Computational Analysis versus Human Intuition: A Critical Comparison of Vector Semantics with Manual Semantic Classification in the Context of Plains Cree

Daniel Dacanay
Antti Arppe
Atticus Harrigan
University of Alberta
4-32 Assiniboia Hall, University of Alberta, Edmonton,
Alberta, Canada T6G 2E7

Abstract

A persistent challenge in the creation of semantically classified dictionaries and lexical resources is the lengthy and expensive process of manual semantic classification, a hindrance which can make adequate semantic resources unattainable for under-resourced language communities. We explore here an alternative to manual classification using a vector semantic method, which, although not yet at the level of human sophistication, can provide usable first-pass semantic classifications in a fraction of the time. As a case example, we use a dictionary in Plains Cree (ISO: crk, Algonquian, Western Canada and United States)

1. Introduction

One of the challenges in the construction of lexical resources such as dictionaries is the dilemma of their structural organisation. While convention would have it that dictionaries are organised alphabetically, this is largely an artefact of custom, and, although widely conventionalised, does little to mimic (or even correspond to) the generally accepted psycholinguistic reality of lexical organisation (Lucas, 2000; Miller et al., 1993). Perhaps the most prominent alternative to alphabetic organisation is semantic classification. Modern semantic dictionaries, far from mere thesauruses, have a variety of practical uses, ranging from improving the accuracy of machine translation and predictive text (Giménez et al., 2005) to creating digital language instruction tools (Lemnitzer and Kunze, 2006).

Likely the most well-known modern attempt at large-scale semantic classification stemmed from Princeton University in the mid-1980s with the creation of WordNet, an ontology of semantic classification based around the relationships of sets of semantically and distributionally proximate lexical items known as synsets, the structure of which Miller claimed to be “consistent with psycholinguistic evidence” of mental semantic organisation (Miller et al., 1993). This structure is a return to the Firthian notion of wording meanings being construed contextually rather than denotationally or (de)compositionally (Firth, 1957; Arppe, 2008). Although initially developed for English, the WordNet approach for semantic classification has since become a staple in modern lexicography, with WordNets of varying size and complexity existing for many prominent global and national majority languages, such as German with GermaNet (Hamp and Feldweg, 1997; Hinrich and Hinrichs, 2010), Finnish with FinnWordNet (Lindén and Niemi, 2014), and Korean with KorLex (Aesun Yoon et al., 2009), among dozens of others. However, while semantic classifications such as these have become relatively commonplace among prominent majority languages in the developed world, they remain a rarity among under-documented or otherwise poorly resourced languages (Bosch and Griesel, 2017). Using existing, conventional lexical resources, we provide here a holistic comparison between a manual method in semantic classification using a WordNet-based ontology and an automatic computational method via vector semantics, with respect to the efficacy and precision of both methods.

Proceedings of the 4th Workshop on the Use of Computational Methods in the Study of Endangered Languages: Vol. 1 Papers, pages 33–43,
Online, March 2–3, 2021.

2. Plains Cree

Plains Cree (*nêhiyawêwin*) is an Indigenous language of the Algonquian family, spoken throughout Alberta, Saskatchewan, and parts of northern Montana. Although exact population figures for Plains Cree are difficult to ascertain, the 2016 Census of Population recorded 33 975 native speakers of 'Cree-Montagnais languages' in Alberta and Saskatchewan (Statistics Canada, 2016). This speaker-base, though largely elderly, makes Plains Cree one of the most widely-spoken Indigenous languages in Canada, both in terms of population and geographic reach, a fact which has no doubt contributed to its comparatively comprehensive documentation both in the context of historical missionary observations (LaCombe, 1874) and contemporary academia (Schmirler et al. 2018; Arppe et al., 2019), with at least four major contemporary dictionary resources, comprehensive descriptions and computational models of morphology and syntax (Arppe et al., 2016), and text corpora in the hundred thousands of words (Arppe et al., 2020). Despite recent efforts at revitalisation, such as Cree language schooling, digital resources for Plains Cree, though existent (Arppe et al., 2018) remain scarce.

As an Algonquian language, Plains Cree is highly polysynthetic, with much of its morphological complexity manifesting in verbal morphology, with verbal prefixes largely supplanting adjectives and adverbs as distinct lexical classes (Wolfart, 1973). As with many American Indigenous languages, verbs make up the largest single portion of the lexicon, constituting as much as 79% of word types in existing corpora (Harrigan et al., 2017). There are substantial differences in the general lexicalisation patterns of Plains Cree and English (see section 5)

3. Fundamentals of WordNet

WordNet largely operates on the "central organizing principle" (Miller et al., 1993) of hypernymy and hyponymy with respect to sets of (near-)synonymous words known as *synsets*. Synsets are defined as being groups of words with closely related, distributionally similar meanings, for which, in any given context *C*, "the substitution of one for the other in *C* does not alter the truth value" (Miller, 1993), while the relationships of hypernymy and hyponymy are defined in WordNet as situations wherein, if *x* is defined as a hyponym of *y*, speakers would consider *x* to be a kind of *y*, with *x* inheriting all basic characteristics of *y* while adding at least one other distinguishing feature both from *y* and from other hyponymic synsets of *y* (Miller et al., 1993). While other supplemental lexical relationships exist, they are largely secondary in the fundamental structure of WordNet, and a skeletal, core WordNet of any given language could retain the basic structure of a full WordNet using only these three relationships.

The use of such a simplification of WordNet's semantic relations significantly reduces the amount of time necessary to semantically classify each word, as only a direct correspondence to the relevant WordNet synset would be necessary for a lexical item in the target language to be considered classified, with first-pass hypernymy and hyponymy relationships constructed indirectly by populating synsets. Using this method, manual classification of dictionary items can provide a basic semantic ontology of the target language at a rate of 400-500 word types daily per annotator, compared with a rate of ~1000 synsets per year reported by Bosch and Griesel during their creation of full WordNets for low-resource South African Bantu languages (Bosch and Griesel, 2017). This skeletal form of WordNet also provides the benefit of requiring substantially less linguistic background knowledge to effectively use, reducing the need for lengthy annotator training sessions. Although the end product will inevitably be one of reduced semantic richness, and despite the fact that this method erroneously assumes the basic semantic hierarchies of English to be identical to those of the target language, these simplifications bolster the pragmatic feasibility of performing semantic classification at all in situations where resources for linguistic analysis are scarce.

It is perhaps prudent to note that there already exists a semantic ontology specifically designed for the classification of Algonquian languages, created by Prof. Marie-Odile Junker and Linda Visitor for the Eastern James Bay Cree Thematic Dictionary in 2013. Unlike WordNet, this ontology was purpose-designed for Cree semantic classification, being structured to more accurately reflect not only the lexicalisation patterns of Algonquian languages, but also their general semantic makeup and hierarchies of their vocabulary. Though certainly a useful tool, this ontology was not used in the semantic classification of Plains Cree for the principal reason of transferability; although WordNet may be less tailored to the semantic specifications of Plains Cree, one of its principal allures is the potential it provides for widespread interlinguistic comparisons of semantic content. As such, even if only a fractional version of WordNet is to be applied, using a WordNet-based ontology to begin with ensures a relative ease of semantic comparison between Plains Cree and other languages with WordNets or pseudo-WordNets. Ultimately, this ease of transferability and comparison proved more appealing than the tailor-made structure of the East Cree ontology.

4. Lexical Resources

The corpus used in this analysis was the lexical database underlying *nêhiyawêwin : itwêwina* or *Cree: Words*, a bilingual Cree-English dictionary compiled in the early 2000s by Arok Wolvengrey (2011). This continually-updated dictionary exists in both in print and digital form, and currently consists of 21,347 word types, spread across nouns, verbs, and various lexical and non-lexical particles. *Cree: Words* [CW] provides its entries both in Standard Roman Orthography (SRO) and syllabics, and provides a wealth of other information, such as derivational breakdowns. Some entries in the CW database already had rudimentary semantic notes in place, however, since these notes were largely non-ontological, and at that only existed for fewer than 1000 total entries, they were ignored.

5. The Manual and Computational Methods

The Manual Method: The process of manually semantically classifying the CW dictionary was fairly straightforward; each Cree entry was provided with one or more correspondences to synsets in the Princeton WordNet, with these correspondences being as specific to the English gloss of the Cree word as possible. Lexical class differences between the English synset and the Cree word were ignored, for example, *têyistikwânêw* (an intransitive verb glossed as 'I have a headache') was given the nominal WordNet correspondence *(n) headache#2*. Similarly, as semantic concepts typically conveyed by adjectives and adverbs in English often take the form of bound morphemes in Cree, correspondences to English adjectival and adverbial synsets were often considered relevant for inclusion among the synsets of an entry, for example, *osâwi-sênipân* (a noun glossed as 'yellow ribbon') was given correspondences both to *(n) ribbon#4* and *(adj) yellow#1*.

In assigning WordNet correspondences manually, care was taken to focus classifications on what were perceived to be the semantically central aspects of entries with lengthy glosses. For example, in the entry *wiýinwâpisk*, glossed as 'a certain kind of stone which feels oily (e.g. mica)', there are six lexical words in the gloss, only three of which, namely *stone*, *oily*, and *mica*, are directly relevant to the meaning of the entry. As such, only *(n) stone#1*, *(n) mica#1*, and *(adj) oily#3* were used as correspondences in the manual classifications, with 'certain', 'kind' and 'feels' being ignored.

In general, Cree lexicalises many broad semantic concepts which English does not; an example of this would be the common suffix combination of the augmentative *-sk* and the verb ending *-âw*, which, when used in combination with a noun, produce a form meaning "*x* is abundant" (e.g. *kinosêw* ('fish') to *kinosêskaw* ('there is an abundance of fish')). In the manual classifications, all such derivations were classified with the closest English gloss of the base noun and *(adj) abundant#1* (e.g. *kinosêskaw* was given the classifications *(n) fish#1* and *(adj) abundant#1*). Other common Cree lexicalisation patterns absent in English include regular diminutives and augmentatives, prefixes describing quality, and the regular occurrence of verbs with implicit instrumentals (for example, *kawiwêpahwêw* ('s/he knocks s.o. down by tool') and *kawiwêpiskawêw* ('s/he knocks s.o. down by foot')).

Many terms relating to culturally significant activities such as hunting and leatherworking lacked any correspondence at all in English aside from the highly general; for example, *misipocikan*, glossed as 'sharp-edged rubbing tool used to soften hides'. In examples such as these, more generalised classifications were used (such as *(n) scraper#1* for *misipocikan*).

For the present investigation, particles of all kinds were ignored throughout the manual classification, following the general line of thought of the original Princeton WordNet that they are "probably stored separately as part of the syntactic component of language" (Miller et al., 1993; Garret, 1982).

In addition to the manual WordNet classifications, each Cree word was also given a correspondence in SIL's Rapid Words ontology, a deliberately simplified semantic classification scheme operating on largely the same three semantic relationships as the skeletal WordNet (Boerger, 2017; Moe, 2003). These classifications were largely done to facilitate future comparative research with an existing Plains Cree dictionary, the community-created *Maskwacîs Dictionary of Cree Words,* which has already been classified with Rapid Words (Reule, 2018), as well as for future trials of Rapid Words as a classification ontology in the vector method.

The Vector Semantic Method: The use of the vector semantic method in the semantic classification of dictionaries, although relatively novel, is not an innovation of this study (Wei Li, 2018; Brixey et al., 2020), and usable semantic results have been obtained through its application in past investigations. Jurafsky and Martin (2019) define vector semantics as the representation of word meanings "as a point in some multidimensional vector space", with semantically related words occurring in "distinct portions of the space", and the exact degree of relatedness between any two given points being ascertainable by cosine distance. This method in turn assumes a distributional approach to lexical semantics, echoing the doctrine of John Firth (Firth 1957) and Zellig Harris (Harris 1954), where the meaning of any given word can be ascertained entirely and exclusively through patterns in context; for example, even one did not know the meaning of a word such as *mopane*, seeing the word occur in the sentences "Mopane is often fried and eaten with onions", "Mopane is the principal source of protein for millions in Southern Africa", and "Aversion to eating insects often turns Westerners away from eating mopane" would be enough to infer that mopane is an edible insect in Southern Africa, even without being explicitly told. The vector semantic method analyses the average contexts of a word across large corpora and enumerates the patterns of the word's co-occurrence with other words into numerical values known as dimensions, representing the average context of that word in terms of a set of constant numerical values (for example, 'mopane' might have a high numerical value for the dimension of 'edibility' based on its frequent use in a dietary context, but a low value for the dimension of, say, items of furniture, as it would rarely co-occur with furniture in context). These dimensions may then be compared with the dimensions of other words via cosine, with more similar dimensions between two words indicating a more similar average context, and thus a more similar meaning.

To obtain vector-based classifications, 300-dimensional word vectors for the Cree entries in the CW dictionary were generated using embeddings created by *word2vec*, a freely available NLP neural network model, trained with the Google News Corpus. These vectors were generated for the contents of the dictionary by averaging the embeddings corresponding to all of the individual words in the English glosses of the Cree entries. Similar 300-dimensional vectors were created for all WordNet entries, again averaging the embeddings for the individual words listed in their synset as well as the explanatory glosses. If any of these individual English words in the CW or WN lexical entries were not in the Google News Corpus, they were excluded. Finally, the semantic similarity of the CW and WN entries was calculated based on the cosine distance between their respective vectors.[1]

[1] The use of the entire dictionary entry hopefully would disambiguate any sense-wise ambiguity associated with head words for the WordNet entries.

5.1 Conditions Necessary for Both Methods

The manual classification method holds the advantage of requiring almost no prerequisite equipment, aside from access to the Princeton WordNet of English, which is freely and publicly available online, and access to a digitised version of the chosen lexical resource in the target language. The simplifications to WordNet also ensure that extensive training in the intricacies of semantic relationships would not be overtly necessary to effectively classify new lexical resources; while a high degree of fluency and an extensive (at least passive) vocabulary knowledge in English are ideal, only a relatively basic understanding of WordNet and semantic relationships in general would be required to effectively use the skeletal WordNet as an ontology. In the experience of the first author, only about a day of practice in its application was necessary to begin classifying entries at a rate of several hundred per day; although this rate does assume a previous understanding of semantic relations.

Perhaps the most demanding aspect of the manual semantic method is the requirement of annotators and of time; although much faster than a traditional WordNet, the classification of a single, medium-sized dictionary such as the CW dictionary would still take a single annotator, working more-or-less full time between one and two months to complete. Although this task may be expedited through the use of several annotators, such collaboration would require substantial co-ordination to ensure a consistent annotating style, particularly as it pertains to dictionary entries with no clear single-synset English translation.

Given that the necessary scripts for generating the dictionary entry vectors and comparing cosines with the WordNet vectors already exist, the computational method takes only as much time as is necessary to run aforementioned scripts. The process of calculating the cosine differences between the CW dictionary vectors and the WordNet vectors can take, on a mid-range laptop with 2 cores and 8gb RAM, at most between four and five days. However, the calculations of cosine differences is an *embarrassingly parallelisable* task, taking 90 minutes when run on 64 cores with 4-8gb RAM each (on the *Cedar* high-performance computing cluster maintained by Compute Canada).

6. Discussion of Results

If the goal of vector-based ontological semantic classification is to be taken as the imitation of human judgement in assigning precise semantic correspondences to the level of the individual word, or in the case of WordNet, to the individual synset, then the results generated by the use of vector semantics on the CW dictionary could be said to be a mixed success with respect to nouns, and a decisive (although not absolute) failure with respect to verbs:

%	Verbs, top	Verbs, median	Nouns, top	Nouns, median
0%	1.0	1.0	1.0	1.0
10%	5.0	11.0	1.0	2.0
20%	18.0	51.7	2.0	4.0
30%	51.6	166.3	4.0	8.0
40%	136.8	448.8	7.0	16.1
50%	333.0	1045.0	15.0	30.5
60%	762.2	2057.3	28.0	60.0
70%	1633.87	4096.4	59.0	139.0
80%	3553.8	8036.9	164.0	375.4
90%	9553.8	17448.6	864.2	1670.4
100 %	137352.0	137352.0	121883.0	121883.0

Table 1, the vector-assigned rank of manual WN classifications in percentiles, both for the single top ranked manual classification, and for the median if there were several.

As shown in Table 1, although it was uncommon for the top-vector-selected item to exactly match the manual classification (315 times altogether for the 11.2k verbs; 726 for the 5.5k nouns), with CW verbs, the manual classification occurs in the top 0.24% (333/137k) 50% of the time when counting only the highest ranked manual classification to occur. For 90% of the CW verbs, the manual classification could be found among the top 17.4k (7%) of the 137k computationally ranked WN entries. For the nouns, the top-ranked manual classification was in the top 0.7% of selections (864.2) 90% of the time, and in the top 15 selections 50% of the time.

The median computationally-assigned position of the human selected classification for Cree verbs was 333, with a mean of 3671; for nouns, the median was 15, and the mean 1194. Even in cases where the manual-selection has a relatively low rank, high-ranking items for most entries tend to have the same basic semantic region as the target Cree word. The reason behind the substantially increased accuracy of noun predictions in comparison with verbs is likely a result of general lexicalisation patterns in Cree, rather than a short-coming in the vector method. While Cree verbs cover a wide range of semantic areas which English verbs do not, and often convey full clause or sentence level meanings, Cree nouns cover more or less the same basic semantic and syntactic concepts as their English counterparts, meaning not only that there is more often a single-synset correspondence for Cree nouns in WordNet, but also that that correspondence is more likely to be lexicalised as a noun in the English WordNet.

The (English) part of speech assigned to the topmost vector classification had a strong tendency to correlate with the part(s) of speech of the manual classification. When the manual classification contained multiple (English) parts of speech, they tended to be represented more or less equally in the vector classification, although verbs and nouns seemed slightly favoured over adjectives and adverbs (see Tables 4 and 5, Appendix)

CW nouns: Manual WN PoS	(n)	(adj)	N/A	(v)	(adv)	
(n)	2648	0	185	0	0	2833
(v)	0	0	0	59	0	59
(adj)	0	31	5	0	0	36
Total	2648	31	190	59	0	2928

CW verbs: Manual WN PoS	(v)	(n)	(adj)	(adv)	N/A	
(v)	2891	0	0	0	6	2897
(adj)	0	0	532	0	24	556
(n)	0	376	5	0	14	390
(adv)	0	0	0	13	0	13
Total	2891	376	537	13	44	3856

Tables 2 and 3, confusion matrices of the PoS of top-ranking vector classifications with PoS of manual classifications for Cree nouns (2) and verbs (3) with a single manual classification.

The close correlation between the proportions of vector-assigned and manually-assigned parts of speech seems to justify the decision to allow multiple synset classifications in the manual method, as the vector method consistently validated the human PoS classifications for the English WordNet correspondences, even replicating their proportions.

6.1 Overspecificity

One consistent peculiarity with the vector semantic method was its tendency to assign highly specific semantic denotations to relatively general terms. For the word *sîsîp*, meaning ‘duck’, although the corresponding WordNet synset, *(n) duck#1*, is only 46th, 29 of the preceding 45 correspondences are specific types of ducks, 5 are bodyparts of the duck, 6 are related to duck hunting, and 3 are miscellaneous, duck-related terms. Thus, although the human-selected correspondence only barely appeared in the top fifty, virtually all of the higher-ranking correspondences were either species of, or activities related to, ducks. This pattern, whereby highly specific variants of a general concept precede the more general catch-word for that concept in terms of perceived semantic relatedness, is visible throughout the vector classifications of nouns and verbs alike, being particularly noticeable in the semantic domains of tool names and broadly referential plant and animal names; for example while the corresponding human-selected WordNet terms for *maskwa* (‘bear’) and *apisimôsos* (‘deer’) occurred in 577th and 71st place, preceded by various related, but overly specific terms, more species-specific animal names such as *môswa* ('moose') and *amisk* ('beaver') generally saw human-like correspondences in higher ranks (5th and 20th respectively).

This general, although non-universal, trend towards overspecificity in high-ranking vector correspondences has been remarked upon before; in their analysis of the vector method on Choctaw, Brixey et al. found that, when seeking to return results for the nominal and adjectival forms of the word 'female', their vector model would return specific female names instead (Brixey et al., 2020). This tendency is almost certainly the result of a fundamental methodological difference between human consideration and vector calculation, namely, while the semantic ideas contained in the mental representation of the term 'duck' are likely fuzzy and variable, and can be seen to consist of a central *prototype* surrounded by relevant *exemplars* (Taylor, 2008), owing to the word 'duck' referring to several quite different, though distinctly related, animals, the semantic vector for *(n) duck#1* is precisely defined, interacting with other vectors at exact points in the multi-dimensional vector space. Given this fundamental methodological difference, it is evident that a semantic judgement based solely off of a sharply defined, precise vector cannot be reasonably expected to simulate a judgement based on fuzzily-defined semantic regions such as those of the human mind (McNamara, 2005).

6.2 Proper Nouns

In addition to its lexicographical elements, WordNet also contains a large number of more encyclopaedic entries, including historical and fictional figures, with 8244 entries, countries and geographic regions, with ~3500 entries, and currencies, with 414 entries. With these three categories alone, up to 5% of the 207 016 word-sense pairs in WordNet, with 7.8% of the total 155 327 words are proper nouns of this encyclopaedic nature. Generally speaking, the nature of these encyclopaedic terms reflects the cultural matrix within which WordNet was created, that is, the academic circles of the Eastern United States, a fact which has been remarked upon by various non-American WordNet creators (Lindén and Carlson, 2010). Despite their relative semantic irrelevance to most vocabulary, these proper nouns are frequent in the vector classifications, often populating the higher-ranking correspondences (see section 6.3) of words where no clear English equivalent can be found. As they appear to both serve little utility and have extremely limited relevance to both the CW dictionary in particular and to Plains Cree in general (at least in the sense of basic, first-pass semantic classification), we decided that it would be reasonable to remove these encyclopaedic entries from the 'skeletal' WordNet entirely, with little adverse effect.

6.3 The 'regift' Problem

Another persistent failing of the vector method was its occasional overassignment of highly specific, but semantically irrelevant terms; the name given to this principle stems from the fact that one such term, *(v) regift#1*, a verb which, for record, occurs 16 times in the 1.9 billion word Corpus of Global Web-Based English (Davies, 2013), occurs in the top 1000 correspondences of verbs in the CW dictionary 7324 times, putting it in the top 1% of vector-based semantic relatedness for over 65% of the 11236 Cree verbs in the CW dictionary, in comparison with, for example, ~53% for the highly general *(v) say#1* (6000). Other common 'regift' words include *(n) dingbat#1* (8082 occurrences among top 1000 verb correspondences, ~71%), *(n) cunt#1* (6989, ~62%) and *(n) gumption#1* (5844, ~52%), as well as many proper nouns for historical and mythological figures, such as *(n) Dido#1* (1085, ~9.6%), *(n) Godiva#1* (5775, ~51%), and *(n) Rumpelstiltskin#1* (8094, ~72%). This error, although present among the Cree nouns, is substantially more noticeable in the classifications of the verbs, mirroring the general trend of noun classifications more accurately reflecting human judgement than their verb counterparts.

One possible reason behind the regift problem is that is that the vectors of words such as 'regift' and 'Rumpelstiltskin' are unusually close to zero; the average vector of the aforementioned seven words is 0.0015 compared with a WordNet average of 0.004, and as such, when a word has an ill-defined vector which is close to the origin, it is automatically considered semantically proximate in that dimension to a 'regift' word. Another explanation is that the 'regift' words are all low-frequency items in the Google News corpus, and as such their embeddings are disproportionately affected by individually unusual usage contexts; for example, given the extreme infrequency of word ‘regift’, if even a single text in the Google News Corpus overused ‘regift’ in a non-standard way, it would be enough to quantifiably impact the average context of the word in the corpus as a whole, thus skewing the dimensional vectors for the word. In either case, this problem could be (and was) at least partially resolved through the removal of proper nouns in the WordNet vector set, given that proper nouns seem to occur disproportionately frequently in this erroneous fashion. Another possible solution to both this and the more general problem of overspecificity in vector classifications would be the use of synset hypernyms as correspondences rather than specific synset members, in essence forcing the vector method to choose more semantically broad concepts rather than highly specific ones. The rationale behind this solution is the extreme semantic specificity of all of the items which seem to exhibit the ‘regift’ pattern; the inclusion of such highly-specific, infrequently occuring vocabulary seems to be at least partially responsible for the ‘regift’ problem, and thus removing or reducing such vocabulary may serve as a partial solution, at the obvious cost of reduced semantic richness.

6.4. Potential Applications

Given the inconsistency of classification quality between nouns and verbs, the degree of practical human usability of the present results is largely reliant on part of speech; with nouns being on average suitably well-classified for possible use as a complement to manual classification, while verbs remain on average too poorly classified for such use.

Even with the relatively successful noun classifications, the current results would be a poor substitute for manual classification given how infrequently the top-ranking vector classification matches its exact human-selected counterpart; rather, the vector-selected noun correspondences would be best used as an aid in manual classification, as opposed to a replacement. Given that the median rank of the ‘correct’ human classification is 15th, one could apply the vector method prior to manually classifying a list of Cree nouns and use the top 15 vector classifications for each entry as a starting point for classification, with a 50 percent chance that the ‘best’ human classification is contained within that list, rather than classifying all noun entries from scratch. This would save time for the manual annotator by preventing them from needing to search the entirety of WordNet for each entry, as well as by providing them with a variety of potentially related synsets, which would also allow annotators with less familiarity with the format of WordNet to more effectively classify entries by virtue of finding the ‘best’ synsets for them automatically (at least, a portion of the time). As such, with the present degree of accuracy, vector semantic classifications appear best suited as an annotation primer to aid a human annotator, although (at least in Plains Cree) this application would only seem suitably efficient for noun classification.

7. Conclusion

The vector semantic method is a usable and resource-non-intensive alternative to manual semantic classification which has proven capable of reliably providing accurate semantic domains for nouns and, to a lesser extent, verbs, in Plains Cree. Although not yet at the level of human-like semantic awareness, the vector semantic method is nonetheless capable of producing relatively accurate first-pass semantic classifications for digital lexical resources without the need for time-consuming and expensive manual annotation, serving both as a valuable step in the democratisation and increased availability of semantic analysis and ontological dictionaries for language communities with limited resources, and as a potential streamliner in the process of creating digital language resources relying on semantic relationships. Within Plains Cree, the vector method is capable of classifying nouns accurately enough to seemingly be usable in its present state as an annotation aid, and verb classifications, although underwhelming from a human-centric perspective and still insufficiently reliable to be usable as a replacement for, or accessory to, manual classification, are nonetheless statistically promising, with human-selected classifications reliably occuring in the top 8% of total classifications. Though showing initial promise, the avenues for improvement for the vector method are manifold, including removing redundant WordNet vectors, generating target-language vectors from monolingual target-language corpora, and testing the semantic accuracy of vector classification with more general semantic categories, or alternatively using an entirely different system of categories altogether.

References

Aboriginal languages in Canada, 2016 Census of Population, Statistics Canada, Government of Canada,27 Oct. 2017, www150.statcan.gc.ca/n1/pub/11-627-m/11-627-m2017035- eng.htm. Accessed 18 Aug. 2020.

Aesun Yoon, Soonhee Hwang, Eunryoung Lee, Hyuk-Chul Kwon, "Construction of Korean Wordnet 「KorLex 1.」". *Korean Institute of Information Scientists and Engineers : software and application 36(1),* 2009, pp. 92-108

Arppe, Antti. *Univariate, bivariate, and multivariate methods in corpus-based lexicography – a study of synonymy*. Department of General Linguistics, University of Helsinki, 2008, p. 7, sites.ualberta.ca/~arppe/Publications/Arppe_Dissertation_Final_Print.pdf.

Arppe, Antti, Jordan Lachler, Trond Trosterud, Lene Antonsen, & Sjur N. Moshagen. Basic Language Resource Kits for Endangered Languages: A Case Study of Plains Cree. In: C. Soria, L. Pretorius, T. Declerck, J. Mariani, K. Scannell & E. Wandl-Vogt (eds). *CCURL 2016: Collaboration and Computing for Under-Resourced Languages: Towards an Alliance for Digital Language Diversity (LREC 2016 Workshop)*. Portoroz, Slovenia: European Language Resource Association, 2016 Retrieved from: http://www.lrec-conf.org/proceedings/lrec2016/workshops/LREC2016Workshop-CCURL2016_Proceedings.pdf

Arppe, Antti, Atticus Harrigan, Katherine Schmirler & Arok Wolvengrey. A morphologically intelligent online dictionary for Plains Cree, Presentation conducted at the meeting of *Stabilizing Indigenous Languages Symposium (SILS)*, University of Lethbridge, Lethbridge, Alberta. 2018, June

Arppe, Antti, Katherine Schmirler, Atticus G. Harrigan & Arok Wolvengrey. *A Morphosyntactically Tagged Corpus for Plains Cree***. In M. Macaulay & M. Noodin (eds), Papers of the 49th Algonquian Conference (PAC49), 49. East Lansing, Michigan: MSU Press. Forthcoming, 2020, pp. 1-16

Boerger, Brenda H. "Rapid Word Collection, dictionary production, and community well-being." *5th International Conference on Language Documentation and Conservation*, Mar. 2017.

Bosch, Sonja E., and Marissa Griesel. "Strategies for building wordnets for under-resourced languages: The case of African languages." *Literator - Journal of Literary Criticism, Comparative Linguistics and Literary Studies*, vol. 38, no. 1, 31 Mar. 2017, p. 8, doi:https://literator.org.za/index.php/literator/article/view/1351/2294. Accessed 12 Sept. 2020.

Brixey, Jacqueline, et al. "Exploring a Choctaw Language Corpus with Word Vectors and Minimum Distance Length." Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), May 2020, pp. 2746-53.

Davies, Mark. (2013) *Corpus of Global Web-Based English: 1.9 billion words from speakers in 20 countries (GloWbE)*. Available online at https://www.english-corpora.org/glowbe/.

Firth, J. R. A Synopsis of Linguistic Theory, 1930–1955. In: Firth, J. R. 1968. Selected Papers of J. R. Firth 1952-1959. London: Logmans, 1957, pp. 168-205.

Garrett, M. F. "Production of Speech: Observations from Normal and Pathological Language Use" in A. Ellis (ed.). *Normality and Pathology in Cognitive Functions*. London: Academic Press. 1982

Giménez, Jesus, et al. "Automatic Translation of WordNet Glosses". TALP Research Center, January 2005.

Hamp, Birgit and Helmut Feldweg. "GermaNet - a Lexical-Semantic Net for German." *Proceedings of the ACL workshop Automatic*

Information Extraction and Building of Lexical Semantic Resources for NLP Applications. Madrid, 1997.

Harrigan, Atticus, et al. "Learning from the computational modelling of Plains Cree verbs." *Morphology*, vol. 27, 30 Oct. 2017, pp. 565-98.

Harrigan, Atticus, et al. "A Preliminary Plains Cree Speech Synthesizer." *Proceedings of the 3rd Workshop on the Use of Computational Methods in the Study of Endangered Languages (ComputEL-3)*, vol. 1, 2019, pp. 64-73, doi:https://journals.colorado.edu/index.php/computel/article/view/421/403.

Harris, Zellig S. "Distributional Structure." *Word*, vol. 10, no. 2-3, 1954, pp. 146-62, doi:https://www.tandfonline.com/doi/pdf/10.1080/00437956.1954.11659520.

Henrich, Verena and Erhard Hinrichs. "GernEdiT - The GermaNet Editing Tool". *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC 2010)*. Valletta, Malta, May 2010, pp. 2228-2235.

LaCombe, Albert. *Dictionnaire de la Langue des Cris*. C.O. Beauchemin & Valois, 1874.

Jurafsky, Dan, and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. 3rd ed., 2019, pp. 94-119.

Lemnitzer, Lothar, and Claudia Kunze. "Using WordNets in Teaching Virtual Courses of Computational Linguistics." *Seminar für Sprachwissenschaft, Universität Tübingen*, Jan. 2003

Li, Wei, et al. *Improving Word Vector with Prior Knowledge in Semantic Dictionary*. Beijing, Key Laboratory of Computational Linguistics, Peking University, 2018.

Lindén, Krister and Jyrki Niemi. 2014. "Is it possible to create a very large wordnet in 100 days? An evaluation". *Language Resources and Evaluation* 48(2), 2014, pp. 7, 191–201. doi:10.1007/s10579-013-9245-0

Lindén, Krister, and Lauri Carlson. "FinnWordNet–WordNet på finska via översättning." *LexicoNordica*, vol. 17, 2010, pp. 3-4, http://www.ling.helsinki.fi/~klinden/pubs/FinnWordnetInLexicoNordica-en.pdf. Accessed 12 Sept. 2020.

Lucas, Margery. "Semantic priming without association: A meta-analytic review." *Psychonomic Bulletin & Review*, Dec. 2000, pp. 618-30

McNamara, Timothy P. Semantic priming: perspectives from memory and word recognition (p. 200). *Psychology Press*. 2005, retrieved January 15, 2011

Miller, George, et al. *Introduction to WordNet: An On-line Lexical Database*. Princeton University, 1993, pp. 1-9

Moe, Ronald. Compiling dictionaries using semantic domains. *Lexikos* 13, 2003, pp. 215-223, http://lexikos.journals.ac.za/pub/article/view/731

Reule, Tanzi. *Elicitation and Speech Acts in the Maskwacîs Spoken Cree Dictionary Project*. Department of Linguistics, University of Alberta, 2018.

Schmirler, Katherine, Antti Arppe, Trond Trosterud & Lene Antonsen (2018). Building a Constraint Grammar Parser for Plains Cree Verbs and Arguments. *Proceedings of the 11th Language Resources and Evaluation Conference (LREC 2018)*, 2981-2988. Miyazaki, Japan: European Language Resource Association. Retrieved from http://www.lrec-conf.org/proceedings/lrec2018/pdf/873.pdf.

Taylor, John R. *Handbook of Cognitive Linguistics and Second Language Acquisition*. Routledge, 2008, pp. 39-66.

Visitor, Linda, Marie-Odile Junker, and Mimie Neacappo, eds. *Eastern James Bay Cree Thematic Dictionary ([Northern/Southern] Dialect)*. Chisasibi: Cree School Board, 2013. Print.

Wolfart, H. Christoph. "Plains Cree: A Grammatical Study." *Transactions of the American Philosophical Society, New Series*, vol. 63, no. 5, Nov. 1973, pp. 1-90.

Wolvengrey, Arok. *Cree: Words*. 11th ed., University of Regina Press, 2011.

Appendix

Manual PoSs	(v)	(n)	(adj)	(adv)	N / A	
(n) + (v)	409	372	0	0	0	781
(adv) + (v)	262	0	0	223	0	485
(adj) + (n)	0	213	177	0	0	390
(adj) + (v)	219	0	127	0	0	346
(n) + (v) + (v)	106	48	0	0	0	154
(adv) + (v) + (v)	85	0	0	35	0	120
(n) + (n) + (v)	37	51	0	0	0	88
(adj) + (v) + (v)	45	0	19	0	0	64
(adj) + (n) + (v)	21	25	16	0	0	62
(adj) + (n) + (n)	0	31	23	0	0	54
(adj) + (adj) + (v)	25	0	21	0	0	46
(adj) + (adj) + (n)	0	13	29	0	0	42
(adv) + (adv) + (v)	22	0	0	17	0	39
(adv) + (n) + (v)	15	10	0	7	0	32
(adj) + (adv) + (v)	12	0	3	9	0	24
Total	1350	802	452	315	0	2919

Table 4, confusion matrix of vector-assigned WordNet PoS for Cree verb entries in which the manual classification had correspondences from multiple different lexical classes

Manual PoSs	(n)	(adj)	N/ A	(v)	(adv)	
(adj) + (n)	249	138	1	0	0	388
(adj) + (n) + (n)	109	12	0	0	0	121
(n) + (v)	28	0	0	28	0	56
(adj) + (adj) + (n)	12	21	0	0	0	33
(adj) + (n) + (n) + (n)	13	3	0	0	0	16
(n) + (v) + (v)	7	0	0	6	0	13
(adv) + (v)	0	0	0	5	6	11
(adj) + (adj) + (n) + (n)	5	4	0	0	0	9
(n) + (n) + (v)	5	0	0	4	0	9
(adv) + (n)	3	0	0	0	5	8
(adj) + (n) + (v)	2	3	0	1	0	6
(adv) + (v) + (v)	0	0	0	3	3	6
(adj) + (adj) + (adj) + (n)	1	4	0	0	0	5
(adj) + (v)	0	2	0	3	0	5
(adj) + (v) + (v)	0	1	0	2	0	3
Total	438	204	1	55	22	720

Table 5, confusion matrix of vector-assigned WordNet PoS for Cree noun entries in which the the manual classification had correspondences from multiple different lexical classes