Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 2610–2616
Marseille, 20-25 June 2022

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

Daniel Chen, Mans Hulden
University of Colorado
{daniel.chen-1, mans.hulden}@colorado.edu

Abstract

Adpositions and case markers contain a high degree of polysemy and participate in unique semantic role configurations. We present a novel application of the SNACS supersense hierarchy to Finnish and Latin data by manually annotating adposition and case marker tokens in Finnish and Latin translations of Chapters IV-V of *Le Petit Prince* (*The Little Prince*). We evaluate the computational validity of the semantic role annotation categories by grouping raw, contextualized Multilingual BERT embeddings using k-means clustering.

Keywords: corpus creation/annotation, evaluation methodologies, morphology, semantics, parsing, grammar, syntax, treebank, lexical semantics, adpositions, case markers, polysemy, Latin, Finnish, computational semantics, natural language processing, embeddings, k-means clustering

1. Introduction

Crosslinguistically, adpositions are among the most polysemous word forms in a grammar. **Polysemy** refers to the semantic phenomenon where one form maps to multiple meanings that are interrelated but not necessarily synonymous. Because of their high polysemy, a singular adposition can cover a wide range of semantic fields while occupying the same syntactic context. They describe both grounded spatial relations and abstract causal relations; they occupy both core[1] semantic roles and non-core semantic roles. Adpositions cannot be trivially substituted for each other, and their lexical semantics are heavily informed by contextual dependencies with their **governor** and **object** (Srikumar and Roth, 2013), or head and dependent.

For case-marked languages, case markers on nouns can individually or jointly occupy the same semantic roles that an adposition individually represents in other languages, such as English. Latin and Finnish are case-marked languages that use both adpositions and case markers, but in three different semantic configurations: 1) the case marker solely represents a semantic role (typically one of the case marker's prototypical roles), 2) the case marker and adposition both represent the same semantic role, 3) the adposition represents a semantic role differing from the case marker's prototypical role: together they convey a gestalt semantic role.

To account for lexical polysemy and identify semantic configurations of adposition and case marker tokens, we annotate semantic roles for those tokens using version 2.5 of the English annotation guidelines of the supersense hierarchy Semantic Network for Adposition and Case Senses (SNACS) (Schneider et al., 2018; Schneider et al., 2020b). In this paper, we present a novel application of SNACS to Finnish and Latin translations of *Le Petit Prince (The Little Prince)* by creating a pilot annotation corpus[2] consisting of Chapters IV and V, two chapters that have existing SNACS annotations (Schneider et al., 2022) in English, German, Korean, Hindi, and Mandarin Chinese for accessible comparison. SNACS rules have previously not been developed for Latin or Finnish.

The pilot annotations are an important step in documenting the complex semantic collocations between case markers and adpositions in these two languages, both of which possess rich case systems. The corpus also provides a clear foundation for developing computational models that can capture the combinatorial semantic properties of highly polysemous adposition and case marker tokens. In the latter half of the paper, we present an experiment analyzing clusters of raw, contextualized Multilingual BERT embeddings (Devlin et al., 2019) of adpositions and words containing case markers, derived through k-means clustering. By examining the extent to which Multilingual BERT groups tokens together according to linguistic features, the experiment evaluates the applicability of a fine-grained semantic annotation schema to a computational model.

(1) Case Marker + Adposition	
*puku-**n**-sa*	***takia***
attire-GEN.SG-3P.POSS	because.of
because of his attire	
(2) Case Marker Only	
*kaukoputke-**lla***	
telescope-ADESSIVE.SG	
with a telescope	

Table 1: Glosses of Annotation Scenarios in Finnish

[1] Core semantic roles fill argument structure slots of a main verb like *give*. For example, a prepositional phrase like *to me* fills the RECIPIENT role in the English ditransitive construction, as in "Give the book to me" .

[2] Annotated corpus and code can be found at `https://github.com/dchensta/adpositions_case`

2. Annotation Methodology

Since Latin and Finnish both use a combination of adpositions and case endings, we annotated the following scenarios, as depicted in Table 1: (1) combination of adposition and case-marked object (2) case-marked word with no neighboring (i.e. preceding or succeeding) adposition. In the first scenario, both the case-marked object and adposition both receive an identical scene role annotation, although their function roles can differ. Adjectives modifying these token types were not included, as they cannot be considered a governor or object. The exception is when an adpositionally marked adjective has no noun object present at all and an implicit noun object along the lines of "thing", "person", or "place" can be construed, e.g. the Latin dative plural word *ignotis*, meaning "to the unknown (places)".

Words marked by cases specifying core semantic roles like AGENT and PATIENT, such as nominative and accusative in Latin and nominative, partitive, and genitive in Finnish, were not annotated, excluding situations where an adposition pairs with a word marked by one of these core cases[3]. More commonly, adpositionless, case-marked words marked with one of these prototypical semantic roles were not annotated in order to focus on the typically non-core semantic roles covered by SNACS. The exception is when a verb like *päässyt* (päästä 1P.SG.PST.CONNEG 'to arrive at') in Finnish requires a specific case (here, allative case) for the object carrying the core role of THEME in the verb's syntactic frame.

2.1. Construal Analysis

The defining hallmark of SNACS annotation is the **construal analysis** (Hwang et al., 2017). Rather than force annotators to adjudicate a single label to assign to an adposition, Hwang et al. (2017) specified a second category to allow for dual annotations of the same token. The two annotations are not required to align, allowing for an adposition or case marker to express multiple dimensions of semantic information reflected by other lexical items in the sentence.

The first annotation is the **scene role**, which corresponds to the most direct contextual role that the adposition or case-marked word plays in the overall semantic space of the sentence. The assigned semantic role specifies the relationship between the governor and object, two entities linked by the adposition or case marker. The second annotation is the **function role**, which delineates the semantic contribution that the adpositional or case marker token itself provides to the overall sentence meaning.

Table 2 depicts the application of the SNACS construal analysis to English, Finnish, and Latin data. For each sentence pair, sentence (a) shows an example of construal analysis in which the scene role and function role are congruent, and sentence (b) shows an example in which they are incongruent. The congruent examples often label the tokens with their most prototypical semantic function: LOCUS for the English preposition *in* in (1a), ANCILLARY for the Finnish comitative case marker *-en* in (2a), and LOCUS for the Latin preposition *in* in (3a).

In sentence **(1b)**, the preposition *through* received a SNACS annotation of LOCUS->PATH, where "the forest" is both the metaphorical PATH of the "runs" event and the static LOCUS (location) of the subject "the road" in the global context of the sentence. This is because a road is not an animate subject that can perform a running event, but an inanimate subject that metaphorically traverses through the forest.

The metaphorical traversal is the semantic content offered by the prepositional token itself, i.e. the function role. The preposition *through* inherently encodes a PATH sense, where PATH is defined as "the ground that must be covered in order for the motion to be complete" (Schneider et al., 2020b). However, the overall function of *through* in sentence **(1b)** is to position the road as being located inside a forest, given that the running motion of the road is metaphorical. This is a clear example of how construal analysis can display the incongruency and shared semantic information of metaphorical language.

In **(2b)**, the inessive case in Finnish inherently encodes a locational sense corresponding to "in" in English. However, since the verb governor is *erehdyn* (erehtyä 1P.SG.PRES 'to err'), the meaning of the inessive case marker *-ssa* can be construed as describing the TOPIC that the speaker is making a mistake in, such that the English prepositional translation can more literally read "I err regarding / with regard to length".

In **(3b)**, Latin often uses a dative form to mark possession if the POSSESSOR is a pronoun. The prototypical role of the dative case is to be a RECIPIENT, or indirect object of a sentence like "Sansa gave Jon a warning.", where Jon is the indirect object that would be marked by the dative case. Thus, *mihi*, 'my, lit. for me', in (3b) is assigned the function role RECIPIENT, which is typical of the dative case, but noted to have a scene role of POSSESSOR, given its status as a pronoun describing the owner of the *nomen*, 'name'.

2.2. Specifications of Corpus

Language	# Annotated Tokens
Finnish	152
Latin	180

Table 3: Number of Annotated Case Marker and Adposition Tokens Per Language

[3]In Latin, the preposition *in* followed by an accusative-cased noun like *urbem* (city) does not mark a core role of THEME, but uses the combination of the preposition and accusative-cased word to mark a non-core role of GOAL

Sentence				Scene Role	Function Role
(1) English					
a. *The road is*	***in***	*the forest.*		LOCUS	LOCUS
b. *The road runs*	***through***	*the forest.*		LOCUS	PATH
(2) Finnish					
a. *Ystävä-ni*	*lähti*	*pois*	***lampai-ne-en***.	ANCILLARY	ANCILLARY
friend-1P.POSS	went	away	sheep-COMIT-3P.POSS		
My friend	went away	with his	sheep.		
b. *Erehdyn*	*myös*	*vähän*	***pituude-ssa***	TOPIC	LOCUS
I err	also	a little	length-INESSIVE		
I also err	a little	in/regarding	length.		
(3) Latin					
a. ***In***	*agr-is*	*ambulav-i.*		LOCUS	LOCUS
in	field-ABL.PL	walk.PERF-1P			
I walked	in the	fields.			
b. *Nomen*	***mihi***	*est*	*Davos.*	POSSESSOR	RECIPIENT
name.NOM.SG	I.DAT.SG	is	Davos		
My name	is Davos				

Table 2: Construal Analysis in English, Finnish, and Latin

Table 3 depicts the number of case markers and adpositions that were annotated, per language. Table 4 depicts the subset of SNACS scene role labels that are covered by the corpus, while Table 5 depicts the subset of function roles.

In general, the function role assignments for Latin were more difficult due to the higher degree of polysemy covered by only 6 cases in Latin, versus 15 cases in Finnish. Thus, function roles for adpositions in Latin had clear standard meanings for the adposition token, whereas function roles for case markers required consideration of the global context, especially for cases like the ablative case that have several inherent / functional standard interpretations that develop even more complex interpretations for their ultimate scene role.

For this corpus, while we annotated for both scene and function role according to construal analysis, we acknowledge that the scene roles are more unambiguous than function roles for case-marked words, given the difficulty of assigning inherent meanings to languages like Latin and Finnish that possess rich case systems.

Scene Roles
GOAL, TOPIC, POSSESSOR, PARTPORTION, INSTRUMENT, QUANTITYITEM, RECIPIENT, EXPERIENCER, SOURCE, LOCUS, EXPLANATION, BENEFICIARY, WHOLE, STUFF, IDENTITY, STIMULUS, MANNER, ANCILLARY, COMPARISONREF, THEME, MEANS, TIME, AGENT, POSSESSION, CHARACTERISTIC, PURPOSE, CAUSER, ORG, GESTALT, DURATION, SOCIALREL, REFERENCE, COST, PATH, SPECIES

Table 4: SNACS scene roles assigned to the corpus of Finnish and Latin translations of *Le Petit Prince*

Function Roles
GOAL, SOURCE, POSSESSOR, PARTPORTION, INSTRUMENT, LOCUS, EXPLANATION, IDENTITY, RECIPIENT, MANNER, COMPARISONREF, INDEFINITE OBJECT (not in original SNACS), ANCILLARY, PATH, STUFF, AGENT, MEANS, TIME, CAUSER, GESTALT, TOPIC, DURATION, REFERENCE, COST, CHARACTERISTIC, BENEFICIARY, ACCUSATIVE OBJECT

Table 5: SNACS function roles assigned to the corpus of Finnish and Latin translations of *Le Petit Prince*

3. Application of SNACS Annotations

In the remainder of the paper, we present an experiment that aligns the linguistic groupings of adposition and case marker SNACS roles with computational representations of those tokens.

3.1. Related Work

Liu et al. (2019) conducted seventeen probing tasks to identify the linguistic knowledge captured by contextual word representations in neural network and transformer models like BERT. One of these tasks was a preposition supersense disambiguation task trained on the English STREUSLE 4.0 corpus (Schneider et al., 2020a), which houses SNACS annotations. Liu et al. achieved a performance higher than the then state-of-the-art by using a linear probing model to run preposition supersense disambiguation using BERT embeddings, improving from 66.89 to 79.61 for scene role annotation and from 78.29 to 90.13 for function role annotation. This was a supervised task showing the usefulness of SNACS annotations in NLP architecture.

Many linguistic evaluations of contextualized word embeddings tend to focus on deep learning models' acquisition of syntactic structures. Tenney et al. (2019) probe for sentence structure across syntactic, semantic, local, and long-range phenomena, and found that while syntactic representations were strong, semantic tasks like semantic role labeling did not benefit greatly from contextualized representations, like the word embeddings that BERT employs. Chi et al. (2020) ran an unsupervised clustering task on the multilingual mBERT, discovering that mBERT natively represents syntactic dependency labels that agree with the Universal Dependencies labels prescribed by linguists. This indicates that semantic minimal pairs for polysemous tokens represent a significant challenge for contextualized representations like mBERT, since adpositions often occupy identical syntactic environments, as do case markers.

3.2. Generating Contextual Embeddings for Alignment Task

To test for computational alignment of adposition and case marker tokens with their SNACS roles, we generated embeddings for each token of interest. Ideally, the contextual embeddings of BERT transformer models can intrinsically collocate the semantic functions of adpositions and case markers, in the same way a linguistic schema like SNACS groups its supersenses.

Each translation of Chapters IV and V of *The Little Prince* was fed into uncased[4] Multilingual BERT (mBERT) sentence by sentence. BERT transformer models are notable for using contextualized word embeddings, which encode sentence-specific context for each word, in a departure from static word embeddings like GloVE (Pennington et al., 2014) and word2vec (Mikolov et al., 2013). The mBERT model (Devlin et al., 2019) is trained on 104 languages, including Latin and Finnish. We chose this model to test the crosslingual applicability of the SNACS supersense hierarchy, whose semantic role labels theoretically are universally applicable to typologically diverse languages.

mBERT does not generate word embeddings per se, but splits each whitespace-separated word into **WordPieces**: morphemes of a word that are automatically segmented by the transformer. This means that some of the morphemes that receive embeddings do not necessarily line up with the adpositions and case marker morphemes. For example, mBERT splits up the Finnish postposition *paitsi* into the WordPieces *pai* and *##tsi*, but another Finnish postposition *kohti* remains intact as a single WordPiece. Thus, alignment is not necessarily a one-to-one process or consistent across tokens with the same POS tag.

However, WordPiece tokenization does allow for convenient analysis of case marker suffixes, which tend to be segmented cleanly and presumably carry the bulk of the adpositional or case-marked meaning. For example, the allative case marker in Finnish, *-lle*, produced a neat cluster of *##lle* WordPiece segments, with some variation of vowels preceding the *-lle*. In the case of prepositions like *paitsi*, which was segmented into *pai* and *##tsi*, we chose to analyze the last segment[5] for alignment, to avoid collocations of a stem like *pai* with noun stems.

For the purposes of this experiment, we chose to analyze only scene role annotations, despite having performed the full construal analysis on the entire corpus. Since we want to align senses that are shared amongst disparate adpositions and case markers, a pair of embeddings with higher cosine similarities should reflect shared scene roles. To provide an English example, an *in* token corresponding to the LOCUS scene role should align to a different cluster than an *in* token corresponding to the TOPIC scene role.

3.3. Clustering Methodology

After running tests that revealed higher cosine similarity scores between WordPiece embeddings that were created through concatenation rather than summation, we chose to concatenate the last four hidden layers. This resulted in each WordPiece embedding being represented by a vector with 3,072 dimensions, the result of the concatenation of four vectors of 768 dimensions: the length of a standard hidden state vector.

The k-means clustering algorithm is an unsupervised machine learning task that takes in a set of data points as input and groups each data point into a user-specified amount of clusters. The scene roles in the SNACS hierarchy were designed to not require global knowledge of all other scene role types to produce an annotation belonging to that scene role category, which describes specific semantic variations that warrant a unique category label. We do not control for granularity, since more fine-grained categories are unique enough (e.g. POSSESSOR and WHOLE both are subsumed by the coarser-grained GESTALT, but each category possesses different animacy obligations and would thus co-occur with different semantic classes of nouns) to be treated as separate categories for the purposes of a baseline clustering task.

The number of unique scene role annotation labels we assigned for both the Latin and Finnish texts was roughly 30 scene roles per language. There are 50 unique SNACS supersenses, so theoretically, mBERT should be allowed the entire range of options to perform its clustering. For both languages, we slowly decreased the number of clusters from 50 to get closer approximations to the mBERT WordPiece embeddings' semantic clustering.

[4]We chose to use an uncased model to prevent cased adpositional words being assigned different embeddings from uncased equivalents, e.g. *In* versus *in*. By uncasing all text, this misleading distinction is avoided.

[5]Luckily, most Finnish and Latin adpositions did not receive spurious segmentations like *paitsi*, given that many Finnish and Latin adpositions are quite short in length (e.g. the preposition *ad* in Latin).

# 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)

Table 6: The three clusters that the inessive case marker *-ssa* in *yksityiskohdi-ssa* is assigned to when varying the number of clusters from 50, to 30, to 15. The indices in parentheses correspond to the index of the WordPiece token in the list of adpositions and case-marked words manually isolated from the text.

Since the mBERT token embeddings possess extraordinarily high dimensionality, we used principal component analysis (PCA) to lower the vector size from 3,072 to 100 dimensions to allow for more efficient processing time for k-means clustering.

4. Results

We annotated 29 scene role labels throughout Chapters IV and V of the Finnish translation of *Le Petit Prince* (de Saint-Exupéry, 2001a), *Pikku Prinssi* (de Saint-Exupéry, 1980), and 32 scene role labels throughout Chapters IV and V of the Latin translation, *Regulus* (de Saint-Exupéry, 2001b). Cluster sizes of 50, 30 and 15 failed to collocate tokens that were annotated with the same scene role. Even with 30 clusters, a closer approximation to the number of scene roles we annotated, we found that clusters were still partitioned strictly morphologically, inhibiting any meaningful collocations.

To illustrate how we analyzed clusters for each annotation at at time, Table 6 depicts the three clusters produced at varying clustering sizes (50 clusters total, 30 clusters total, 15 clusters total) that were assigned to the Finnish word *yksityiskohdi-ssa*, marked with inessive case. From just this example, it is clear that decreasing the total number of clusters is necessary to increase the scope of a cluster containing *yksityiskohdi-ssa* to include other inessive markers with longer morphological alternations, such as *-uksesssa*, *-assa*, and *-uudessa*, none of which appear in the cluster containing just *-ssa* at levels of 50 clusters total and 30 clusters total. This was the case for all adposition and case marker tokens, where semantic groupings were not nearly inclusive enough until the total number of clusters possible for the algorithm was abstracted down to 15.

4.1. Alignment of Joint Adpositions and Case Markers

In both languages, adpositions were almost always found in separate clusters than their case-marked noun objects. For example, alignment was strongest within the category of PARTPORTION, a scene role that is represented almost exclusively by adpositions in both Finnish and Latin. Thus, individual clusters for adpositions were far more frequent than clusters that pair adpositions with their dependent case marker. For example, the Latin preposition *de* selects an ablative-marked noun object. *De* exists in a cluster populated exclusively by *de* and *a* adpositions, both of which select ablative-marked noun objects. Notably, this cluster was completely disparate from the cluster containing other words marked by the ablative case marker *-e*.

4.2. Adpositional Phrases with Multiple Tokens

muusta	*kuin*
another	like
but, except	

Table 7: Gloss of Postpositional Phrase in Finnish

Some alignment errors indicate potential biases in the raw embeddings. As shown in Table 7, the postpositional phrase *muusta kuin* consists of the noun *muusta* and the postposition *kuin*. Both words together create a joint meaning of "except", which we annotate as PARTPORTION. For this specific instance of the postpositional phrase, a contextually aware embedding would have grouped *kuin* with *muusta*, but mBERT simply classified the *muusta* token with other words marked by the elative *-sta* marker.

This example indicates that mBERT's contextual representations are not capturing polysemous context for postpositional phrases like *muusta kuin* that consist of multiple adpositional tokens. Besides this example, mBERT did generally assign different adposition tokens that we labeled as PARTPORTION to the same cluster. The misalignment of *muusta kuin* indicates that mBERT prioritizes prototypical, word-specific semantics for multiword Finnish adpositions.

4.3. Morphological Alternations of Case Markers

SOURCE and TOPIC showed little to no cluster alignment in both Finnish and Latin. Some Finnish words with ablative and elative case were assigned identical scene role and function roles, indicating that they

were occupying the prototypical semantic role of the case marker (e.g. elative case has the function role SOURCE). Morphologically distinct instantiations of the elative case were not grouped together. For example, *siitä*, the elative form of the pronoun "it", is in a different cluster than *josta*, a wh-word meaning "from which" that is marked by the agglutinative elative case marker *-sta*.

In Latin, the second declension ablative singular case marker *##ulo* from *regulo* was grouped into a cluster containing primarily noun stems. Meanwhile, the first declension ablative singular ending *##ia* in *adansonia* was properly grouped with other ablative singular endings, while the preposition *ab* belonged to a small cluster consisting almost exclusively of *ab* and *in*, another ablative-selecting preposition. While prepositions again cluster well together, the strict partitioning of mBERT clusters, even with regard to morphological variations of the same case marker, showed an even more semantically insensitive encoding of adpositions and case markers than the mismatched clusters of the PARTPORTION examples.

TOPIC is represented by the genitive, elative, and inessive cases in Finnish. All three were in different clusters, so there was no alignment. In the case of the elative-cased *ystävä-stä-nne*, which had an additional possessive suffix *-nne* appended after the elative case marker morpheme *-stä-*, the word was not even grouped together with other clusters containing primarily elative case suffixes; it was grouped with noun stems because the WordPiece token that received an embedding was *##stän*, only 1 letter off from the prototypical elative case marker morpheme. These examples indicate that mBERT clusters similar token embeddings via strict morphological partitioning. The model does not account for morphological alternations of trivial minimum edit distance of the same syntactic structure, much less semantic similarity across distinct case markers that occupy the same semantic role.

4.4. Broader Implications

The choice of the multilingual BERT model certainly has implications for the k-means clustering task that using a monolingual, language-specific model may not have had. The ability to generalize to 104 languages might yield a more homogeneous language neutrality that is unable to account for semantic similarities that cross morphological boundaries within just one language. The monolingual Finnish BERT model, FinBERT (Virtanen et al., 2019) trains the transformer model on significantly more data than mBERT, which only covers 3% of FinBERT's training data. This results in better WordPiece segmentation, more contextual information to draw from, and a theoretically more accurate representation of semantic similarities between the more accurate WordPieces. However, our initial experiments with running a k-means clustering using FinBERT embeddings yielded clusters as similarly disorganized and non-discriminatory as the mBERT clusters. Given the unpromising results, we did not pursue an in-depth analysis of the FinBERT clusters for the purposes of this paper.

Libovický et al. (2020) ran experiments to test the language neutrality of mBERT, and found that crosslingual alignment of function words was widely accurate. Since function words like adpositions are very frequent in a language, they regarded them as part of the centroid of a language and thus the most indicative of language-specific features.

Subtracting the centroid successfully suppressed the language-specific phenomena exhibited by function words by decreasing performance on a language identification task from 93.5% to 86.7%. This accuracy score is still relatively high, showing how largely neutral mBERT is, to the point where removing adpositional semantics barely decreased performance.

This high neutrality seems to have produced somewhat similar cluster types among both Latin and Finnish data. For example, both languages have clusters that group together multiple case markers (inessive, adessive; ablative in Finnish and genitive, dative, and ablative in Latin) as well as reserving clusters for specific morphological instantiations of case endings , like elative *-sta* in Finnish and ablative *-e* in Latin. For both languages, similarities between WordPiece embeddings appeared to be attributed solely to morphological similarity, with morphological alternations of same case marker not being assigned to the same cluster.

Using principal component analysis to reduce dimensionality runs the risk of oversimplifying the semantic sensitivities of contextualized embeddings like mBERT. Given that the higher cosine similarities existed for the 3,072 dimensions representing 4 concatenated layers as opposed to the 768 dimensions representing 4 summed layers, it is clear that the downgrade from 3,072 to 768 already compromised the similarities of vectors representing polysemous tokens. Future work can explore alternative methods for reducing dimensionality, such as kernel PCA.

5. Conclusion

We conducted a novel application of the SNACS supersense hierarchy to Finnish and Latin, neither of which has undergone SNACS analysis before. The corpus of pilot annotations for translations of *The Little Prince* identify unique lexical and combinatorial properties of how adpositions and case markers encode semantic roles. Since other translations of *The Little Prince* have been annotated for SNACS, like Korean, German, and Hindi, the novel adaptation for Latin and Finnish can provide more insight into the crosslingual applicability of the hierarchy.

The corpus also lends itself to an investigation of the computational tractability of semantically complex groupings like SNACS. Using an unsupervised k-

means clustering algorithm, we tested for alignment of annotated SNACS semantic role labels with the raw, contextual Multilingual BERT embeddings of disparate tokens that receive the same SNACS annotation.

Overall, alignment using mBERT embeddings was not successful. Scene roles that spanned morphologically distinct case markers and adpositions could not be matched up with a cluster containing these diverse items. Instead, clusters were almost always partitioned strictly morphologically. Adpositions sometimes clustered together, but mostly within clusters consisting only of that adposition, and not with semantically similar adpositions or case markers that had a different morphological composition. These results indicate that mBERT is somewhat naive in representing adpositions and case markers, failing to account for the semantic variations displayed by these highly polysemous tokens.

Future work can leverage the annotations to create programs that can automatically produce SNACS annotations for Finnish and Latin data. Another experiment of interest is whether BERT embeddings can be mathematically retrofitted to better reflect the complex semantic content carried by adpositions and case markers.

6. Bibliographical References

Chi, E. A., Hewitt, J., and Manning, C. D. (2020). Finding Universal Grammatical Relations in Multilingual BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5564–5577, Online, July. Association for Computational Linguistics.

de Saint-Exupéry, A. (1980). *Pikku Prinssi*. WSOY.

de Saint-Exupéry, A. (2001a). *Le Petit Prince*. Clarion Books.

de Saint-Exupéry, A. (2001b). *Regulus*. Mariner Books.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Hwang, J. D., Bhatia, A., Han, N.-R., O'Gorman, T., Srikumar, V., and Schneider, N. (2017). Double Trouble: The Problem of Construal in Semantic Annotation of Adpositions. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (* SEM 2017)*, pages 178–188.

Libovický, J., Rosa, R., and Fraser, A. (2020). On the Language Neutrality of Pre-trained Multilingual Representations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1663–1674, Online, November. Association for Computational Linguistics.

Liu, N. F., Gardner, M., Belinkov, Y., Peters, M. E., and Smith, N. A. (2019). Linguistic Knowledge and Transferability of Contextual Representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1073–1094, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space.

Pennington, J., Socher, R., and Manning, C. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October. Association for Computational Linguistics.

Schneider, N., Hwang, J. D., Srikumar, V., Prange, J., Blodgett, A., Moeller, S. R., Stern, A., Bitan, A., and Abend, O. (2018). Comprehensive Supersense Disambiguation of English Prepositions and Possessives. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 185–196, Melbourne, Australia, July. Association for Computational Linguistics.

Schneider, N., Hershcovich, D., Mannion, R. A., Prange, J., Somers, C., Blodgett, A., and Gessler, L. (2020a). Streusle. STREUSLE: a corpus with comprehensive lexical semantic annotation (multiword expressions, supersenses), `https://github.com/nert-nlp/streusle`.

Schneider, N., Hwang, J. D., Bhatia, A., Srikumar, V., Han, N.-R., O'Gorman, T., Moeller, S. R., Abend, O., Shalev, A., Blodgett, A., and Prange, J. (2020b). Adposition and Case Supersenses v2.5: Guidelines for English. *arXiv:1704.02134 [cs]*, March. arXiv: 1704.02134.

Schneider, N., Srikumar, V., and Hwang, J. (2022). Xposition. existing language corpora with SNACS annotations, `http://flat.nert.georgetown.edu/`.

Srikumar, V. and Roth, D. (2013). Modeling Semantic Relations Expressed by Prepositions. *Transactions of the Association for Computational Linguistics*, 1:231–242.

Tenney, I., Xia, P., Chen, B., Wang, A., Poliak, A., McCoy, R. T., Kim, N., Van Durme, B., Bowman, S. R., Das, D., and Pavlick, E. (2019). What Do You Learn From Context? Probing for Sentence Structure in Contextualized Word Representations.

Virtanen, A., Kanerva, J., Ilo, R., Luoma, J., Luotolahti, J., Salakoski, T., Ginter, F., and Pyysalo, S. (2019). Multilingual is not enough: BERT for Finnish.