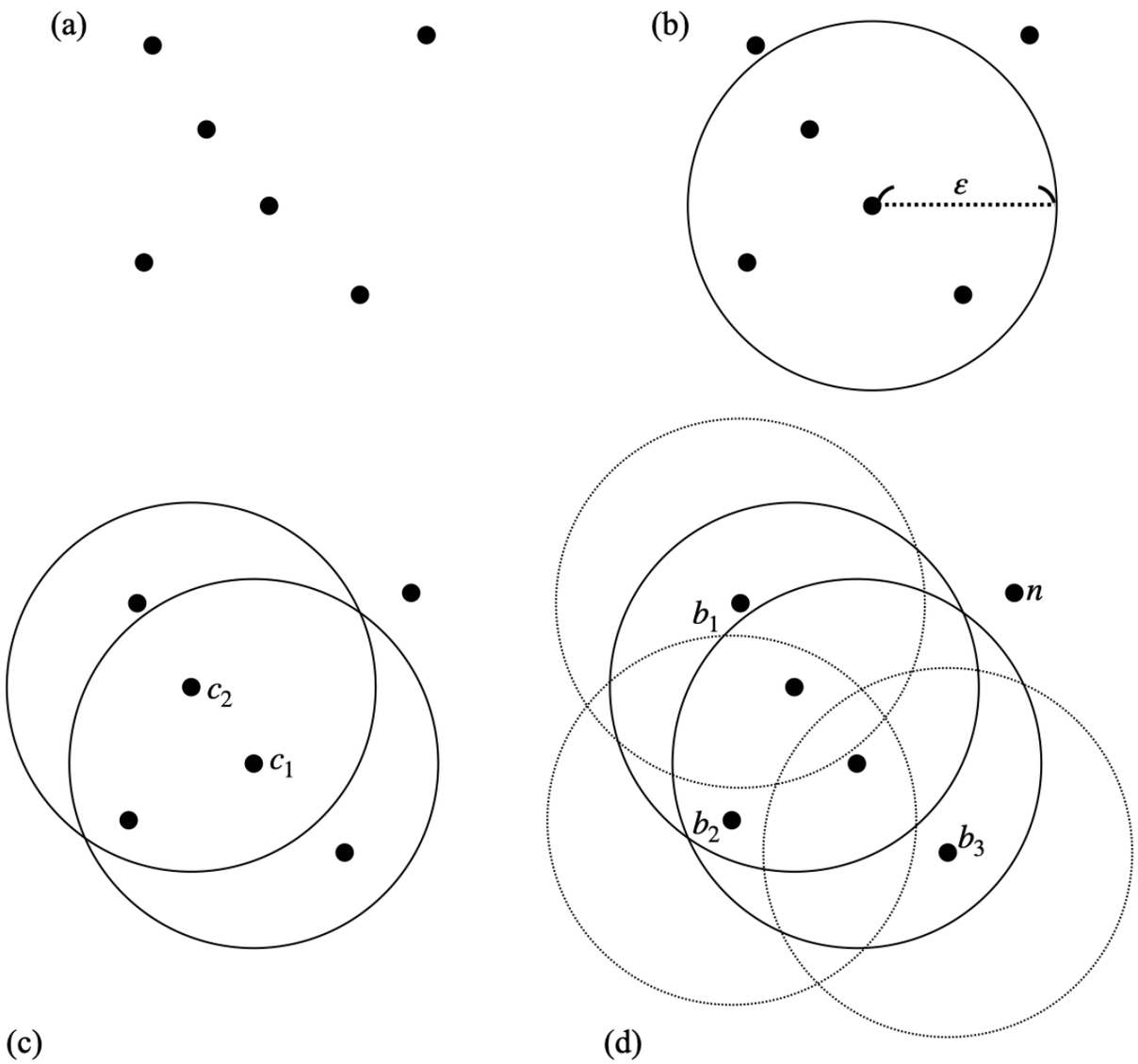
②埋め込み表現を得る



Devlin et al. (2019, 6)

- BERTは，「文中の各語彙との関連の強さ」も加味してトークンごとに埋め込み表現(意味の長さと方向)を得る
- 語彙が持つ文脈による意味の違いを，自動処理にかけることができる
- 分析には，事前学習モデルのBERT multilingual base model (Devlin et al., 2019)を用いた
- 分析する埋め込み表現は，最終層の768次元を，tSNE (Maaten and Hinton, 2008)で次元圧縮したもの

③語義数を推定する



Ester et al. (1996, 228)

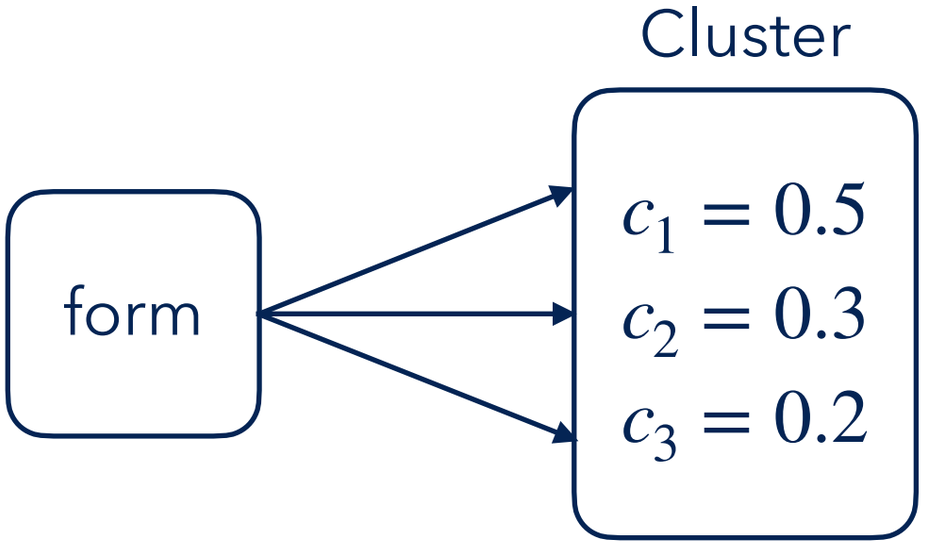
- DBSCANは，「各データから半径εの円内に何個のデータが含まれるか」を元に，探索的にクラスターを決定する手法
- 予め設定した数を超えたデータ数を含むデータ(e.g., (c)のc_1, c_2)，及びそれらの円に含まれるデータ(e.g., (d)のb_1, b_2, b_3)が同一クラスターに分類される
- それに該当しないデータ(e.g., (d)のn)は外れ値として，クラスターに属さない

④エントロピーの計算をする

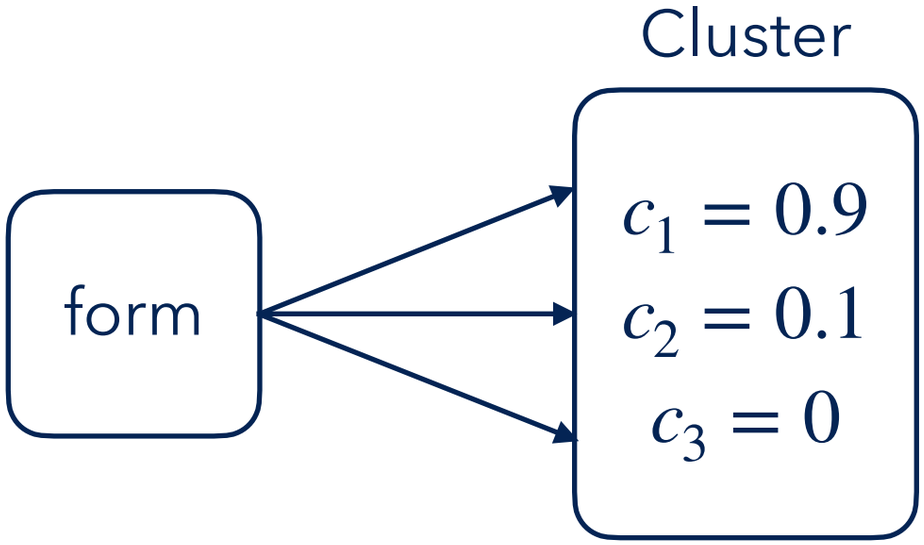
n個の事象のうちi番目の事象が起こる確率がp_iである時,

シャノンエントロピーHは以下で定義される (Shannon, 1948):

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

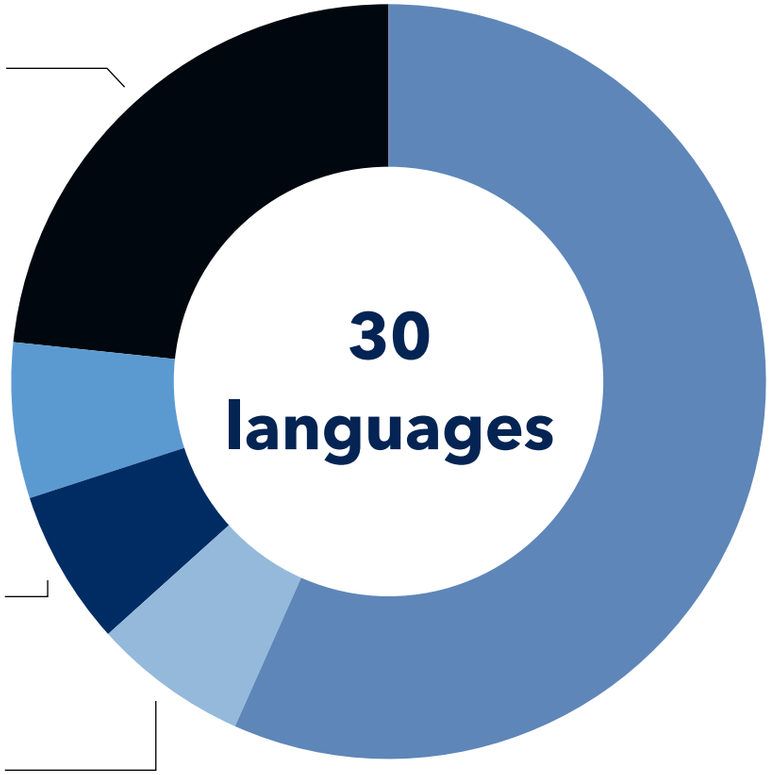


$H \approx 1.4854$



$H \approx 0.4689$

データ



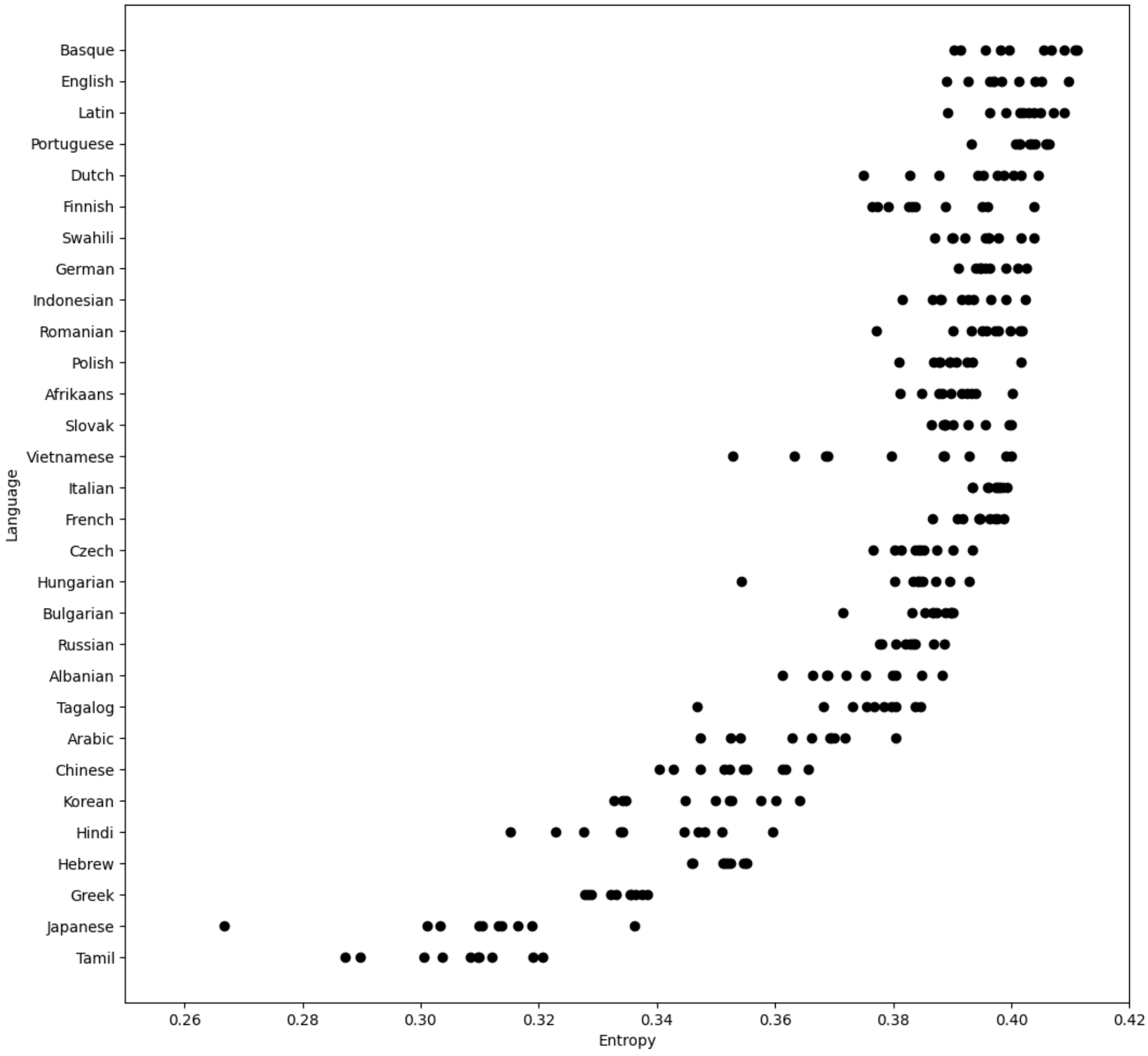
Dryer and Haspelmath (2013) 参照

- ランダムに抽出したWikipediaの100記事を1試行分のデータとして利用し，計10回ずつの試行を行った
- 対象とする言語は30言語

【対象言語一覧】

Afrikaans, Albanian, Arabic, Basque, Bulgarian, Chinese, Czech, Dutch, English, Finnish, French, German, Greek, Hebrew, Hindi, Hungarian, Indonesian, Italian, Japanese, Korean, Latin, Polish, Portuguese, Romanian, Russian, Slovak, Swahili, Tagalog, Tamil, Vietnamese

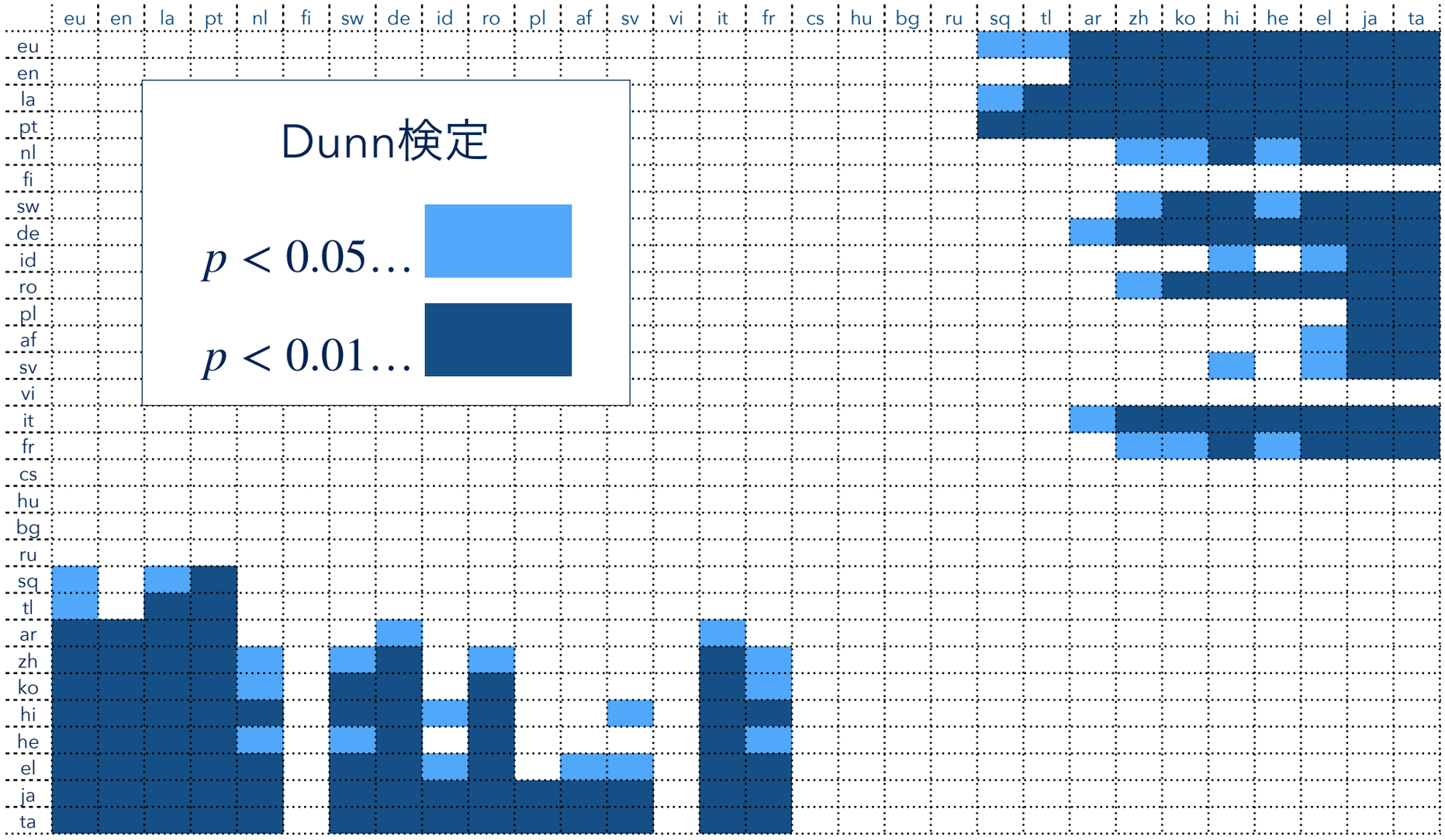
3.結果・考察



- 最大値を示したのがバスク語 (≈ 0.411),
  最小値を示したのが日本語 (≈ 0.266)
- それぞれ, ≈ 1.329/sw, ≈ 1.202/sw
- 比較的狭い範囲に全ての言語が収まっている

Kruskal–Wallis 検定	
統計量	p値
250.371	4.317E-37**

- Kruskal–Wallis 検定により，有意差あり
- Dunn検定により，エントロピーの高い言語と低い言語との間で，有意差が認められた
- 印欧語がエントロピーの高いものとして分類される傾向がある



4.結論と今後の課題

結論と今後の課題

【結論】

- 対象とした30言語の多義性は，概ね1.202/sw~1.329/swという狭い範囲に収まっている一方で，多義性には高い言語と低い言語とで有意差がある
- 印欧語の多義性が，より高い傾向があることが示唆される

【今後の課題】

- 対象とする言語がまだ少ない
- WordPieceアルゴリズムは，スペースによる分ち書きをベースにしているため，そのシステムを持たない言語への適用妥当性の検証が必要である

参考文献・謝辞

【謝辞】

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. (2018). B*ERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. arXiv [cs.CL]. http://arxiv.org/abs/1810.04805

Dryer, M. S., & Haspelmath, M. (Eds.). (2013). *Wals online (v2020.4)* [Data set]. Zenodo. https://doi.org/10.5281/zenodo.13950591

Ehret, K., & Szmrecsanyi, B. (2016). An informationtheoretic approach to assess linguistic complexity. In R. Baechler & G. Seiler (Eds.), *Complexity, isolation, and variation* (Vol. 57, pp. 71–94). De Gruyter.

Ester, M., Kriegel, H., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, 226–231

Ferguson, C. (1959). Digrossia. *Word*, *15*, 325–340.

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

HuggingFace. (n.d.). WordPiece tokenization. https://huggingface.co/learn/nlp-course/en/ chapter6/6

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.

Mikolov, T., Yih, W.-T., & Zweig, G. (2013). Linguistic regularities in continuous space word representations. aclanthology.org. https://aclanthology.org/N13-1090.pdf

van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of machine* learning research, 9(86), 2579–2605.

von Humboldt, W. F. (1836). *Über die verschiedenheit des menschlichen sprachbaues und ihren einfluss auf die geistige entwickelung des menschengeschlechts* [*on the diversity of human language structure and its influence on the intellectual development of humankind*]. S. Calvary.

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. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

ご清聴ありがとうございました

GitHub

Slide