Cross-Corpora Analysis of Spatial Language: The Case of Fictive Motion

Ekaterina Egorova
Massey Geoinformatics Collaboratory, Massey University, Auckland, New Zealand
e.egorova@massey.ac.nz

Niloofar Aflaki
Massey Geoinformatics Collaboratory and School of Natural and Computational Sciences, Massey University, Auckland, New Zealand
n.aflaki@massey.ac.nz

Cristiane K. Marchis Fagundes
Department of Geomatics, Federal University of Paraná, Curitiba, Paraná, Brazil
cristiane.fagundes@ufpr.br

Kristin Stock
Massey Geoinformatics Collaboratory, Massey University, Auckland, New Zealand
k.stock@massey.ac.nz

Abstract

The way people describe where things are is one of the central questions of spatial information theory and has been the subject of considerable research. We investigate one particular type of location description, fictive motion (as in, *The range runs along the coast*). The use of this structure is known to highlight particular properties of the described entity, as well as to convey its configuration in physical space in an effective way. We annotated 496 fictive motion structures in seven corpora that represent different types of spatial discourse – news, travel blogs, texts describing outdoor pursuits and local history, as well as image and location descriptions. We analysed the results not only by examining the distribution of fictive motion structures across corpora, but also by exploring and comparing the semantic categories of verbs used in fictive motion. Our findings, first, add to our knowledge of location description strategies that go beyond prototypical locative phrases. They further reveal how the use of fictive motion varies across types of spatial discourse and reflects the nature of the described environment. Methodologically, we highlight the benefits of a cross-corpora analysis in the study of spatial language use across a variety of contexts.

2012 ACM Subject Classification Information systems → Content analysis and feature selection

Keywords and phrases spatial language, spatial discourse, fictive motion, location, cross-corpora analysis

Digital Object Identifier 10.4230/LIPIcs.COSIT.2019.9

Category Short Paper

Funding *Ekaterina Egorova*: This study was financed by the Swiss National Science Foundation, grant number P2ZHP2$_1$7475
Niloofar Aflaki: This study was financed in part by Ordnance Survey-UK
Cristiane K. Marchis Fagundes: This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) Finance Code 001 and by Conselho Nacional de Desenvolvimento Científico e Tecnológico - Brasil (CNPq)

Acknowledgements We gratefully acknowledge data provided by Manaaki Whenua - Landcare Research NZ (in the form of descriptions from the National Soils Database) to assist in this research.

1 Motivation and Background

Any language is a network of options and, whether consciously or not, its speakers regularly make linguistic choices in everyday communication situations [23]. Spatial language in particular offers multiple options for the subtle windowing of attention of a listener to particular aspects of described scenes [21]. Borrowing an example by Matlock, imagine the scene behind *The table goes from the kitchen wall to the sliding glass door* [9]. In hearing this sentence, our minds automatically draw a table that is long and narrow – we do not visualise a round kitchen table or a small square coffee table. This interpretation is evoked by the use of fictive motion (henceforth FM), a linguistic structure that includes motion information in the description of the location of a static entity [20].

Understanding patterns of FM use is important for spatial information theory for several reasons.

First, multiple studies in cognitive linguistics and psychology have shown how the use of FM in language reflects the focus of attention of the speaker, while also affecting the mental representation of a scene as constructed by the listener [13, 9, 11, 12]. From the speaker's side, the use of FM signals the conceptual primacy of a spatial entity and its configuration in space, since FM offers an efficient way for conveying information about the physical layout of a scene [13, 9]. Further, FM can highlight and even construe spatial properties of the described entity, as in the example with the table above, where the table "becomes lengthened through dynamic construal" [12, p. 548]. Other properties reported in previous studies on FM include vertical orientation (as in, *route plunges*) and complex shape or structure (as in, *glacier spills*, *mountains roll*) [4]. For the listener, FM and the semantics of the verb induce a mental simulation of motion, which results in a particular mental representation of a spatial scene [11]. A number of experiments with participants have demonstrated that people process semantically equivalent expressions with FM and without FM differently [10, 8, 13]. For example, they would draw a longer tattoo if the latter was described with *The tattoo runs along his spine*, as opposed to *The tattoo is next to the spine* [11]. Studying patterns of FM use can thus provide insights into particular aspects of the construction and communication of mental representations of space.

Second, FM is described as "pervasive across languages" and is known to often occur "when people are describing physical space" [8, p. 1390]. Given this pervasiveness, the importance of studying patterns of FM use has been acknowledged in the line of work that explores spatial concepts found in various types of spatial discourse [15]. Developing a spatial language annotation scheme, Pustejovsky and Yocum introduce the MOTION SENSE attribute to distinguish between different interpretations of motion events, one of them being FICTIVE [16]. In a corpus-based study, Egorova et al. further report on types of scenes and spatial concepts encoded by FM: actual motion of the observer (as in, *The second icefield led much more quickly than anticipated*), general encyclopedic knowledge (as in, *The range runs east west across the central part of the Tibet plateau*), vistas (as in, *Far off, a great red buttress rose steeply*)[4]. Building upon [4], another study develops a rule-based approach for the automated extraction and classification of FM from text, demonstrating the non-trivial nature of such tasks but pointing out that exploring patterns of use of figurative language such as FM is necessary if we want to develop algorithms that exhibit spatial awareness [3].

Fictive motion has been thoroughly studied in cognitive linguistics, but research is largely based on introspection (e.g. [21]) or general corpora such as BNC or novels (e.g. [19]). To the best of our knowledge, no studies have actually examined the frequencies and patterns of its use across different types of spatial discourse. Conducting a cross-corpus analysis of FM

can contribute to our knowledge of spatial description strategies used in different contexts. It can also make a step towards a more nuanced automated annotation of spatial information in text, which is crucial given that the correct interpretation of motion has a "lasting effect on the interpretation of a text with respect to spatial information" [15, p. 992]. To address this gap and explore the opportunities, the following research questions have been formulated for this study:

- RQ1: How frequent is FM in various types of spatial corpora?
- RQ2: How do motion verbs in FM and their semantic classes differ across corpora?
- RQ3: What does this tell us about spatial discourse production in different contexts?

2 Data and Methods

2.1 Corpora

Four corpora were used in our analysis, and one of the four – **Nottingham Corpus of Geospatial Language (NCGL)** [18] – was divided into four subcorpora, allowing analysis across different domains.

The **NCGL: News** sub-corpus includes 1592 geospatial sentences from 14 news web sites from the USA, Australia, New Zealand and South Africa.

The **NCGL: Travel and Tourism** sub-corpus includes 3380 sentences from 9 web sites including travel blogs (e.g. Seat61), tourism agency sites (e.g. Tourism NZ) and tourism publishers (e.g. Lonely Planet).

The **NCGL: Outdoor pursuits** sub-corpus contains 1822 sentences from 6 sites, mainly focused on walking (e.g. Arizona Trails, BBC Walks).

The **NCGL: Local History** sub-corpus contains 2104 sentences and focuses on local history, harvested from 11 sites, mostly from the UK (with one site from Australia).

The **Geograph** corpus includes descriptions from Geograph, an online project that collects geographically representative images (and their descriptions) for every square kilometre of the British Isles.[1] Descriptions used in this study refer to images of six neighbouring squares within the urban area of London and are represented by 3153 sentences.

The **Where am I?** corpus was created using human subjects experiments [17], in which respondents were shown an array of photos of a particular location, and were asked to imagine that they had witnessed an accident and to describe the location to emergency services. 178 native English speakers responded to the experiment, which resulted in a corpus of 737 sentences.

The **National Soils Database**[2] (**NSD**) is a collection of descriptions of locations of soil specimens gathered by Manaaki Whenua - Landcare Research in New Zealand. Our corpus consists of a subset numbering 1389 sentences.

2.2 Fictive motion annotation and analysis

FM annotation involved two steps: automated annotation of FM candidates in the corpora and manual annotation of FM among the candidates. In the first step we automatically identified sentences containing motion verbs, based on part-of-speech tags and lemmas of motion verbs as compiled from two sources [4, 6]. In the second step the candidates were split into

[1] https://www.geograph.org.uk

[2] https://soils.landcareresearch.co.nz/soil-data/national-soils-data-repository-and-the-national-soils-database/

Table 1 Characteristics of corpora and FM in them.

Corpus	N of sentences	N of FM candidates	N of FM	FM per candidate
Local history	2104	239	114	0.48
Outdoor pursuits	1822	478	192	0.4
News	1592	291	27	0.09
Travel and tourism	3380	575	87	0.15
NSD	1389	107	5	0.05
Geograph	3153	312	27	0.09
Where am I?	737	249	44	0.18
Total	14177	2251	496	n/a

three sub-corpora and FM structures were annotated in each sub-corpus by one of the first three authors of the paper. A motion event was considered fictive if the noun linguistically represented as a moving entity was a static entity. To measure the inter-coder agreement, we randomly sampled 5% of candidate sentences from each corpus, and the resulting corpus of 112 candidate sentences was independently annotated by the three mentioned annotators. Further, we performed a cross-corpora comparison of FM using a combination of quantitative and qualitative methods. First, we calculated the ratio between the number of FM structures and the number of FM candidates as a proxy for evaluating the pervasiveness of FM in each corpus. Second, we classified the verbs according to their semantics and compared the distribution of classes across corpora in order to examine how the semantics of verbs reflects the nature of spatial discourse. The motion verbs' classification scheme was borrowed from [22] where path verbs include "Source-originating" (e.g. *leave*), "Goal-oriented" (e.g. *reach*), "Vertical" (e.g. *ascend*), "Trajectory" (e.g. *cross*) and "Change in direction verbs" (e.g. *turn*), while manner verbs include "Complex shape trajectory" (e.g. *wind*) and "Trajectory of unspecified shape" (e.g. *roll*) verbs.

3 Results and Interpretation

In total, 496 FM structures were identified in all corpora. The average pairwise Cohen's Kappa, a standard measure of agreement [2], is 0.78, which is a good positive indication of the reliability of the annotation.

3.1 Fictive motion frequency

The highest proportion of FM in relation to the number of FM candidates[3] is found in **Local history** – 0.48 (see Table 1). This might reflect the nature of the corpus – focussing on the history of rural England, it is rich in descriptions of vistas of local (both built and natural)

[3] This metrics essentially represents the ratio between the number of all motion verbs in the corpus and the number of verbs used in FM. We chose to report this ratio (instead of verbs used in FM per number of sentences) for two main reasons. First, it avoids the problem arising from the presence of multiple FM in one sentence. Second, this ratio is more revealing about the use of motion verbs, and is more relevant for the line of work that captures spatial information in text and distinguishes between various interpretations of motion events [16].

landscapes. Additionally, it has a largely poetic flavour and creatively deploys FM to convey nuanced properties of the environment (as in, *To the left is the mansion, skirted by the gloomy cedars, and beyond, the lake expanding into a noble sheet of water is embosomed in magnificent woods*). Another corpus with a high ratio of FM is **Outdoor pursuits** (0.4). From the perspective of spatial discourse, texts in this corpus are mostly descriptions of trails or narratives about completed walks, whereby FM appears to offer an effective way of communicating spatial information about trail-like features (as in, *The path goes between trees by the side of the lake*). The ratio of FM in **Travel** was found to be quite low (0.15) which can be explained by the fact that the corpus mostly describes travelling by trains and buses. Thus, while spatial descriptions including motion events abound, they mostly include means of travel (e.g. *ferry*, *train*) and factive motion (as in, *The Livorno train heads down the Rhine Valley in the early evening, past castles, Rhine river barges and vineyards*).

The smallest ratios are found in NSD (0.05), News (0.09), Geograph (0.09) and Where am I? (0.18). In the case of **NSD** (0.05), this might be explained by the nature of geographic information that it contains – namely, precise quantitative metrically-grounded descriptions of small-scale locations. In a rather similar way, FM is rarely used in **Where am I?** (0.18), which might also be explained by the necessity to describe one's own position as precisely as possible. This results in a high frequency of prototypical locative phrases and descriptions of landmarks (as in, *There's a pedestrian crossing and a disabled parking spot in front of the school building. The building is brick and concrete fame with blue walls*). Furthermore, the scenario given to respondents in the Where am I? survey was rather utilitarian and urgent in nature, allowing little room for consideration of different modes of expression. The sample from **Geograph** (0.09) also represents descriptions of urban vistas (captured in images), but since there is no task of describing the location, descriptions of space are rather scarce – instead, the focus is often on people and events (as in, *A guard stands to attention as the people walk by*). Finally, in the case of **News** (0.09) the low frequency of FM reflects the focus on events and their locations, mostly represented by the first- and second-order political entities such as countries and regions (as in, *Net traffic will travel to the satellite through Hughes' Earth station near Los Angeles*).

3.2 Verbs in fictive motion

57 different motion verbs occurred in 496 FM structures in the corpora. The most frequent verbs were *run* (86 inst.), *lead* (73 inst.), *pass* (50 inst.), *go* (47 inst.), *cross* (34 inst.), *turn* (31 inst.), *take* (26 inst.), *follow* (25 inst.), *climb* (17 inst.), *wind* (16 inst.). Among the verb classes, the most prominent class is "Trajectory of unspecified shape", followed by "Vertical" and "Trajectory" classes. A cross-corpora comparison of the distribution of classes in the three largest sub-corpora of fictive motion (**Outdoor pursuits**, **Local history**, and **Travel**) invites for several observations.

"Complex shape trajectory", while almost absent in **Local history** (2.63%), has slightly higher proportions in the two other corpora (7.35% for **Outdoor pursuits**, 8% for **Travel**) where it is represented by verbs such as *wind*, *meander*, *snake*, *wrap*, *twist*. This class of verbs is mostly used to describe the shape of water bodies (as in, *The Nile snakes through upper Egypt*) or trails (as in, *The track meanders through gullies*).

"Vertical" verbs are similarly frequent in **Local history** (14%) and **Outdoor pursuits** (15.1%). However, in **Outdoor pursuits** the two dominant verbs are *climb* and *drop*, usually collocated with a trail-like entity. In contrast, in **Local history** this class is overly represented by *rise* that is frequently used to describe human-built parts of the landscape (as in, *The spire of Edwinstowe Church rises gracefully from among the old oaks*).

"Trajectory of unspecified shape" is the most frequent class in both **Local history** (35%) and **Travel** (42.5%). Verbs in this group mostly encode the spatial extension of an entity (as in, *A small wood stretched from Jenny Burton's Hill to near her cottage*). In **Outdoor pursuits**, in contrast, "Trajectory of unspecified shape" represents 18% only, while the most frequent class is "Trajectory" (verbs such as *cross*, *follow*, *traverse*). This reflects the focus on the path of the motion event in the context of outdoor activities, where locomotion is an important part of navigation, as in *The footpath initially follows the right hand field boundary.*

4 Conclusion and Outlook

A cross-corpora analysis of fictive motion has provided us with several insights that have important implications and invite for further investigations.

First, our findings suggest that the relative frequency of FM in a particular type of spatial discourse depends on aspects such as the scale of described scenes (we found more FM in the descriptions of vistas of landscapes in **Local history** and spatial layouts of trails in **Outdoor pursuits**), required precision of spatial information (we saw less instances of FM in **NSD**, where preference is given to metric locative phrases), as well as the main theme of spatial descriptions (we saw less instances of FM in **News**, where spatial information mostly relates to the location of events, and not spatial entities). Second, the semantic classes of verbs found in FM further reflect the peculiarities of each corpus, both from the perspective of the described environment and from the perspective of spatial information in focus. Example of the former is the low frequency of "Trajectory of unspecified shape" verbs in **Local history**, which might be a result of the absence of features such as large winding rivers in the described area. Example of the latter is the high frequency of "Trajectory" verbs in **Outdoor pursuits**, which might reflect the focus on the path in the context of walking and hiking.

These findings have practical implications for several lines of work within the spatial information theory. For the line of work developing spatial annotation schemes and capturing spatial information in text [15, 14], this study highlights the fact that taking FM into account is especially legitimate when working with corpora similar in their characteristics to **Local history** and **Outdoor pursuits**. It also provides insights which are of key importance for the development of parsers that are capable of distinguishing between factive and fictive motion in text [3]. The findings are further relevant for the development of spatial language generation systems [5, 7, 24], given that the use of FM has the capability of inducing a more effective processing of spatial information through the simulation of motion [8]. Finally, for the line of research looking into spatial language use across various contexts, the study brings an important message of the utility of a cross-corpus analysis.

In future work, we plan to enhance our understanding of FM use through more controlled, hypothesis-driven studies. In the next step, we aim at performing a more systematic analysis of verbs' classes and types of spatial entities occurring in FM, as well as at exploring potential conventionalised FM structures. More broadly, further work is also required for a better understanding of how we could model the representations of spatial scenes encoded by FM, and how situatedness and context impact FM use and interpretation [1].

References

1 John A Bateman, John Hois, Robert Ross, and Thora Tenbrink. A linguistic ontology of space for natural language processing. *Artificial Intelligence*, 174(14):1027–1071, 2010.

2 Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

3 Ekaterina Egorova, Ludovic Moncla, Mauro Gaio, Christophe Claramunt, and Ross S Purves. Fictive motion extraction and classification. *International Journal of Geographical Information Science*, 32(11):2247–2271, 2018.

4 Ekaterina Egorova, Thora Tenbrink, and Ross S Purves. Fictive motion in the context of mountaineering. *Spatial Cognition & Computation*, 18(4):259–284, 2018.

5 Mehdi Ghanimifard and Simon Dobnik. Knowing when to look for what and where: Evaluating generation of spatial descriptions with adaptive attention. In Laura Leal-Taixé and Stefan Roth, editors, *Computer Vision – ECCV 2018 Workshops. ECCV 2018. Lecture Notes in Computer Science, vol 11132*, pages 153–161. Springer, Cham, 2018.

6 Beth Levin. *English verb classes and alternations: A preliminary investigation*. University of Chicago press, Chicago, London, 1993.

7 Tomoyuki Maekawa and Wataru Takano. Spatial representation of context-dependent sentences and its application to sentence generation. *Advanced Robotics*, 31(15):780–790, 2017.

8 Teenie Matlock. Fictive motion as cognitive simulation. *Memory & Cognition*, 32(8):1389–1400, 2004.

9 Teenie Matlock. The conceptual motivation of fictive motion. In Günter Radden and Klaus-Uwe Panther, editors, *Studies in linguistic motivation*, pages 221–248. Mouton de Gruyter, Berlin, New York, 2004.

10 Teenie Matlock. Depicting fictive motion in drawings. In June Luchjenbroers, editor, *Cognitive Linguistics Investigations: Across languages, fields and philosophical boundaries*, volume 15, pages 67–85. John Benjamins Publishing, Amsterdam, 2006.

11 Teenie Matlock. Abstract motion is no longer abstract. *Language and Cognition*, 2(2):243–260, 2010.

12 Teenie Matlock and Till Bergmann. Fictive Motion. In E Dąbrowska and D. Divjak, editors, *Handbook of Cognitive Linguistics*, pages 546–562. Berlin, Mouton de Gruyter, 2014.

13 Teenie Matlock and Daniel C Richardson. Do eye movements go with fictive motion? *Proceedings of the Annual Meeting of the Cognitive Science Society*, 26(26), 2004.

14 Ludovic Moncla, Mauro Gaio, Javier Nogueras-Iso, and Sébastien Mustière. Reconstruction of itineraries from annotated text with an informed spanning tree algorithm. *International Journal of Geographical Information Science*, 30(6):1137–1160, 2016.

15 James Pustejovsky. ISO-Space: Annotating static and dynamic spatial information. In Nancy Ide and James Pustejovsky, editors, *Handbook of Linguistic Annotation*, pages 989–1024. Springer, Dordrecht, 2017.

16 James Pustejovsky and Zachary Yocum. Capturing motion in ISO-Spacebank. In *Proceedings of the 9th Joint ISO-ACL SIGSEM Workshop on Interoperable Semantic Annotation*, pages 25–34, 2013.

17 Kristin Stock and Luciene Delazari. Where am I? The challenges of interpretation of natural language descriptions of geographical location. In *SAI Computing Conference (SAI), 2016*, pages 1335–1338. IEEE, 2016.

18 Kristin Stock, Robert C Pasley, Zoe Gardner, Paul Brindley, Jeremy Morley, and Claudia Cialone. Creating a corpus of geospatial natural language. In Thora Tenbrink, John Stell, Antony Galton, and Zena Wood, editors, *Spatial Information Theory. COSIT 2013. Lecture Notes in Computer Science, vol 8116*, pages 279–298. Springer, Cham, 2013.

19 Dejan Stosic and Laure Sarda. The many ways to be located in French and Serbian: the role of fictive motion in the expression of static location. In Brala M. Vukovic and Gruic L. Grmusa, editors, *Space and Time in Language and Literature*, pages 39–60. Cambridge Scholars Publishing, 2009.

20 Leonard Talmy. Fictive motion in language and "ception". In Paul Bloom, Merrill F Garrett, Lynn Nadel, and Mary A Peterson, editors, *Language and Space*, pages 211–277. MIT Press, Cambridge, MA, 1996.

21 Leonard Talmy. *Toward a cognitive semantics*. MIT Press, Cambridge, MA, 2000.

22 Piia Taremaa. Fictive and actual motion in Estonian: Encoding space. *SKY journal of linguistics*, 26:151–183, 2013.

23 Thora Tenbrink. Cognitive discourse analysis: Accessing cognitive representations and processes through language data. *Language and Cognition*, 7(1):98–137, 2015.

24 Jette Viethen and Robert Dale. The use of spatial relations in referring expression generation. In *INLG'08 Proceedings of the Fifth International Natural Language Generation Conference, Salt Fork, Ohio, June 12 - 14, 2008*, pages 59–67. Association for Computational Linguistics, Stroudsburg, PA, USA, 2008.