Visualising Changes in Semantic Neighbourhoods of English Noun Compounds over Time

Malak Rassem, Myrto Tsigkouli, Chris Jenkins,
Filip Miletić and Sabine Schulte im Walde

University of Stuttgart
Germany

Institut für
Maschinelle
Sprachverarbeitung

Motivation & Contributions

- **Noun compounds (NCs)**: a type of multiword expressions (MWEs) whose meanings are semantically idiosyncratic to some degree, i.e., not necessarily fully predictable from the meanings of their parts
  - *fairy tale*
  - *gold mine*
- **Perspective**: diachronic development of NC compositionality
- **Contributions**:
  - carefully crafted **semantic vector space** to represent compounds and their constituents across time slices of the cleaned corpus of historical American English CCOHA (Davies, 2021; Alatrash et al., 2020)
  - **semantic neighbours** for compounds and their constituents, relying on both (i) time-specific and dynamic as well as (ii) static present-day representations
  - **temporal compound–constituent visualisation tool**: adaptation of deterministic multi-dimensional scaling and two-dimensional plotting (Hilpert, 2016)

Data

1. **Corpus: CCOHA**
   - cleaned version of the Corpus of Historical American English
   - reduced, coarse-grain part-of-speech set
   - segmentation into timeslices: 1810–1829, 1830–1859, 1860–1889, 1890–1919, 1920–1949, 1950–1979, 1980–2009
2. **Noun compound targets**
   - noun-noun compounds from Cordeiro et al. (2019)
   - 195 compounds and their constituents occurring in all time slices of CCOHA

Semantic Space

- **Goal**: identify an appropriate set of vector-space dimensions
  → regarding the semantic interpretations of the dimensions
  → regarding the notorious sparse-data problem in historical corpus data
- **SSPs (semantic space points)**:
  - nouns appearing with a frequency >500 in the entirety of the CCOHA, i.e., not just within individual timeslices
  - exclusion of top 50 most frequent nouns to eliminate potential semantic hubs
  - resulting SSPs: 9,345 unique nouns
- **TSCs (timeslice-specific co-occurrences)**:
  - co-occurrence frequencies within a ±10-word window
  - only content words: nouns (NN), verbs (VV), adverbs (RR), adjectives (JJ)
  - transformation into vectorised formats

Semantic Neighbourhoods

- **Semantic proximity**: comparison of TSC vector representations of NCs to those of SSPs within the same timeslice, using *cosine*
- **Semantic neighbours**: five most similar SSPs
- **Static semantic space**: TSC vectors of semantic neighbours from last timeslice

Target	Timeslice	Five Nearest Neighbours
credit card	1830–1850	—
	1920–1940	rationing, gallon, shuttle, questionnaire, invitation
	1980–2000	reservation, card, cash, credit, check
credit	1830–1850	exchange, money, bank, account, circulation
	1920–1940	loan, bank, account, banker, reserve
	1980–2000	card, visa, account, cash, greeting
card	1830–1850	game, paper, trick, minute, stranger
	1920–1940	paper, game, ball, box, trick
	1980–2000	check, credit, paper, line, trick

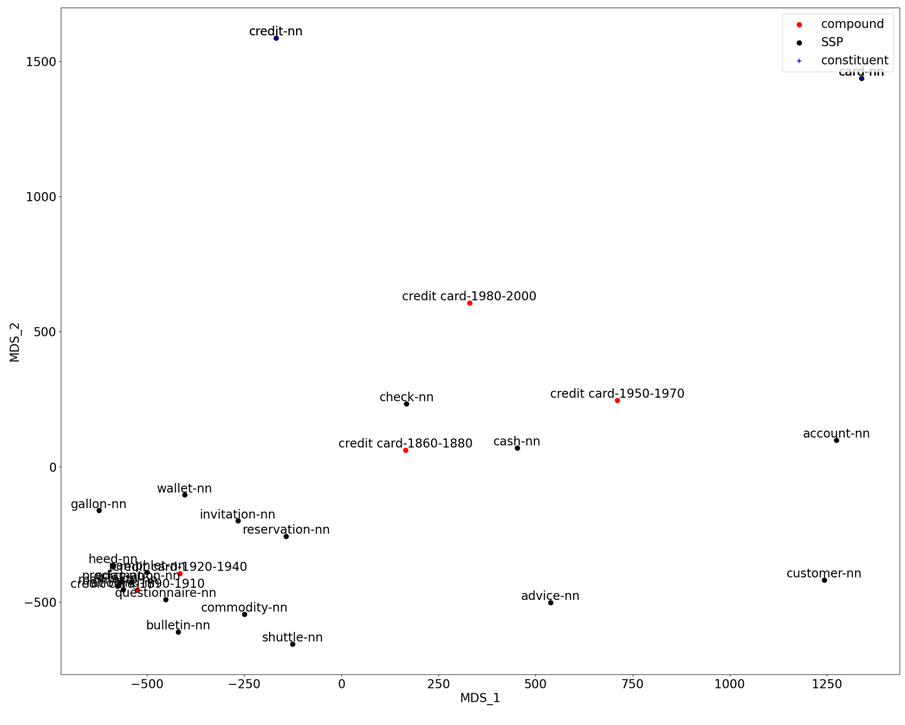
Target Compound	Timeslice	Five Nearest Neighbours
field work	1830–1850	fugitive, wedding, afternoon, vogue, feast
	1920–1940	field, career, imagination, knowledge, fashion
	1980–2000	program, instructor, project, training, success
food market	1860–1880	chancery, jurisdiction, decision, appeal, judge
	1920–1940	parity, commodity, rise, boost, sell
	1980–2000	supermarket, grain, food, specialty, convenience
ghost town	1920–1940	suburb, inhabitant, dweller, resident, outskirts
	1980–2000	town, ruin, village, tourist, neighborhood
rat race	1920–1940	novel, autobiography, estate, poem, biography
	1980–2000	stuff, kid, trouble, folk, idea

Diachronic Visualisation of Time-Specific Compounds in Semantic Space

1. **Own-compound approach**
   - TSC vectors of compound for every timeslice
   - TSC vectors of constituents and neighbours only from the last "static" timeslice
   - application of metric multidimensional scaling (MDS)
   - results in sub-optimal plots, where compounds tend to cluster together and away from the SSPs regardless of timeslice
2. **Projected-compound approach**
   - difference: exclude compound's own TSC vectors
   - instead: weighted averages of respective five time-specific nearest neighbours' coordinates; weights: cosine scores
   - effect: reflect compound's relative positions to its neighbours' semantic fields
   - results in better illustration of temporal semantic NC shifts

→ `https://www.ims.uni-stuttgart.de/data/dia-neighbour-nn`

Visualisation of compound *credit card* over time



Visualisation of compound *gold mine* over time