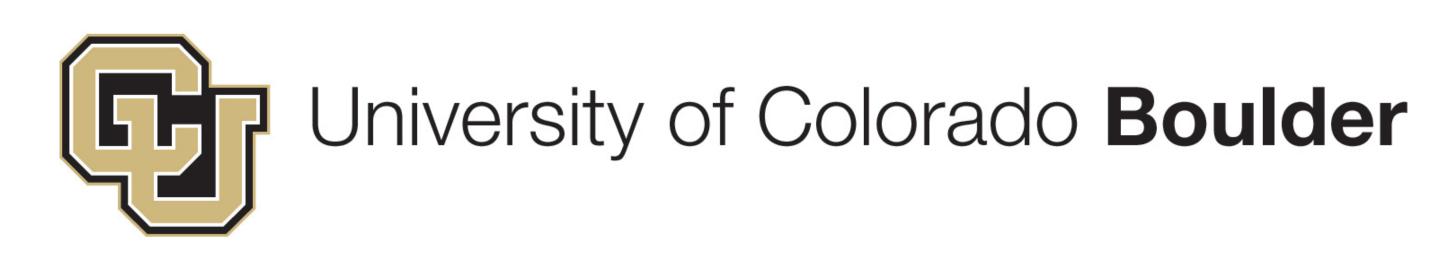


https://github.com/dchensta/adpositions_case



My Case, For an Adposition: Lexical Polysemy of Adpositions and Case Markers in Finnish and Latin

Daniel Chen, Mans Hulden

Creation of SNACS Corpus

- Adpositions and case markers cover wide range of semantic fields: grounded spatial relations, abstract causal relations
- Useful to create annotation guidelines for disambiguating different senses of tokens using context of sentence

SNACS - Semantic Network for Adposition and Case Senses (Schneider et al 2018)

- Supersense hierarchy: four levels of granularity
 - Previous languages with existing SNACS corpora: English, German, Korean, Hindi, Mandarin Chinese
- New SNACS corpus for Finnish and Latin: languages with rich case systems
 - Source texts: translations of Le Petit Prince
 - ~30 unique scene roles for each language



K-Means Clustering Analysis

 All adpositions and case-marked words are represented by Multilingual BERT (BERT base, uncased) embeddings generated from automatically tokenized WordPieces.

In tensor([0.0778, -4.9119, 4.1403, 2.4959, -3.2524])
##bus tensor([-1.8946, -1.4443, 3.4854, 4.3752, -1.0381])
specie tensor([-2.1599, -2.1298, 3.4818, 4.2119, -0.9580])
##bus tensor([-5.1346, -1.0579, 3.8172, 5.0954, 1.2067])
##e tensor([-2.8429, -1.3406, 3.0980, 1.3174, 1.5849])
##ibus tensor([-3.9265, -2.2255, 0.4299, 2.7338, 2.8088])
##o tensor([-2.5140, -0.3774, 0.2082, 3.9227, 3.1609])

- Concatenate the last 4 hidden layers => each WordPiece token has 3,072 dimensions.
 - Achieves higher cosine similarity between tokens
- Use principal component analysis (PCA) to reduce dimensionality to 100 dimensions, for more efficient processing.
- K-Means Clustering Algorithm: identifies k (user-specified) cluster centers (prototype / mean of data points), assigns each data point to a cluster.

Construal Analysis

Finnish

152 adposition + case marker annotations for *Pikku Pinssi*

Sentence (from Pikku Prinssi, Chapter IV)	Scene Role	Function Role
a. Ystävä-ni lähti pois lampai- <u>ne</u> -en. friend-1P.POSS went away sheep-COMITATIVE-3P.POSS My friend went away with his sheep.	ANCILLARY	ANCILLARY
b. Erehdyn myös vähän pituude- <u>ssa.</u> I err also a little length-INESSIVE I also err a little in/regarding length.	TOPIC	LOCUS

Latin

180 adposition + case marker annotations for *Pikku Pinssi*

Sentence (from Regulus, Chapter IV)	Scene Role	Function Role
a. Quotientscumque cum eis As_often_as with them.ABL.PL	RECIPIENT	ANCILLARY
de nov-o amic-o loque-ris About new-2.ABL.SG friend-2.ABL.SG talk-2P.PRES.SG.ACT As often as you talk with them about the new friend	TOPIC	TOPIC
b mihi admiration-em magn-am movere non poterat me.DATIVE admiration-3F.ACC great-3F.ACC to_move not was_able.3P.PSThe was not able to move great admiration for me.	EXPERIENCER	GOAL

Results + Analysis

# Clusters	Cluster
50	##assa(36), ##uksessa(38), ##ssa(127), ##ssa(167), ##sina(192), ##ssa(233), ##na(268), ##ssa(314)
30	##ksi(22), ##assa(36), ##alla(53), ##oilla(93), ##lla(96), ##sena(111), ##ssa(127), ##a(136),
	##ssa(167), ##lta(189), ##sina(192), ##lla(217), ##lla(223), ##a(227), ##ssa(233), ##lla(243),
	##la(258), ##na(268), ##ssa(314)
15	##kella(17), ##ksi(22), ##ulla(25), ##kella(30), ##assa(36), ##uksesssa(38), ##alla(53), ##oilla(93),
	##lla(96), ##sena(111), ##ssa(127), ##a(136), ##a(155), ##uudessa(157), ##ssa(167), ##lta(189),
	##sina(192), ##lla(217), ##lla(223), ##kill(226), ##a(227), ##ssa(233), ##lla(243), ##nna(248),
	##llaan(253), ##la(258), ##na(268), ##tajana(290), ##ssa(314)

- 1. Adpositions cluster well together, since they are separate words that typically make up the entirety of their own WordPiece token.
- 2. Adpositions are found in separate clusters from their case-marked noun objects => mBERT identifies no semantic similarity between them.
- NP P
 No muusta kuin

"except" in Finnish

- 3. Multiword adpositions are not found in the same cluster.
- 4. Clusters do not group morphological alternations of the same case marker together.
 - Finnish: elative case markers -sta and -stanne are grouped in different clusters
 - Latin: first declension ablative singular *-ia* is with other ablative singular endings, but second declension ablative singular *-ulo* is with noun stems.
 - => mBERT cluster partitions are strictly orthographical. Can't quite begin to account for semantic variation when morphological variation is not accounted for.