豊かな記憶が創る複数形の諸相

皆川 翔
慶應義塾大学大学大学院
tsunx2.pg8.517@keio.jp

概要

英語の複数形は，英語学習者がその初期に学ぶ基本的な文法事項の１つである．しかし単数形に形態素‘s’をつけるという操作は後付けの文法的な説明であり，それは必ずしも英語話者がどのように複数形を運用しているのかという説明にはならない．今研究は大規模な言語データを用いて，複数形という文法を創りあげる背後に後述する「豊かな記憶」が複雑なネットワークを構築していることを単数形と複数形の使用頻度差と類似度の観点から検証し，複数形の習得と運用のプロセスを具体的に可視化することを目的としている．

1 はじめに

我々は言語を習得する際，言葉そのものだけでなく，その言葉が使用された文脈，環境，五感などあらゆる豊かな情報を記憶（rich memory: [1, 2]）に蓄え，新しい言語情報と照合を繰り返している．事例基盤モデル（exemplar-based model: [1, 2, 3, 4, 5]）は，出会った豊かな記憶を全て一度記憶し，過去の事例と新しい事例を照合する言語モデルを築いている．

当然，事例基盤モデルにおいて使用頻度が高い言語事例は記憶の前面に出てきやすいし，逆に頻度が低いものは照合に困難を強いることもある．こういった頻度の問題は言語習得，運用を考える上で欠かせない要素であろう．しかし頻度だけでは説明できない事例も存在し，特にそれは複数形の習得と運用に顕著に現れる．頻度だけでなく，豊かな記憶が複数形を構築していくのである．

1.1 事例基盤モデルと複数形

これまで言語学において事例基盤モデルは用法基盤モデル（usage-based model）を具体的に実装したものとして捉えられてきた[6]．というのも，用法基盤モデルは概念的な整理は行っているが，カテゴリー判断のプロセスがどのように実行されているとは言い難いからである[7]．用法基盤モデルと事例基盤モデルの大きな違いは，言語情報をスキーマに見出すか，個々の事例に見出すかである．スキーマは個々の事例の集積であるが，ある程度の抽象化を許している．多方事例基盤モデルは個々の事例に着眼しているため，抽象化の影響を受けないより微細な言語情報の変化に敏感なモデルであると言える．さらに個々の事例は，テキスト情報にとどまらず，その事例が使用された環境や五感といった豊かな記憶が埋め込まれている．複数形の習得と運用について議論するためには，単数形と複数形の微細な変化を捉えないと説明できないことが出てくるのである．

1.2 事例基盤モデルと頻度効果

言語の習得と運用において頻度効果は重要な要素である．頻度効果とは言語処理や意味理解において，よく使われる単語や表現が，より迅速かつ効率的に処理されるという現象である．Tomasello も子どもの言語習得を考える上でもこの頻度効果は欠かせないと述べている[8]．

しかし，用法基盤モデルが広く扱われるようになるにつれて，この頻度効果は絶対的な地位を築いてしまったように思える．言語とその使用を考えるとき，頻度という結果に落ち着くのは必然であるのだが，良くも悪くもその流れの中で頻度絶対視といった動きが見て取れる．

複数形の習得について考えるとき，この単純頻度効果だけでは説明できない事例が出てくる．頻度効果と複数形習得の関係性について探るために重要なのは，単数形との頻度差である．頻度効果に従うのなら，複数形と単数形の頻度差が大きくなれば２つの使い分けは容易であるということになりそうだ．datum/data といったペアは圧倒的に data を使用することから頻度効果で説明できそうである．しかし，単数形と複数形の使用が小さいものも存在する．たとえば car と cars は単数形と複数形に使用差がないにもかかわらず，ある程度英語を学習した人なら難なく使い分けることができる．これは頻度効果では説明し難い現象であろう．

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

頻度効果の背後には言語をその使用と結び付ける用法基盤モデルの発展があった．スキーマベースの理論と頻度は相性が非常に良い．用法基盤モデルを具体実装する事例基盤モデルが台頭してきた今日，この頻度効果もそれ相応にアップデートする必要があるのではないだろうか．事例基盤モデルでは豊かな記憶を事例が持ち，事例ベースで微細な違いにも敏感になることができるモデルだ．この理論を汲むのであれば，頻度は認知補助装置であり言語の実態ではないことを改めて考える必要があるだろう．

1.3 分布意味論から Word2Vec

事例基盤モデルと相性が良い理論の一つに分布意味論がある．分布意味論は，単語の意味がその使用される文脈によって形成されるという理論で, HarrisやFirthによって提案された[9, 10]．この理論は，Word2Vec[11, 12]のような密な単語ベクトル表現を学習することで，単語間の意味的な関係を捉える言語処理モデルの基礎となっている．

Word2Vecは, ニューラルネットワークを用いて単語を数値ベクトルで表現する技術で，単語間の意味的な関連性や類似性を数値的に捉えることができる．Word2Vecには，CBOW（Continuous Bag of Words）モデルとSkip-gramモデルの2つの形式が存在する．CBOWモデルは周囲の単語からターゲット単語を予測するニューラルネットワークで，Skip-gramモデルはその逆の操作を行うニューラルネットワークだ（図1）．本研究では，事例基盤モデルと相性の良い分布意味論を基にしたWord2Vecを用いて単数形と複数形の単語の意味ベクトルを求め，その差異を生じさせる要因を検討する．

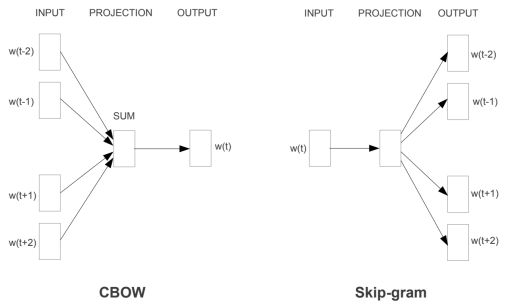

図 1 Word2Vec [11]

2 方法論

Word2Vec を使用するにあたって，そのベクトル計算のもととなるコーパスを指定する必要がある．今研究では少しでも多くの情報をもとに計算を行うため，Wikipedia の dump データ[i]をもとに Word2Vec のモデルを構築した．Word2Vec のモデルを学習させる際, ウィンドウと次元数を指定する必要がある．ウィンドウとは，意味ベクトルの計算を行うために使用する前後の文脈の指定であり，次元数とは意味ベクトルを計算するために，どの程度高次元な座標を用いるかという意味探索の詳細さである．今研究ではウィンドウ数を前後 5 単語に指定し，次元数は 500 次元[ii]で行った．

次に単数形と複数形の頻度差と 2 つの類似度の関係を探るため，コーパス内の全名詞の単数形と複数形の使用頻度差とそのペアのコサイン類似度を計算した．この結果によって頻度差がどのような影響を与えているのか全体像を掴む目的がある．

最後に複数形の意味ベクトルから単数形の意味ベクトルの引き算を行い，その差分の意味ベクトルに類似する単語を上位 5 例抽出した．この操作を行うことで，複数形にあって単数形にない要素を可視化することができる[iii]．これによって複数形を運用する背後にはどのような知識ネットワークがあるのかを考察していく．

3 結果と考察

3.1 単数形と複数形の頻度差と類似度

まず，単数形と複数形の使用頻度差と類似度の関係性について求めた．なお使用頻度差は，絶対的な頻度差ではなく，相対的な頻度差（F）を

$$F = \frac{|singular - plural|}{\max(singular, plural)}$$

で求めた．というのも，たとえば単数形の頻度 50，複数形の頻度 60 の場合と，単数形の頻度 5000，複数形の頻度 5010 の場合の 10 の差の重みが異なるからである．以上をもとに単数形と複数形の使用頻度差を縦軸に，コサイン類似度を横軸にプロットしたものが図 2 である．

[i] https://dumps.wikimedia.org/enwiki/
[ii] 100 次元，300 次元，500 次元，1000 次元のモデルを学習させた結果, Wikipedia のデータでは 500 次元から意味表現が飽和したため 500 次元に設定した．
[iii] 逆に単数形から複数形の意味ベクトルを引くと単数形にあって複数形にない要素を可視化できる．今回はその両方を行った．

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

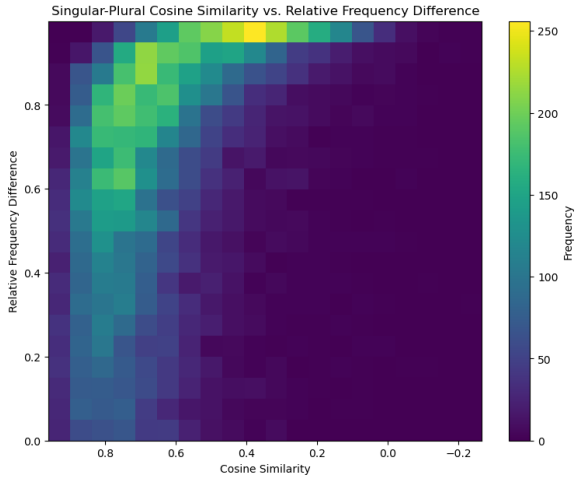

図 2 単数形と複数形の使用頻度差と類似度の関係

この結果を見ると，単数形と複数形の使用頻度とその類似度の関係パターンは３つしか存在しないことが分かる．

1. 頻度差が大きくて，類似度が高い
2. 頻度差が大きくて，類似度が低い
3. 頻度差が小さくて，類似度が高い

この結果で最も顕著なのは，単数形と複数形の使用頻度差が小さいものは類似度が低くなることがほとんどないということだ．これは頻度効果の観点から見れば当然の結果だろう．重要なのはその結果は必ずしも言語使用の現実を映し出しているとは言えないということだ．頻度効果に従えば，特に 3 のように頻度差も小さく，類似度が高い場合は単数形と複数形の使い分けが難しいと言えるかもしれない．しかしながら，その頻度差と類似度にかかわらず，我々は単数形と複数形を明確に使い分けている．以上の結果をもとに，なぜ 3（もしくは 1）の場合でも単数形と複数形を使い分けられているのかについて考察していく．

3.2 複数形−単数形

次に複数形から単数形の意味ベクトルの引き算を行い，その差分ベクトルに類似する単語を上位 5 例抽出した．補足として，抽出した上位 5 例全てと差分ベクトルのコサイン類似度が 0.8 以上の事例に限定して検証した．というのも，差分ベクトルに類似する単語のコサイン類似度が低いと，その単語がそれぞれの差分を表しているとは言い難いからである．以上を踏まえ抽出した結果が表 1 である．

まず複数形から単数形を引いた結果の中の insects から insect を引いたベクトルに類似する単語を抽出すると，birds，invertebrates，wasps，moths，foraging とある．興味深いのは，その差分に類似するものとして虫以外を表す単語も入っているという点であろう．複数形にするだけでも単純に insect の数が増えるのではなく，鳥や無脊椎動物といった虫の天敵となりうる生物や foraging という虫が複数いる狩猟採集という環境を想起するということがわかる．insect と insects の相対的な頻度差は約 0.56 であり単複には差があるともないとも言えないが，類似度に関しては約 0.91 と非常に高い．頻度差は多少あるが，類似度が高いペアを使いこなせる背景には，このような想起されるイメージの複雑な変化があるのだろう．同様に forests から forest を引いた差分ベクトルも生態系や生息地，熱帯雨林，マングローブの数が増えるだけでなく， arid といった形容詞があることから複数形にすることで乾燥したイメージが想起されることが考えられる．

また単複の相対的な頻度差が 0.13 と小さく，類似度もそれ相応に約 0.83 と高い lizards と lizard の差分ベクトルに類似する単語には rodents，carnivores が抽出されている．rodents はトカゲの餌としてネズミなどを想定していると考えられ，複数形にすることで，それ相応に rodents の数も増え，結果的に carnivores としてのトカゲの一面が強調されているのだろう．さらに bonobo も出現していることから単数形 lizard と複数形 lizards には，想起される生態系の変化が起こっていると考えられる．

次に muffins から muffin を引いた差分ベクトルに類似する単語を抽出すると，プリンや調味料，トルティーヤなどの単に数が増えているだけでなく，flavored という形容詞も見て取れることから muffins と muffin で見た目だけでなく風味も変わっていることが示唆される．単数形と複数形の相対的な頻度差は，約 0.27 と小さく，類似度は約 0.72 と比較的高いがその 2 つを使い分けられている背景にはこのような数の変化に加え風味の変化といった，味覚や嗅覚に関するイメージの変化が関わっていることが考えられる．

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

表 1 複数形 - 単数形

Plural	Singular	Frequency	Similarity	word1	word2	word3	word4	word5
insects	insect	0.56467	0.908260345	birds	invertebrates	wasps	moths	foraging
forests	forest	0.558585	0.852162004	ecosystems	habitats	rainforests	mangroves	arid
muffins	muffin	0.272727	0.716795802	flavored	puddings	flavoring	condiment	tortillas
lizards	lizard	0.134122	0.828244209	parasites	rodents	bonobos	primates	carnivores

表 2 単数形 - 複数形

Singular	Plural	Frequency	Similarity	word1	word2	word3	word4	word5
opposition	oppositions	0.991688	0.340409338	unionists	Tories	supporters	unpopularity	moderates
sorrow	sorrows	0.678474	0.756948233	grief	shame	sadness	despair	misery
wit	wits	0.795635	0.607918203	subtlety	witty	humor	elegance	sensibility
pandemic	pandemics	0.97392	0.607026637	covid-19	lockdown	lockdowns	covid	coronavirus

3.3 単数形-複数形

次に表 2 が表すように，単数形から複数形を引く操作をすることで，単数形にはあって複数形にはない要素を抽出していき，その背後にある知識ネットワークを明らかにする．この操作では主に抽象名詞を抽出することができた．

まず opposition から oppositions を引いた差分に類似する単語に supporters, moderates とある．これは複数形には supporters，moderates の要素が単数形に比べると弱いということになる．つまり，複数形を運用する背景に，支持者の数的な増加や大胆さという感覚情報が知識として加わっているということだ．結果的に複数形にすることで，単数形の意味が強まっていることを表しているのであろう．

しかし，必ずしも複数形にすることで単数形の要素が強まるとは限らない．逆に単数形の方が元の意味をとどめ，複数形にすることで元の意味から逸れてしまう例も存在する．たとえば，sorrow から sorrows を引いた差分には grief，shame，sadness，despair，misery が抽出されている．これは単数形には悲しみや恥，絶望や惨めさがあるが複数形にはないことを示している．他にも wit から wits を引いた差分には subtlety，witty，humor，elegance，sensibility が抽出されている．これらは単数形の方が本来の意味をとどめている事例であると言えるだろう．

pandemic から pandemics を引いた差分ベクトルに類似する単語は時流を捉えている．差分の単語には，covid-19，lockdown，lockdowns，covid，coronavirus が現れている．これは新型コロナウイルスの影響を受けその文脈で pandemic という単数形がより多く使われたためであろう．これは単数形にして記述することに意義が見出されている例であり，その意味で特定の環境下における単数形の使用が複数形の使用との区別を容易にしていると言える．

6 結語

複数形はただ単に数の概念をもとに構築される計量的な操作ではなく，その背後に「豊かな記憶」が存在し，単数形と複数形のそれぞれのイメージや五感の変化がその理解や運用を可能にしている．我々はその豊かな記憶を一度全て記憶することで単数形と複数形に使用頻度差がなく，類似度が高い微々たる変化が起きている状況下でも使い分けが可能になるのである．かといってこれは頻度を軽視するものではない．事実，使用頻度差によって単数形と複数形が使い分けられている例も多く存在するのだ．

しかし，頻度そのものは必ずしも言語を運用する知識の直接的な説明にはならない．重要なのは，頻度を考慮に入れた上でその使用の背後にある複雑な知識ネットワークを具体的に解明することであろう．そういったハイブリッドな言語モデルとして事例基盤モデルはますます現実味を帯びてくる．今研究が示したのは，複数形の理解と運用という全文法事象の 1 つにすぎないが，その背後にある「豊かな記憶」は複数形に限らず全言語現象の背後にある知識ネットワークの姿であると筆者は考えている．

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

参考文献

[1] Port, R. 2007. How words are stored in memory: Beyond phones and phonemes. *New Ideas in Psychology*, *25*, 143-170.

[2] Port, R. 2010. Rich memory and distributed phonology. *Language Sciences, 32*, 43-55.

[3] Bod, R. 2006. Exemplar-based syntax: How to get productivity from examples. *The Linguistic Review, 23*, 291-320.

[4] Bod, R. 2009. From exemplar to grammar: A probabilistic analogy-based model of language learning. *Cognitive Science, 33*, 752-793.

[5] Wedel, B. 2006. Exemplar models, evolution, and language change. *The Linguistic Review*, *23*, 247-274.

[6] Langacker, R. 2009. A dynamic view of usage and language acquisition. *Cognitive Linguistics, 23*, 627–640.

[7] 吉川正人. 2010. 「用法基盤」から「事例基盤」へ:妥当な言語記憶のモデルを求めて. 言語処理学会第16回年次大会発表論文集, 962−965.

[8] Tomasello, M. 2003. *Constructing a language: A usage-based theory of language acquisition*. Harvard University Press.

[9] Firth, J. R. 1957. A synopsis of linguistic theory 1930-1955. Studies in linguistic analysis.

[10] Harris, Z. S. 1954. Distributional Structure. *Word*, *10*(2-3), 146-162.

[11] Mikolov, T., Chen, K., Corrado, G., & Dean, J. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

[12] Mikolov, T., Yih, W., & Zweig, G. 2013. Linguistic Regularities in Continuous Space Word Representations. In *HLTNAACL Vol. 13*, 746-751.

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).