埋め込みベクトルを用いた動詞の意味の粒度分析と共起関係

森下裕三
桃山学院大学
ymorishi@andrew.ac.jp

概要

本研究では，埋め込みベクトルと頻度情報を基に移動動詞の粒度を定量化し、それと着点との共起傾向を検証する。これまでの研究では，移動事象における着点志向性が認知的に優位であり，言語にも反映されることが示されてきた。しかし，着点と共起しやすい動詞については主観的な議論にとどまっていた。本研究では，英語の移動動詞を対象に，COCA から抽出した用例を基に埋め込みベクトルを生成し、その分散の平均値と頻度から粒度を算出する手法を提案する。

1 はじめに

Firth (1957) のよる研究を萌芽として，コーパスを使用した言語学的な研究のみならず，近年の自然言語処理の分野でも分布意味論 (distributional semantics) は大きな影響力をもっている。これは「語の意味は周囲に生起する語によって決まる」という分布仮説に基づいて発展してきたものである。コーパス言語学の分野では，20 世紀後半以降も，分布意味論を発展させた意味的韻律 (semantic prosody) という考え方に基づいて研究が進められている (e.g., Louw 1993, Stubbs 1995, Hunston 2007)。また，自然言語処理の分野でも，分布意味論はニューラル言語モデルや大規模言語モデルの発展へと引き継がれ，文法化 (grammaticalisation) のような言語学的な課題を解決するための糸口になるのではないかと注目されている (e.g., Hamilton et al. 2016)。

認知言語学の分野でも，分布意味論とは独立した形で，使用基盤モデル (usage-based model) と呼ばれる考え方が注目を集めている (Langacker 1988; 2000)，多くの大規模コーパスが構築され普及してきたこともあり，さまざまな言語現象にアプローチするための手段とされている。認知言語学では，私たち人間の認知と言語との関連を探ることを目的としている。このような研究分野において，私たちの認知に見られる非対称性が言語にどのようにあらわれるのかを明らかにしようとするものがあり，使用基盤モデルの枠組みから，コーパスを利用した研究が数多く発表されている (e.g., Stefanowitsch & Rohde 2004, Stefanowitsch 2018, Guse 2022)。

私たちは，誰かが，あるいは何かが移動するような事象を認識する際に，どこから移動したのかという情報よりも，どこへ移動したのかという情報に注意を向ける。このような私たちの認知の偏りについては，Ikegami (1987) による言語学の分野からの議論に端を発し，コーパス言語学のみならず心理言語学の分野でも実証的な研究が報告されてきた (e.g., Lakusta & Landau 2005, Papafragou 2010)。これらの研究で議論になっているのは，移動事象を言語化することにともなう着点 (Goal) と起点 (Source) の非対称性である。たとえば，Kopecka (2012) によるポーランド語の研究では，どこかから物を移動させる場合と比べて，どこかへ物を移動させる場合の方がより詳細に言語化する傾向があることが示されている。

コーパス言語学や心理言語学による研究によって明らかになってきたことは，移動事象を言語化する際に着点志向性がみられるということである。そして，動詞については，この着点と共起しやすいものもあれば，着点とは共起しにくいものもあるということも明らかになっている (e.g., Stefanowitsch 2018)。では，着点と共起しやすい移動動詞にはどのようなものがあるのだろうか。Stefanowitsch (2018) は，汎用的な移動動詞の方が着点と共起しやすい傾向があると主張している。

この汎用的な移動動詞が，着点と共起しやすいという言語学的な仮説を検証するために，本研究では，それぞれの移動動詞の粒度 (granularity) を埋め込みベクトルの分散の平均値と頻度を組み合わせた値によって算出できると主張する。言い換えると，

この手法によって算出された移動動詞の粒度が低いほど，その移動動詞はより汎用的なものだと考えられ，着点とより共起しやすい傾向がみられるということである。

2 先行研究

1 節で述べたように，移動事象に対する私たちの認知には偏りがみられ，認知言語学的な前提から，そのような認知の偏りは言語にも反映されると考え多くの研究が発表されてきた。本節では，本研究と関連性の強い Stefanowitsch & Rohde (2004)，および Stefanowitsch (2018) によるコーパスを利用した先行研究，ならびに移動表現と語の粒度についての研究について批判的に検討する。これらの研究では，英語の移動表現において，どのような移動動詞が，どういった経路表現と共起しやすいのかをコーパスから得られたデータをもとに記述している。その上で，特に着点と共起しやすい移動動詞の性質について議論している。

Stefanowitsch & Rohde (2004) による研究では，調査の対象としている移動動詞の選定が恣意的であるものの，通過点 (TRAJECTORY) や起点 (SOURCE) といった着点以外の経路表現と共起する傾向にある移動動詞についても広く議論されている。使用基盤モデルの枠組みから英語の移動表現の分析をしたものとしては先駆的である。

一方，Stefanowitsch (2018) による研究では，先の研究をさらに進め，構文文法 (Construction Grammar) の枠組みから洗練された統計的手法を利用することによって，英語の移動表現において，着点と共起しやすいのは，汎用的な移動動詞であるということを明らかにした。この発見は，私たちの移動事象における認知の非対称性と言語の関係について議論する上で示唆的である。しかし，Stefanowitsch は汎用的な移動動詞として *go* や *move* を挙げながら，汎用的な移動動詞とはどのようなものを指すのかについて具体的に定義していない。

この汎用的な移動動詞と関連のある概念として粒度というものを挙げることができる。移動表現における研究では，この粒度という概念に言及しながら，着点へのバイアスについて議論している研究がある。たとえば，Guse (2022) は，Tutton (2013: 50) および Wnuk (2016) や Stathi (2017) による定義を引用しながら，事象や語がもつ情報の緻密さとして粒度という概念を定義し，粒度の低い動詞ほど着点と共起しやすいと主張している。Stathi (2017) によると，移動動詞であれば *walk* は *go* よりも粒度が高く，*saunter* は *walk* よりもさらに粒度が高いとしている。ここで挙げられている例をみる限り，移動動詞の粒度については直感に反するものではない。

しかし，この定義では，英語の語彙に存在する多くの移動動詞の粒度をうまく捉えることはできない。なぜなら，英語は衛星枠付言語 (satellite-framed language) と呼ばれる類型に属し，移動の様態を語彙化した動詞が非常に多いからである (cf. Talmy 2000)。たとえば，英語の動詞について語彙意味論の観点から議論した Levin (1993) は，英語の移動動詞を 100 以上も挙げている。他にも実験的手法によって複数の言語の移動動詞について分析している Slobin et al. (2014) も，やはり英語の移動動詞を 70 以上も挙げている。このように，数多く存在する英語の移動動詞の粒度を比較し，どの移動動詞が，どの移動動詞よりも粒度が高いのかを議論することは容易ではない。実際に，*saunter* と *stroll* の粒度を比較して，どちらの方がより粒度の高い動詞なのかを直感によって判断することなどできそうにない。

先行研究において議論されてきたこれらの点を踏まえ，移動動詞の粒度を定量的に算出する手法を提案し，実際に英語の移動表現において，粒度の低い動詞ほど着点と共起しやすいのかを検証することが本研究の目的である。

3 移動動詞の粒度を定量化する手法

2 節で議論したように，移動事象の認知における非対称性と言語との関係を解き明かすためには，移動動詞の粒度を定量化する必要がある。本節では，BERT (Devlin et al. 2019) の `bert-base-uncased` を利用した文脈付き埋め込みベクトルと，コーパスにおける各移動動詞の頻度を考慮した値を粒度の定量化の算出に利用する。埋め込みベクトルだけでなく，頻度の情報も定量化に組み込むのは，使用基盤モデルでは，言語学的な現象の分析に頻度が重要であることが知られているからである (e.g., Bybee 2007, cf. Arona et al. 2017)。

まず，COCA (Corpus of Contemporary American English) と呼ばれるコーパス (Devies 2008–) から，語形に基づきランダムに次の 16 種類の移動動詞を含む文を抽出する。

(1) *amble*, *come*, *go*, *jog*, *run*, *sashay*, *saunter*,

scamper, *scurry*, *scuttle*, *sprint*, *stride*, *stroll*, *trot*, *walk*, *run*

なお、これらの動詞は，Levin (1993) および Slobin et al. (2014) が挙げた移動動詞のリストに共通する 44 種類の動詞から選んだものである。

続いて，抽出した文から目視によって物理的な移動をあらわし，かつ動詞が定形 (finite form) のもので，移動の経路が省略されていないものを 100 例ずつ選び出す。その上で，各移動動詞を含む文の埋め込みベクトルを作る。この時，COCA のテキストデータに含まれる @ @ @ @ @ @ @ @ @ @ が含まれる文はすべて削除する。残った用例を埋め込みベクトルに変換し，各埋め込みベクトルの分散の平均値に次のように調整頻度を加えることで各動詞の調整埋め込みベクトルとした。

$$\alpha + \log_{10}(1 + \text{frequency}) \qquad (1)$$

結果として，この値が高ければ高いほど，その移動動詞の粒度は低いということになる。

4 結果

物理的な移動をあらわす定形の移動動詞のみを分析の対象としたにもかかわらず，それぞれの動詞の各例のベクトルの分散には違いが見られた。代表的な動詞について，t-SNE によって次元を削除して可視化したものを以下に示す。

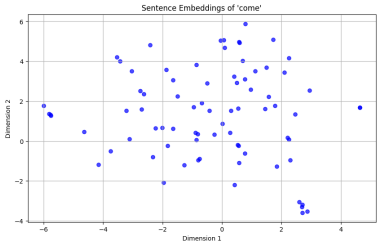

図 1 物理的移動をあらわす *come* のベクトル

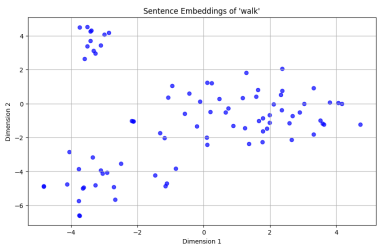

図 2 物理的移動をあらわす *walk* のベクトル

図 1 と図 2 を比較するだけでも，それぞれの動詞の粒度の違いがわかる。分散が大きい *come* は，図 1 からもわかるように，どの例もそれぞれ違った傾向を示している。一方，より埋め込みベクトルの分散が小さい *walk* は，*come* と比べると分布のばらつきが小さい。

続いて，定量化された各移動動詞の粒度と着点との共起にどれほどの相関がみられたのかを確認する。結果は次の図 3 に示す通りである ($r = 0.70$, $p < 0.01$)。

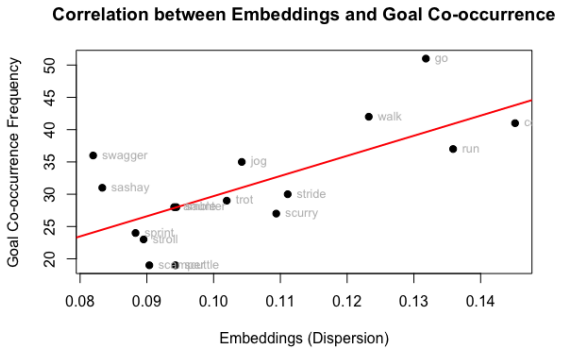

図 3 各移動動詞の粒度と着点との共起関係

本研究で提案した移動動詞の粒度の定量化手法はおおむね評価できる結果を得ることができた。また，今回の結果を踏まえ，やはり先行研究で議論されてきた汎用的な移動動詞や粒度の低い移動動詞とされてきたものは，着点と共起しやすいことを改めて確認することができた。

しかし，*go* よりも *come* の粒度が高いこと，それに着点との共起が比較的多くみられる *swagger* の粒度が高いことなど検討の余地は残されているように思われる。

5 考察

本節では，4 節で得られた分析結果を踏まえ，各動詞の埋め込みベクトルの分散の平均値と頻度の情報を組み合わせた手法によって妥当な結果が得られた理由について議論する。

本研究の分析手法から妥当な結果が得られた理由として考えられるのは，動詞の埋め込みベクトルと各動詞の頻度情報を組み合わせたことにある。Zipf (1945) や Hamilton et al. (2016) による議論からも明らかなように，頻度の高い語ほど多義的で，さまざまな種類の語と共起しやすい。このことを本研究との関連で述べると，頻度の高い語ほど埋め込みベクトルの分散の平均値は低くなるということになる。実際に，Morishita (to appear) は，本研究と同じ動詞を対象として，COCA における各動詞の頻度のみに

着目し，頻度の高い移動動詞と着点との共起には強い相関がみられることを示している。

しかし，これまでも述べてきたように，言語学的な関心は，移動事象における私たちの認知の非対称性と言語との関連である。つまり，移動動詞であっても次のように物理的な移動をあらわさない用例を分析に組み込んでしまうため，頻度のみを指標にする手法には問題がある。

(2) a. ”It’s not going to happen overnight,” Rosa DeLauro told me. (COCA, 2015)
b. Eventually, over many months, the Groeschens came to realize the old Tom might never truly return. (COCA, 2007)

また，次の例からもわかるように，特に埋め込みベクトルの分散が小さい動詞は，共起する語，具体的には主語として生起する名詞句などにより強い制限がみられる傾向がある。

(3) a. The wolves trotted down the trail away from the cave, followed by the horses and their riders. (COCA, 1994)
b. The dog trotted over and looked at us with smiling eyes. (COCA, 2007)

このような理由から，各移動動詞の頻度情報と埋め込みベクトルの分散の平均値を組み合わせた手法によって妥当な結果が得られたのだと考える。

また，本研究での提案を応用することによって，移動の経路を語彙化した経路動詞 (path verb) と呼ばれるものの粒度も比較することが可能になる。Stefanowitsch (2018) による研究でも，汎用的な移動動詞のひとつとして挙げられていた *enter* などの粒度を，移動の様態を語彙化した動詞の粒度と比較することもできるだろう。Slobin et al. (2014) による研究では議論されていなかったため，本研究でも分析の対象から外していた移動動詞の *drop* などは，次の例からもわかるように共起する語に特徴的な性質がみられるため，他の移動動詞の粒度と比較していく必要がある。

(4) a. The book dropped from her hand, and for a moment she struggled to get away, then she figured out what was going on and leaned into him instead. (COCA, 2017)
b. A tree frog dropped from a branch onto my upturned neck, and I brushed it away without so much as a squeal. (COCA, 1996)

6 おわりに

本研究では，英語の移動表現にみられる着点志向性とかかわる語の粒度の定義という言語学的な課題に対して，自然言語処理の分野で発展してきた技術を応用することで解決を試みた。これまでの言語学的な研究成果も踏まえた上で，移動動詞の粒度を定量化するために，頻度情報も組み合わせた埋め込みベクトルの分散の平均値という指標を提案した。この指標の妥当性は，粒度の低い移動動詞が着点と共起しやすいという傾向を確認することによって示された。

しかしながら，本研究で分析の対象とした動詞は，直示的な移動動詞である *come* と *go* を含む 16 種類のみである。既に述べたように，英語の語彙のは多くの移動動詞があり，今後はこれらの移動動詞についても同様の手法によって検証を進める必要があるだろう。

7 謝辞

本研究の一部は，JSPS 科研費 20K13069 の助成を受けたものである。

参考文献

Arona, S., Y. Liang & T. Ma. 2017. A simple but rough-to-beat baseline for sentence embeddings. *Paper presented at 5th International Conference on Learning Representations*, ICLR 2017, Toulon, France.

Bybee, J. L. 2007. *Frequency of use and the organization of language*. Oxford University Press.

Davies, M. 2008–. The Corpus of Contemporary American English (COCA). Available online at https://www.english-corpora.org/coca/.

Devlin, J., M.-W. Chang, K. Lee, & K. Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Associatio. for Computational Linguistics: Human Language Technologies, vol. 1*, pp. 4171–4186. Association for Computational Linguistics.

Firth, J. R. 1967. A synopsis of linguistic theory 1930–55. In *Studies in linguistic analysis*, 1–31. Blackwell.

Guse, L. 2022. Source-Goal asymmetry in German: A corpus study comparing intentional and non-intentional motion events. L. Sarda & B. Fagard (eds.) *Neglected aspects of motion-event description: Deixis, asymmetries, constructions*, 173–185. John Benjamins Publishing Company.

Hamilton, W. L, J. Leskovec, & D. Jurafsky. 1996. Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1489–1501.

Hunston, S. 2007. Semantic prosody revisited. *International Journal of Corpus Linguistics 12*, 249–268.

Ikegami, Y. 1987. 'Source' versus 'Goal': A case of linguistic dissymmetry. In R. Dirven & G. Radden (eds.), *Concepts of case* (Studien Zur Englischen Grammatik 4), 122–146. Tübingen: Narr.

Kopecka, A. 2012. Semantic granularity of placement and removal expression in Polish. In A. Kopecka & B. Narashimhan (eds.), *Events of putting and taking: A crosslinguistic perspective*, 327–346. John Benhamins.

Lakusta, L. & B. Landau. 2005. Starting at the end: The importance of goals in spatial language. *Cognition, 96*. 1–33.

Langacker, R. 1988. A usage-based model. In B., Rudzka-Ostyn (ed.), *Topics in cognitive linguistics*, 127–161, John Benjamins.

Langacker, R. 2000. Dynamic usage-based model. In M., Barlow & S., Kemmer (eds.) *Usage-based models of language*, 1–65, CSLI Publications.

Levin, B. 1993. *English verb classes and alternations: A preliminary investigation*. The University of Chicago Press.

Louw, B. 1993. Irony in the text or insincerity in the writer?: The diagnostic potential of semantic prosodies. In M. Baker, G. Francis, & E. Tognini-Bonelli (eds.), *Text and technology: In honour of John Sincair*. 157–176. John Benjamins.

Morishita, Y. to appear. Frequency and congruency: A new perspective on motion verb and path expression co-occurrence. *Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation*.

Papafragou, A. 2010. Source-goal asymmetries in motion event representation: Implications for language production and comprehension. *Cognitive Science, 34*, 1064–1092.

Slobin, D. I., I. Ibarretxe-Antuñano, A. Kopecka & A. Majid. 2014. Manners of human gait: A crosslinguistic event-naming study. *Cognitive Linguistics 25*. 701–741.

Stathi, K. 2017. Granularity effects in event descriptions: A cross-linguistic study. Poster presented at the workshop 'Event Representations in Brain, Language, and Development', Max-Plank-Institute for Psycholinguistics, Nijmegen, 29 October 2017.

Stefanowitsch, A. 2018. The goal bias revisited: A collostructional approach. *Yearbook of the German Cognitive Linguistics Association 6*. 143–166.

Stefanowitsch, A. & A. Rohde. 2004. The goal bias in the encoding of motion events. In G., Radden & K.-U., Panther (eds.), *Studies in linguistic motivation* (Cognitive Linguistics Research 28), 249–267. Mouton de Gruyter.

Stubbs, M. 1995. Collocations and semantic profiles: On the cause of the trouble with quantitative studies. *Functions of Languages 2*, 1–33.

Talmy, L. 2000. *Toward a cognitive semantics, Volume II: Typology and process in concept structuring*. Cambridge, MA/London, England: The MIT Press.

Tutton, M. 2013. Granularity, space, and motion-framed location. In M. Vulchanova & E. van der Zee (eds.), *Motion encoding in language and space*, 149–165. Oxford: Oxford University Press.

Wunk, E. 2016. Semantic specificity of perception verbs in Maniq. Ph. D. Thesis, Radboud University.

Zipf, G. K.. 1945. The meaning-frequency relationship of words. *Journal of General Psychology 33*, 251−256.