Tracing the shifting collocates of complex prepositions with diachronic word embeddings: the Prep NP_{IL} of construction

Friday, August 20th, 2021 – Universiteit Antwerpen

Guillaume Desagulier
gdesagulier@univ-paris8.fr
CxG: how it started

- Construction Grammar (CxG) emerged in reaction against modular linguistic theories.
- CxG pioneers looked for non-canonical, idiomatic patterns to show that some conventional form-meaning pairings occupy the middle ground in the lexicon-grammar continuum.

  - They’re not engaged, let alone married. (Fillmore et al. 1988)
  - Waiter, what is this fly doing in my soup? (Kay & Fillmore 1999)
  - The richer the people, the bigger the crates they erected. (Kay et al 1999)
  - She elbowed her way through the crowd. (Goldberg, 1995, Israel 1996, Perek 2016, 2020)
CxG: how it is going

patterns that are more canonical and less idiomatic have received less attention because they are potentially analyzable in terms of general syntactic mechanisms, and therefore not necessarily describable in terms of constructions

some exceptions:

- *is to* (Goldberg & van der Auwera 2012)
- *may* (Hilpert 2016)
- *it BE ADJ to V_{inf} that* (Desagulier 2021)

even when such generalizations are available a CxG reading of analyzable patterns offers a deeper, usage-based account of their conceptual and functional complexity
case study:

4 prepositional patterns that denote a relation of internal location between a located entity (a trajector) and a reference entity (a landmark).

(1) He stops suddenly in the middle of the stage and seems to consider. (COHA, 1815–FIC–FalseShame)
(2) Marvin walked to the chalk mark in the center of the ring. (ibid., 1934–FIC–CaptainCaution)
(3) (...) we were in the heart of Norwalk. (ibid., 1827–FIC–Novels)
(4) We see St Eustace praying in the midst of the river. (ibid., 1980–FIC–RiddleyWalker)

the Prep NP\textsubscript{IL} of construction
Grea (2017): *sein, milieu, parmi, centre, coeur*

- **AU SEIN DE**: bounded lm
- **AU MILIEU DE**: coincidence between tr and lm
- **PARMI**: plural lm
- **AU CENTRE DE**: functional dependence between tr and lm
- **AU CŒUR DE**: intensive lm
the debate around complex prepositions

- complex prepositions are attested in 19th- and 20th-century grammars (Hoffmann 2005: 26):
  - ‘phrasal prepositions’ (Earle 1873),
  - ‘prepositional formulæ’ (Mätzner 1873),
  - ‘group prepositions’ (Sweet 1892).
- some P-NP-P sequences function like prepositions.
- yet, the full preposition status is denied because of F/M mismatch
complex prepositions do not exist

- Huddleston (1988: 126)
  - Prep1 NP Prep2: a “mismatch” between the lexicon & grammar bc of lexicalization
  - e.g. for the sake of X: coalescence due to lexicalization but in fact: [for] + [the sake of X]
- Huddleston & Pullum (2002: 618)
  - Prep1 NP Prep2: “idiomatic and fossilized expressions headed by a preposition”
    - idiomatic = non-compositional
    - fossilized = lexicalized
  - syntactic constraints: not all syntactic manipulations are allowed
complex prepositions do not exist

has “some intuitive appeal” (Huddleston and Pullum 2002)

complex-preposition analysis
complex prepositions do not exist

the NP *the stage* is considered the complement of the head preposition *in*.

This analysis is favored if the expression alternates with a genitive construction

*in the middle of X → in its middle*
complex prepositions do not exist

The expression is divided into the head PP *in the middle* and the complement PP *of the stage*

This analysis is chosen if fronting is possible

*the stage in the middle of which he was*

layered-head analysis
complex prepositions do not exist

• Seppänen et al. (1994)
  • concede that some Prep1 NP Prep2 expressions display a certain degree of syntactic ‘fixity’ and idiomaticity. Yet, the preservation of conservative syntactic parsings prevails
  • 4 constituency tests
    - fronting: He was in the middle of the stage > Of what stage was he in the middle?
    - coordination: he was in the midst of his friends and of old acquaintances
    - ellipsis: A: He was in the middle of the stage. B: of a stage? He is not even an actor!
    - interpolation: He was in the middle, or perhaps center, of the stage.
  • one successful test is enough to rule out the constituent status of complex preps
  • no gradient as far as constituency is concerned

• Hoffman (2005, 47): such tests yield sentences that are highly artificial and hardly ever used, based on the BNC

• Beckner & Bybee (2009): chunking and categorization have semantic effects and change incrementally over time.
complex prepositions exist

• Kruisinga (1931, 345–346): a gradient of morphosyntactic unity

• Quirk et al. (1985, 671):
  ▪ “[i]n the strictest definition, a complex preposition is a sequence that is indivisible both in terms of syntax and in terms of meaning.”
  ▪ “there is no absolute distinction between complex prepositions and constructions which can be varied, abbreviated, and extended according to the normal rules of syntax”

• Based on Quirk and Mulholland (1964), Quirk et al. (1985) develop a scale of ‘cohesiveness’
the Prep $NP_{IL}$ of construction

$NP_{tr} \ldots$ Prep $NP_{IL}$ of $NP_{lm}$

'\text{tr is included in lm}'

form

meaning
the Prep NP\textsubscript{IL} of construction network
research questions

• has Prep NP_{IL} of undergone constructional change?
  more productive?
  more schematic?
  (Perek 2020)

• what are the functional constraints at work at the (sub)schematic levels?
data & usage

• all instances of the $\text{Prep NP}_{IL} \аЁ construction were extracted from the 406M-word COHA (Davies 2010).

• manual filtering to remove the following:

(5) With all his singularities, there was in the heart of Uncle Tim a depth of religious sincerity (...) (COHA, 1843-fic-MayflowerSketches)

• 15,635 tokens of the construction, along with their $\text{NP}_{lm}$ collocates

• “prepositions are indifferent as to the nature of their trajector”(Langacker 2012, 117): trajector NPs were ignored.
data & usage

relative token frequencies per million words

same w/ locally weighted smoothing
productivity: vocabulary growth curves

• productivity is generally measured with:
  • type frequency (V): the number of distinct occurrences
  • hapax legomena (V1): the number of types that are realized by unique occurrences

• problem: the decades vary in token size, which means we cannot compare decades fairly

• we use vocabulary growth curves instead: the data are scanned, and V & V1 are recorded cumulatively at regular intervals
productivity: vocabulary growth curves

vocabulary growth curves (Baayen 1989, 1992, 1993; Evert & Baroni 2007)
interim answer

• has Prep NP_{IL} of undergone constructional change?
  more productive ⬗
  more schematic  ?

• schematicity
  the more semantically diverse the NP_{im} collocates are, the more schematic the constructional meaning
from count-based to predictive DSMs

- 3 computational implementations of the distributional hypothesis (Harris, 1954; Firth, 1957) in CxG:
  - Hilpert (2016) may, COCA & COHA, 3 periods (1800s–1860s, 1870s–1920s, 1930s–1990s)
  - Perek (2016) the hell-construction, COCA & COHA, 4 periods (1930s — 1940s, 1950s — 1960s, 1970s—1980s, and 1990s—2000s)
  - Perek (2018) the way-construction, COHA, 4 periods (1830—1870s, 1880s—1920s, 1930s — 1960s, 1970s — 2000s)

- the vector space is then augmented with diachronic frequency information (words are plotted based on period-relevance, or density curves are plotted)
from count-based to predictive DSMs

• Hilpert (2016) and Perek (2016, 2018) use DSMs that rely on count models (whose vectors are generated from co-occurrence counts).

• The resulting vectors are long (with as many dimensions as there are collocates) and sparse (most of their cells are zeros). This is likely to affect the quality of the vector representations, even after weighting and dimensionality reduction have been applied (Baroni et al. 2014).

• SGNS, one of the two predictive models of the word2vec toolkit (Mikolov, Yih, et al. 2013), outperforms PPMI and SVD in the discovery of new shifts and the visualization of changes (Hamilton et al. 2016).

• I adopt SGNS, which predicts a word’s context given the word itself.

• word meanings are represented in the form of short, dense vectors (series of numbers based on the distributions of the words).

• despite limitations (Desagulier 2019, Fonteyn 2021), word vectors can approximate meaning.
Building the reference vector space

- I further trained word2vec on COHA
  - hyperwords (Levy et al. 2015) > SGNS > cds 0.75; win ±5; neg 15; dim 300 (Levy & Goldberg 2014); all words whose frequency is less than 5 in the COHA were discarded
- I extracted the vectors of the most distinctive NP_{lm} of the Prep NP_{IL} of NP_{lm} construction (G^2 >= 10)
- I reduced the 300-dim vectors two 2-dim vectors with t-SNE (Maaten & Hinton 2008)
- I projected these NP_{lm} collocates onto a 2D map
Projecting diachronic information

• data divided into four arbitrary periods:
  period 1: 1810s–1860s
  period 2: 1870s–1910s
  period 3: 1920s–1970s
  period 4: 1980s–2000s

• density curves were projected for each subconstruction and each period
• the denser the curves, the higher the frequency in the whole COHA
Prep the MIDST of

• originally: **space-related coincidence**

(5) As a sort of mockery of the valuables, there was a wooden platter in **the midst of the table**, containing a crust of bread and a red herring. (COHA, 1867–FIC–Ambrose Fecit)

(6) in the **midst of the forest** (Webster, https://www.merriam-webster.com/dictionary/midst)

(7) at a ruin in the **midst of the forest** he was traversing, (COHA, FIC—1839—CaptainKyd)
Prep the MIDST of

- increasing preference for NPs$_{lm}$ that are compatible with an intensive reading

(8) indeed, be supposed that a youth, reared in the midst of boundless forests (COHA, 1832—NF/ACAD—ProgressGenius)

(9) she had lived at a remote "township," in the midst of Canadian forests; (COHA, 1843—FIC—IdomenValeYumuri)

(10) Wretched villages, which I had left in the midst of impenetrable forests, had assumed the appearance of flourishing towns. (COHA, 1826—MAG—NorthAmRev)
Prep the HEART of

• a clear shift towards NPs$_{lm}$ that invite an intensive reading

(11) Her condominium was located midway up in a high-rise building in the heart of the city.
(12) a so-called "' floating palace " on the river in the center of the city (…)

![Diagram](image-url)
Prep the HEART of

• a clear shift towards NPs\textsubscript{lm} that denote conflicts and disputes

(13) But in its crude way, the incident touched on issues and emotions -- power, pride and humiliation -- that lie at the heart of the conflict between Arabs and Jews. (COHA, 1992—NEWS—WashPost)
Prep the MIDDLE/CENTER of

- favor NPs\(\text{lm}\) that denote bounded places (literally) or that can serve as a locational reference point
- differences?

(14) Don't leave it there in the middle of the floor, " she said sharply. incongruous presence
(15) All eyes within were focused on a couple waltzing in the center of the floor to low music. exact loc°
Prep the MIDDLE/CENTER of

- favor NPs$_{lm}$ that denote bounded places (literally) or that can serve as a locational reference point
- differences?

(16) in the middle/center of a chapter/scene/sentence/tirade
interim answers

• has Prep NP<sub>IL</sub> of constructionalized over time?
  more productive ✓
  more schematic
  • midst ✓ (except it goes the other way around)
  • heart ✓ (very clearly)
  • middle (less clear)
  • center (less clear)
What is left to do

• a more systematic period-wise comparison of distinctive collocates (especially when these appear in several constructions)
• a more systematic comparison to French equivalents (Grea 2017)
• investigating whether $\text{Prep NP}_{\text{IL}} \text{ of}$ has undergone constructionalization
What I hope to have shown

• we have reasons to believe that Prep NP\textsubscript{IL} of counts as a construction
• F-M mappings reconfigure at the subschematic level (midst & heart)
• on the one hand, one of its subconstructions has gained in terms of productivity and schematicity (heart)
• the three other constructions (midst, middle, center) display more conservative distributions over time
Thanks for your attention!

https://corpling.hypotheses.org/3532