動詞派生前置詞の文法化の定量化

永田亮 [1] 川崎義史 [2] 大谷直輝 [3] 高村大也 [4]
[1] 甲南大学 [2] 東京大学 [3] 東京外語国語大学 [4] 産業総合技術研究所
nagata-dprep @ ml.hyogo-u.ac.jp.

概要

動詞派生前置詞とは，動詞から派生し，前置詞のように振る舞う語句（例：following）を指す．動詞派生前置詞が，動詞の性質を失い，前置詞の機能を果たすようになる文法化現象について様々な仮説が言語学で知られている．しかしながら，仮説検証の規模について改善が必要なことが指摘されている．本稿では，動詞派生前置詞の文法化度を自動的に定量化する手法を提案し，大規模なコーパスを対象にして，文法化に関する三つの仮説の検証を行った．その結果，提案手法は全ての仮説を支持する結果となった．

1 はじめに

本稿では，動詞派生前置詞の文法化に関する仮説を計算言語学的なアプローチで検証することを試みる．動詞派生前置詞とは，動詞から派生し，前置詞のように振る舞う語句を指す（例：*considering*, *according to*）．また，文法化とは，内容語が文法的な機能を果たすようになる変化のことを指す [1]．

2 節で詳細を述べるように，言語学では，動詞派生前置詞の文法化に関する様々な仮説が提案されている．主な仮説は次の 3 点に要約される：

1. 動詞派生前置詞は完全には文法化しておらず，動詞と前置詞の中間カテゴリに位置する [2, 3]
2. 文法化の度合いは連続的であり，動詞派生前置詞により異なる [2, 3, 4]
3. 文法化の度合いは，前置詞性と動詞性という二つの極性に基づいて判定可能である [3, 4]

言語学において動詞派生前置詞に関する豊富な知見があるものの，Hayashi [3] が指摘するように，分析対象の種類と事例数という点で従来研究には改善の余地がある．より頑健な理論の構築のためには，多様な事例に対して仮説の検証を行うことが重要であるが，従来研究は事例分析や内省に基づくため，分析の規模を大きくすることは容易でない．関連して，従来の文法化度の判定には一部矛盾するところもある．例えば，文献 [2] と文献 [3] では，*following* に対して正反対の結論を下している．

本稿では，より頑健な理論の構築を目指して，動詞派生前置詞の文法化度を定量化する計算言語学的な手法を提案する．まず，複数の手法を吟味し，言語学の知見と最も整合する手法を明らかにする．その手法を用いて，上述の三つの仮説の検証を行う．更に，従来研究にみられる文法化度の不一致の解消を試みる．最後に，文法化度の可視化を行い通時的変化について分析を行う．

2 関連研究

言語学では，動詞派生前置詞の文法化に関する研究が盛んに行われている（例えば，文献 [2, 3, 4, 5, 6, 7, 8] など）．多くの研究では，質問紙法またはコーパス中の事例に対する言語学的なテストにより文法化度を定量化する．両アプローチとも，動詞派生前置詞が at や on などの典型的な前置詞の特性を満たすかどうかの判定に基づき文法化度を決定する（一部の研究では動詞の特性も考慮する）．例えば，Kortmann ら [2] は，言語学的なテストを用いて，19 種類の動詞派生前置詞を 5 段階の文法化度で分類している．また，Hayashi [3] は，前置詞の特性を満たすかどうかを問う二種類の質問紙法を用いて，37 種類の動詞派生前置詞に文法化度（0〜10 の連続値）を割り当てている．参考として，両研究における文法化度の判定結果を付録 A に示す．

このような研究により，1 節で述べた三つの仮説の提案と検証が進められているが，依然，十分でないことを Hayashi [3] が指摘している．特に，分析の規模（動詞派生前置詞の種類と事例数）を大きくすることが求められている．理想的には，大規模なコーパス中の多様な事例に対して検証を行うことが望ましい．本稿は，計算言語学的なアプローチを用いることで，そのような検証を実現する．

3　文法化度の定量化手法

本稿では，文脈なし単語ベクトル（具体的には，word2vec [9]）に基づいた手法と文脈付き単語ベクトル（BERT [10] の最終層のベクトル）に基づいた手法を提案する．両手法に共通する手順として，動詞派生前置詞の認識がある．構文解析器（spaCy[1)]）を用いて，次の二つの条件を満たす単語を動詞派生前置詞とした：(1) 表層形が付録 A の表 3 または表 4 に示す動詞派生前置詞と一致（大文字小文字の違いは無視する）；(2) 依存構造ラベルが，*prep* (前置詞) または *advcl* (副詞節)．以降の処理用に，入力コーパス中で，この条件を満たす語句にタグを付与する．

3.1　文脈なし単語ベクトルに基づいた手法

2 節で述べたように，従来研究では，対象の動詞派生前置詞が，前置詞の特性を持つかどうかの判定により文法化度を決定する．本節の手法もこの考え方を踏襲するが，より直接的に定量化を行う．具体的には，動詞派生前置詞と前置詞との間の類似度を文法化度とする．更に，Hayashi [3] に従い，動詞との類似度も考慮する．

ここで，類似度をどのように定義するかが問題となる．本稿では，単語ベクトル間の余弦類似度を利用した，次の 3 種類を提案する．

平均前置詞：前置詞の単語ベクトルの平均を利用する．本稿では，付録 B に示す前置詞 52 種類についての平均ベクトルを用いる．この前置詞の平均ベクトルと動詞派生前置詞の単語ベクトル間の余弦類似度を文法化度とする．したがって，対象とする動詞派生前置詞が平均的な前置詞にどれぐらい類似しているかで文法化度を定量化することになる．なお，複数の語からなる動詞派生前置詞については，見かけ上一単語となるように事前処理（例：*according to* であれば *according_to* とする）を行う（以降の類似度でも同様である）．

平均動詞：平均前置詞のように，動詞の平均ベクトルと動詞派生前置詞の単語ベクトル間の余弦類似度を文法化度とする．ただし，動詞との類似性は動詞性に対応するため，余弦類似度の符号を反転した値を文法化度とする．平均は，最頻 200 の動詞を対象とする．入力コーパス中に出現する動詞の原型と変化形について合算した頻度で最頻 200 の動詞を決定する（ただし，動詞派生前置詞の派生元の動詞（例：*consider*）は含めない最頻 200 件とする）．平均は，同様に，200 の動詞における原型および変化形全ての単語ベクトルを考慮したものとする．

平均前置詞＆平均動詞：上述二つの指標の平均を文法化度とする．この指標は，「動詞派生前置詞は動詞性を失い，前置詞性を獲得しつつある」という文献 [3, 4] の考え方を反映したものになっている．

3.2　文脈付き単語ベクトルに基づいた手法

基礎となる考えは，「機能語の文脈は多様であるため，その文脈付き単語ベクトルの方向も多様であり，もし動詞派生前置詞が文法化しているのであれば，同様な傾向を示すはずである」というものである．この考えに基づき，文脈付き単語ベクトルの方向のばらつきを文法化度とする．

幸いなことに，ベクトルの方向のばらつきは，von Mises-Fisher 分布 [11] を通じて定量化が可能である．この分布は，d 次元の単位ベクトル $\mathbf{x}$ に対して，$f(\mathbf{x};\boldsymbol{\mu},\kappa) \propto \exp\left(\kappa\boldsymbol{\mu}^{\mathsf{T}}\mathbf{x}\right)$ と定義される．ここで，$\boldsymbol{\mu}$ ($\|\boldsymbol{\mu}\| = 1$) と κ ($\kappa \geq 0$) は，それぞれ平均方向と集中度と呼ばれるパラメータである．

本稿で重要となるのは，集中度 κ である．von Mises-Fisher 分布では，単位ベクトル $\mathbf{x}$ は平均方向を中心に集中度 κ で等方的に分布すると考える．言い換えれば，κ は，ベクトルの方向の集中度を表しており，上述の考えに合致する．集中度 κ の最尤推定は，近似的に，

$$\kappa \approx \frac{l(d - l^2)}{1 - l^2}, \tag{1}$$

になることが知られている [11]．ここで，l は，観測データの平均ベクトルのノルムである（d は上述の通り，ベクトルの次元である）．ただし，文脈が多様なほど文法化が進んでいると考えるので，集中度の逆数 $1/\kappa$ を文法化度とする．なお，文脈付き単語ベクトルは単位ベクトルとは限らないため，ノルムが 1 となるように事前に正規化しておく．

4　評価実験

文献 [2, 3] に記載された文法化度を用いて各手法の評価を行う．それぞれ，付録の表 3 と表 4 に記した動詞派生前置詞のうち，下記コーパス中で頻度が 30 以上となる 16 種類と 32 種類を評価対象とした．評価尺度はスピアマンの順位相関係数とした．

単語ベクトル取得のためのコーパスとして CCOHA [12] を用いた．CCOHA は，1820〜2019 年

1)　spaCy 3.5.1 (https://spacy.io/)

表 1　人手による文法化度と提案手法による文法化度との間のスピアマンの順相関係数：カッコ内の数値は p 値.

手法	Hayashi	Kortmann
平均前置詞	0.37 (0.03)	0.22 (0.42)
平均動詞	0.09 (0.64)	-0.15 (0.58)
平均前置詞＆平均動詞	**0.69** (0.1×10^{-6})	**0.23** (0.39)
文脈付き単語ベクトル	0.14 (0.43)	-0.07 (0.81)
Hayashi	—	0.19 (0.47)

の文書を収録した歴史コーパスである．本評価実験では，2000 年代の文書を用いた（残りのコーパスは 5 節での通時的考察で利用する）．

文脈なし単語ベクトルの取得には，Gensim の *word2vec*[2] を用いた．文脈付き単語ベクトルについては，*bert-large-uncased*[3] を用いた．最終層の出力ベクトルを単語ベクトルとした．ハイパパラメータを含む実験設定の詳細は付録 C に示す．

表 1 に結果を示す．参考として，人の判定同士（Kortomann ら [2] と Hayashi [3]）の順位相関係数も示している．表 1 より，前置詞の情報に基づいた手法がより高い相関を示すことがわかる．このことは，前置詞性に基づいて動詞派生前置詞の文法化が定量化できるという言語学の仮説 [2, 3] を支持する．ただし，Kortmann との比較においては，相関係数の値が全体的に小さい．これは，Kortmann の文法化度は五段階しかなく，同順位となる動詞派生前置詞が多数あるためである．このことは，人同士の比較においてもみられる．同順位となるものがなくなるように，各カテゴリからランダムに一つの動詞派生前置詞を選んだ場合，**平均前置詞＆平均動詞**の相関係数は 0.7（$p = 0.19$）となった（Hayashi に対しては 0.9（$p = 0.04$））．

また，動詞の情報のみは有効ではないが，前置詞と組み合わせると相関が高まることも確認できる．これは，文献 [3, 4] の「動詞派生前置詞の文法化度は，前置詞性と動詞性の両方を考慮して測られるべきである」という主張と整合する．このことをより明確にするために，図 1 に，Hayashi の文法化度と**平均前置詞**および**平均前置詞＆平均動詞**の文法化度を散布図として表す．この図より，前置詞の情報のみでも，ある程度の相関を示すが，動詞の情報を考慮することでより適切に文法化度の調整が行われていることがわかる．特に，人の判定で文法化度が低い動詞派生前置詞（例：*facing*）においてこの傾向が顕著である．文法化度が低い動詞派生前置詞は，動詞の特性を比較的強く持っており，そのことが単語ベクトルに反映される．そのため，平均動詞ベクトルとの類似度を引くことで文法化度のスコアが下がる．一方で，文法化度が高い動詞派生前置詞は，動詞との類似度は低いため，文法化度のスコアは相対的に高いままである．

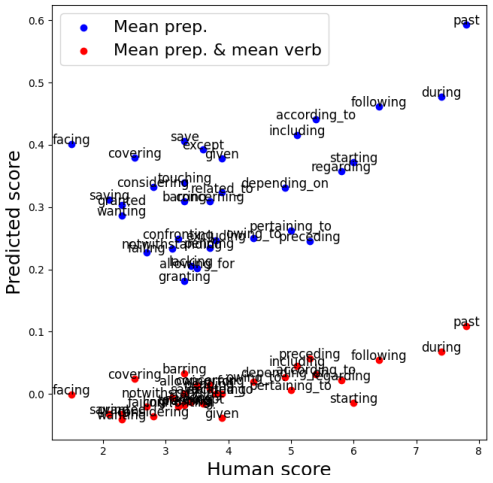

図 1　Hayashi [3] および提案手法の文法化度の関係.

文脈付き単語ベクトルに基づいた手法については，人の判定との相関は低い．この理由として，動詞派生前置詞は完全に文法化されておらず，機能語ほど文脈の多様性が高くないためということが考えられる．適切に文法化度を推定するためには，文脈なし単語ベクトルに基づいた手法のように前置詞の情報など他の情報が必要となる可能性が高い．ただし，この手法では，前置詞や動詞など他の単語の情報は直接使っていないため，更なる改良を加えることで，他の単語にも適用可能な文法化度の一般的な尺度となる可能性もある．

5　考察

前節の評価実験により，文脈なし単語ベクトルに基づいた手法が人の分析に基づいた文法化度に対して一定の相関を示すことをみた．特に，前置詞，動詞，それぞれの平均ベクトルを用いることが効果的であることを確認した．この手法により，動詞派生前置詞とその事例に対して，更に幅広い文法化の分析が行えることが期待される．本節では，**平均前置詞＆平均動詞**（以降，本節では提案手法と省略）を用いて，三つの分析事例を示す．

まず，従来の言語学に見られる動詞派生前置詞の

2）Gensim 4.3.1: https://radimrehurek.com/gensim/models/word2vec.html

3）Hugging Face transformers: https://huggingface.co/docs/transformers/model_doc/bert

表 2 Kortmann [2] と Hayashi [3] で文法化度の判定が異なる動詞派生前置詞と提案手法の定量化結果.

	Kortmann	Hayashi	提案手法
following	2	6.4	5.4×10^{-2}
according to	3	5.4	3.2×10^{-2}
pending	4	3.7	0.8×10^{-2}
concerning	4	3.7	0.2×10^{-2}
except	4	3.6	-0.2×10^{-2}

文法化度の不一致について解消を試みる．具体的には，表 2 に示した 5 種類を吟味する．同表により，*following* について，Kortman ら [2] と Hayashi [3] で大きく判断が異なることがわかる．前者では文法化度が二番目に低いカテゴリに分類されている．一方，後者の文法化度 6.4 は，37 種類の動詞派生前置詞のうち，*past*，*during* に次いで三番目に高い値であり，最も文法化度が高い部類に属する．提案手法においても，この三種類の動詞派生前置詞について Hayashi と同一の順位付けとなった．その他の動詞派生前置詞についても，表 2 の第四カラムに示すように，提案手法は全て Hayashi の結果を支持する．

次に，1 で述べた仮説 1「動詞派生前置詞は完全には文法化しておらず，動詞と前置詞の中間カテゴリに位置する」，2「一部の動詞派生前置詞は他の動詞派生前置詞より文法化が進んでおり，文化の度合いは連続的である」の吟味を行う．そのため，前置詞の文法化度を動詞派生前置詞と共にヒストグラムとして表す（図 2）．対象としたのは付録 B に示した 52 の前置詞である．ただし，文法化度の算出の際には，当該前置詞を除いた他の前置詞についての平均ベクトル（と動詞の平均ベクトル）を用いた．図 2 より，動詞派生前置詞の文法化度は，前置詞に比べて低いレンジに分布することが分かる．ただし，中には前置詞と同程度以上の文法化度を示すものも存在し，文法化度が連続的であることもわかる．この結果は，上の二つの仮説を支持する．

最後に，通時的な分析の例を示す．CCOHA の文書を 10 年ごとに分割し，文法化度を算出した（全てのサブコーパスで頻度が 30 以上となる 31 種類の動詞派生前置詞を対象とした）．図 3 に，その結果を年代と文法化度の関係を表すグラフとして示す（可読性のため，Hayashi の文法化度の高い順に 4 グループに分割している）．図 3 から，第一のグループ（上部の四つの線）は，緩やかに上昇しているように見える．これら既に文法化度が高い動詞派生前置詞は，文法化の最終段階にあり，文法化が加速されているという可能性がある．その他については，大域的には平坦な線となっている．いずれの場合も，文法化度が（大域的に）減少することはみられない．利用可能な情報からは断定はできないが，文法化の一方向性 [13] に対応する可能性もある．時間的により幅広いコーパス（例：Early English Books Online[4]）を用いた分析など更なる調査が待たれる．

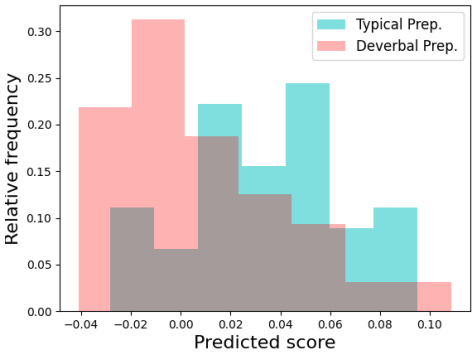

図 2 前置詞と動詞派生前置詞の文法化度の分布.

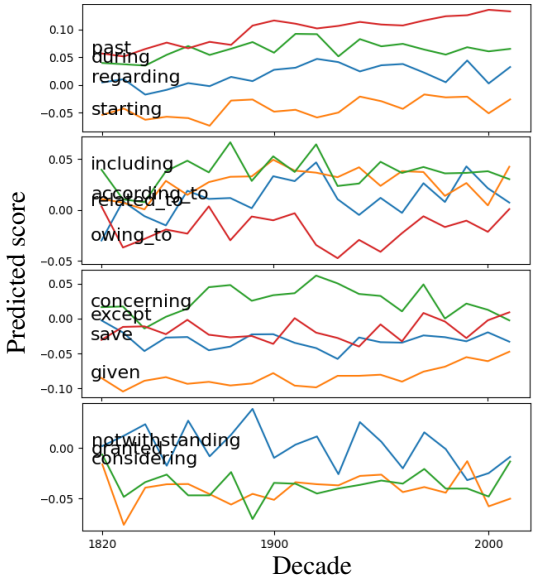

図 3 動詞派生前置詞の文法化度の通時的変化.

6 おわりに

本稿では，動詞派生前置詞の文法化度を定量化する手法を提案し，三つの仮説の検証を行った．評価実験の結果，前置詞と動詞それぞれの平均ベクトルに基づいて定量化した文法化度は人による判定と一定の相関を示すことを確認した．更に，提案手法は，言語学で知られる三つの仮説を全て支持することを報告した．最後に，提案手法を用いた文法化の通時的変化に関する分析の可能性についても報告した．

4) https://quod.lib.umich.edu/e/eebogroup/

謝辞

本研究の一部は JSPS 科研費 JP23K12152 により実施した.

参考文献

[1] Paul Hopper. **On some Principles of Grammaticalization**, pp. 17–35. John Benjamins Publishing Company, 1991.

[2] Bernd Kortmann and Ekkehard König. Categorial reanalysis: The case of deverbal prepositions. **Linguistics**, Vol. 30, pp. 671–698, 1992.

[3] Tomoaki Hayashi. Prepositionalities of deverbal prepositions: Differences in degree of grammaticalization. **Papers in Linguistic Science**, Vol. 21, pp. 129–151, 2015.

[4] Teruhiko Fukaya. **The Emergence of *-ing* Prepositions in English: A Corpus-Based Study**, pp. 285–300. Taishukan, Tokyo, 1997.

[5] Arne Olofsson. A participle caught in the act. on the prepositional use of following. **Studia Neophilologica**, Vol. 62, No. 1, pp. 23–35, 1990.

[6] Tomohiro Kawabata. On the development of considering: The prepositional conjunctive and adverbial usages. **Studies in Modern English (The Twentieth Anniversary Publication of the Modern English Association)**, pp. 138–152, 2003.

[7] Matti Rissanen. **Despite or notwithstanding? On the development of concessive prepositions in English**, pp. 191–203. Gunter Narr Verlag, Germany, 2002.

[8] Jacqueline Visconti. **Conditionals and subjectification: Implications for a theory of semantic change**, pp. 169–192. John Benjamins Publishing Company, Amsterdam/Philadelphia, 2004.

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

[10] 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, June 2019. Association for Computational Linguistics.

[11] Arindam Banerjee, Inderjit S. Dhillon, Joydeep Ghosh, and Suvrit Sra. Clustering on the unit hypersphere using von Mises-Fisher distributions. **Journal of Machine Learning Research**, Vol. 6, No. 46, pp. 1345–1382, 2005.

[12] Reem Alatrash, Dominik Schlechtweg, Jonas Kuhn, and Sabine Schulte im Walde. CCOHA: Clean corpus of historical American English. In **Proc. of the 12th Language Resources and Evaluation Conference**, pp. 6958–6966, 2020.

[13] Paul J. Hopper and Elizabeth Closs Traugott. **Grammaticization**. Cambridge University Press, New York, second edition, 2003.

付録

A　動詞派生前置詞カテゴリ

言語学で提案されている動詞派生前置詞カテゴリを示す．表 3 は，Kortmann ら [2] によるもの，表 4 は，Hayashi [3] によるものである．

表 3　Kortmann ら [2] による動詞派生前置詞の文法化度：値が大きいほど文法化が進んでいる．

Level of Prepositionality	Instances
1	facing, lining, preceding, succeeding
2	considering, failing, barring, following
3	according to, allowing for, owning to, notwithstanding
4	during, pending, except, concerning
5	past, ago, bar

表 4　Hayashi [3] による動詞派生前置詞の文法化度：値が大きいほど文法化が進んでいる．

Instance (prepositional score)
past (7.8), during (7.4), following (6.4), starting (6.0), regarding (5.8), according to (5.4), preceding (5.3), succeeding (5.2), including (5.1), pertaining to (5.0), depending on (4.9), owing to (4.4), related to (3.9), given (3.9), respecting (3.9), excluding (3.8), concerning (3.7), pending (3.7), except (3.6), allowing for (3.5), lacking (3.4), barring (3.3), granting (3.3), save (3.3), touching (3.3), confronting (3.2), notwithstanding (3.1), considering (2.8), failing (2.7), covering (2.5), granted (2.3), wanting (2.3), saving (2.1), bar (1.9), facing (1.5), excepting (1.4), bating (1.3)

B　対象とした前置詞

文脈なし単語ベクトルに基づいた手法で対象とした前置詞は次の通りである：*aboard*, *about*, *above*, *across*, *after*, *against*, *along*, *amid*, *among*, *around*, *as*, *at*, *before*, *behind*, *below*, *beneath*, *beside*, *besides*, *between*, *beyond*, *by*, *despite*, *down*, *for*, *from*, *in*, *inside*, *into*, *like*, *near*, *of*, *off*, *on*, *opposite*, *outside*, *over*, *round*, *since*, *through*, *to*, *toward*, *towards*, *under*, *underneath*, *unlike*, *until*, *up*, *upon*, *via*, *with*, *within*, *without*.

C　評価実験の詳細設定

分析対象コーパス CCOHA に次のような前処理を行った．ノイズと思われる文書は除外した．具体的には，「@@年.txt」（例：@@1525.txt）のように年とファイル名と思われる文字列を含む文書は分析対象外とする．また，文書中のタグ（<P></P> など）は除去した．更に，伏字が含まれている文（CCOHA では，著作権の制限により，一定の割合で文章の一部が伏字になっている）も除外した．

文脈なし単語ベクトルの取得には，Gensim の *word2vec*[5] を用いた．ハイパパラメータは次の通りである：最小頻度 5；ベクトルサイズ 200；窓幅 10[6]；それ以外はデフォルトの値を使用．

5）　Gensim 4.3.1: https://radimrehurek.com/gensim/models/word2vec.html
6）　この他，5 と 20 も試したところ，結論の概要は変わらなかったため本稿では 10 の結果を報告する．