Leveraging English Word Embeddings for Semi-Automatic Semantic Classification in Nêhiyawêwin (Plains Cree)

Atticus G. Harrigan
University of Alberta
4-32 Assiniboia Hall
University of Alberta, Edmonton
atticus.harrigan@ualberta.ca

Antti Arppe
University of Alberta
4-32 Assiniboia Hall
University of Alberta, Edmonton
arppe@ualberta.ca

Abstract

This paper details a semi-automatic method of word clustering for the Algonquian language, Nêhiyawêwin (Plains Cree). Although this method worked well, particularly for nouns, it required some amount of manual postprocessing. The main benefit of this approach over implementing an existing classification ontology is that this method approaches the language from an endogenous point of view, while performing classification quicker than in a fully manual context.

1 Introduction

Grouping words into semantic subclasses within a part of speech is a technique used widely throughout quantitative and predictive studies in the field of linguistics. Bresnan et al. (2007) use high level verb classes to predict the English dative alternation, Arppe et al. (2008) uses verb class as one of the feature sets to help predict the alternation of Finnish *think* verbs, and Yu et al. (2017) use polarity classifications (*good* vs *bad*) from pre-defined lexica such as WordNet (Miller, 1998). In many cases, classifications within word classes allow researchers to group words into smaller cohesive groups to allow for use as predictors in modelling. Rather than using thousands individual lexemes as predictors, one can use a word's class to generalize over the semantic features of individual lexemes to allow for significantly more statistical power.

While extensive ontologies of word classifications exist for majority languages like English (Miller, 1998), German (Hamp and Feldweg, 1997), and Chinese (Wang and Bond, 2013), minority languages, especially lesser resourced languages in North America generally do not boast such resources.[1] Where such ontologies do exist, for example in Innu-aimun (Eastern Cree) (Visitor et al., 2013), they are often manually created, an expensive process in terms of time. Alternatively, they may be based upon English ontologies such as WordNet. This opens the window to near-automatic ontology creation by associating definitions in a target language and English through a variety of methods. This is especially important, given the amount of time and effort that goes into manually classifying a lexicon through either an existing ontology (be it something like Rapidwords[2] or even Levin's like classes (Levin, 1993)). Moreover, there is a motivation based in understanding a language and its lexicalization process on its own terms, though how to do this with a lesser resourced language remains unclear.

2 Background

We begun word classification in preparation for modelling a morpho-syntactic alternation in Nêhiyawêwin verbs. One hypothesis we developed for this alternation, based on Arppe et al. (2008), is that the semantic classes of the verbs themselves as well as their nominal arguments would inform the verbal alternation. Due to constraints of time, we investigated methods to automatically classify both verbs and nouns in Nêhiyawêwin. Although statistical modelling remains the immediate motivator for the authors, semantic/thematic classifications have a wide range of benefits for language learners and revitalization, particularly in online lexicographic resources, where one may want to view all words to do with a theme, rather than simply finding translations of single English words.

In creating a framework for automatic semantic classification we make use of Word2vec (Mikolov et al., 2013a) word embeddings. Word embeddings are words represented by n-dimensional vectors. These vectors are ultimately derived from a word's

[1] There is one attempt at semantically classifying Nêhiyawêwin through automatic means found in Dacanay et al. (2021). This work makes use of similar techniques as desccribed in this paper, differing mainly in its mapping of Nêhiyawêwin words onto Wordnet classes.

[2] See http://rapidwords.net/

Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas, pages 113–121
June 11, 2021. ©2021 Association for Computational Linguistics

context in some corpus through the Word2vec algorithm. Unfortunately, the Word2vec method is sensitive to corpus size. We initially attempted to create basic word and feature co-occurrence matrices based on a 140,000 token Nêhiyawêwin corpus (Arppe et al., 2020) to create word vectors using Principal Components Analysis, but in the end found the results to be not practically useful. Similarly, an attempt at both tf-idf and Word2Vec using only the Nêhiyawêwin dictionary produces mostly ill-formed groupings, though in these cases preprocessing by splitting verbs and nouns was not performed. Regardless, the poor performance was most certainly due simply to the paucity of data. Although the available corpora are small, Nêhiyawêwin does have several English-to-Nêhiyawêwin dictionaries, the largest being Wolvengrey (2001). Although a bilingual Nêhiyawêwin-English dictionary, it is one formed from an Indigenous point of view, based on vocabulary from previous dictionaries, some of which have been compiled by Nêhiyawêwin communities from their own perspectives, or gleaned from a number of texts collections rather than attempting to find Nêhiyawêwin word matches for a pre-defined set of English words. This results in dictionary entries such as `sakapwêw: it roasts over a fire (by hanging, with string on stick)`. Definitions such as this take into account the nuanced cultural understanding reflected in the word's morphology.

3 Methodology

To address the issue of corpus size, we attempted to bootstrap our classification scheme with pretrained English vectors in the form of the 3 million word Google News Corpus, which represents every word with a 300-dimensional vector.[3] We make use of the English definitions (sometimes also referred to as glosses) provided in Wolvengrey (2001) and fit to each word its respective Google News Corpus vector. This dictionary makes use of lemmas as headwords, and contains 21,717 entries. The presumption is that the real-world referents (at least in terms of denotation) of English and Nêhiyawêwin words are approximately comparable, in particular when taking the entire set of words in an English definition. Stop words were removed, and where content words were present in definitions in Wolvengrey (2001) but *not* available in the Google News Corpus, synonyms were used (one such example might be the word *mitêwin*, which is unavailable in the corpus and thus would replaced with something like *medicine lodge* or deleted if a synonym was given in the definition as well). Because the Google News Corpus is based in American spelling, while Wolvengrey (2001) is based in Canadian spelling, American forms (e.g. *color, gray*) were converted into Canadian forms (e.g. *colour, grey*). If such preprocessing is not performed, these words are simply unavailable for clustering, as they lack a matching vector.[4] Where a Nêhiyawêwin word had more than one word sense, each sense was given a separate entry and the second entry was marked with a unique identifier. Finally, where needed, words in the Nêhiyawêwin definitions were lemmatized.

Once every word in Wolvengrey (2001) definitions matched an entry in the Google News Corpus, we associated each word in a Nêhiyawêwin definition with its respective Google News Vector. That is, given a definition such as `awâsisihkânis: small doll`, the resulting structure would be:

$$\texttt{awâsisihkânis} = \begin{bmatrix} 0.159 \\ 0.096 \\ -0.125 \\ \vdots \end{bmatrix} \begin{bmatrix} 0.108 \\ 0.031 \\ -0.034 \\ \vdots \end{bmatrix}$$

Because all word-vectors in the Google News Corpus are of the same dimensionality, we then took the resulting definition and averaged, per dimension, the values of all its constituent word-vectors. This produced a single 300-dimensional vector that acts as a sort of naive sentence vector for each of the English glosses/definitions:

$$\texttt{awâsisihkânis} = \begin{bmatrix} 0.134 \\ 0.064 \\ -0.080 \\ \vdots \end{bmatrix}$$

Mikolov et al. (2013b) mention this sort of naive representation and suggests the use of phrase vectors instead of word vectors to address the representation of non-compositional idioms; however,

[3] This corpus was trained on a large corpus of 100 billion words. Available at https://code.google.com/archive/p/word2vec/

[4] In reality, there were only a handful of cases where words occurred in the dictionary but not in the Google News Corpus. Because there are so few examples of this, even simply leaving these items out would not substantiqally change clustering results.

given the way Wolvengrey (2001)'s definitions are written (e.g. with few idiomatic or metaphorical constructions), and for reasons of computational simplicity, we opted to use the above naive implementation in this paper.

After creating the sentence (or English definition) vectors, we proceeded to cluster definitions with similar vectors together. To achieve this, we created a Euclidean distance matrix from the sentence vectors and made use of the `hclust` package in R (R Core Team, 2017) to preform hierarchical agglomerative clustering using the Ward method (based on the experience of (Arppe et al., 2008) in using the method to produce multiple levels of smaller, spherical clusters). This form of clustering is essentially a bottom-up approach where groupings are made by starting with individual labels with the shortest distance, then iteratively at a higher level making use of the clusters that result from the previous step or remaining individual levels; this second step is repeated until there is a single cluster containing all labels. This method of clustering creates a cluster tree that can be cut at any specified level after the analysis has been completed to select different numbers of clusters, allowing researchers some degree of flexibility without needing to rerun the clustering. This method is very similar to what has been done by both Arppe et al. (2008), Bresnan et al. (2007), and Divjak and Gries (2006). The choice of what number of clusters was made based on an evaluation of the effectiveness of the clusters, based on an impressionistic overview by the authors.

For our purposes, we focused on the semantic classification of Nêhiyawêwin nouns and verbs. Nêhiyawêwin verbs are naturally morphosemantically divided into four separate classes: Intransitive verbs with a single inanimate argument (VII), Intransitive verbs with a single animate argument (VAI), transitive verbs with an animate actor[5] and an inanimate goal (VTI), and verbs with animate actors and goal (VTA). For verbs, clustering took place within each of these proto-classes. Among the VIIs, 10 classes proved optimal, VAIs had 25 classes, VTIs with 15 classes, and VTAs with 20 classes. The choice to preprocess verbs into these four classes was as not doing so resulted in a clustering pattern that focused mainly on the difference between transitivity and the animacy of arguments. Any more or fewer classes and HAC clusters were far less cohesive with obvious semantic units being dispersed among many classes or split into multiple classes with no obvious differentiation. Similarly, verbs were split from nouns in this process because definitions in Wolvengrey (2001) vary significantly between verbs and nouns.

Nouns are naturally divided into two main classes in Nêhiyawêwin: animate and inanimate.[6] For our purposes we divide these further within each class between independent (i.e. alienable) and dependent (i.e. inalienable) nouns to create four main classes: Independent Animate Nouns (NA), Dependent Animate Nouns (NDA), Independent inanimate Nouns (NI), and Dependent Inanimate Nouns (NDI). The reason for this further division is due to the morphosemantic differences between independent and dependent nouns in Nêhiyawêwin. While independent nouns can stand on their own and represent a variety of entities, they are semantically and morphologically dependent on some possessor. We opted to pre-split NDIs and NDAs into their own classes, so as not to have the clustering focus on alienablity as the most major difference.[7]

4 Results

In all cases, clusters produced by this procedure needed some amount of post-processing. For nouns, this post-processing was minimal and mostly took the form of adjustments to the produced clusters: moving some items from one class to another, splitting a class that had clear semantic divisions, etc. For the verbs, this processing was often more complex, especially for the VAI and VTA classes. Items were determined to not belong in one class or another based on it's central meaning of the action or entity. If the majority of group members pertained to smoking (a cigarette), a word describing smokiing meat (as food preparation) would not be placed in this group, as the essence of the action and its intended purpose diverged significantly from the rest of the group.

[5] As discussed in Wolvengrey (2005), Nêhiyawêwin sentences are devoid subject and objects in the usual sense. Instead, syntactic roles are defined by verbal direction alignment. For this reason, we use the terms *actor* and *goal* instead of *subject* and *object*.

[6] Although this gender dichotomy is *mostly* semantically motivated (e.g. nouns that are semantically inanimate are part of the inanimate gender) this is not always the case as in the word *pahkwêsikan*, 'bread', a grammatically animate word.

[7] Preliminary results for words not seperated by their conjugation class or declension did, in fact, create clusters based around these obvious differences. This likely due to the way definitions were phrased (e.g. dependent nouns would have a possessive determiner or pronoun).

Although most clusters produced somewhat cohesive semantic units, the largest clusters for the VAI and VTA classes acted as, essentially, catch-all clusters. Although computationally they seemed to have similar vector semantics, the relationship between items was not obvious to the human eye. Postprocessing for these clusters took substantial amounts of time and essentially comprised of using more cohesive clusters as a scaffold to fit words from these catch-all clusters into. In most cases, this resulted in slightly more clusters after postprocessing, though for VAIs this number was significantly higher, and for the NDIs it was slightly lower. Table 1 lists the number of cluster directly from HAC and from postprocessing.

Postprocessing grouped together words based on the most core semantic property of the word class: nouns were generally grouped based on the entity or state they represented, and verbs were generally grouped based on the most basic form action they represented. This is why, for example, `AI-cover` includes words for both covering and uncovering. In some cases a final class may seem like something that could be subsumed under another (e.g. `AI-pray` or `AI-cooking` might be understood as subsets of `AI-action`); however, in these cases, the subsumed class was judged to be sufficiently separate (e.g. *cooking* is an action of transforming resources into food for the purposes of nourishment, while verbs of `AI-action` are more manipulative, direct actions done for their own sake. Further, the automatic classification already grouped words in these ways, further justifying their separation. Finally, some grouping seem more morphosyntactic (e.g. `AI-reflexive`), though we argue that reflexivity, performing an action inwards, is in and of itself a salient semantic feature, and the inclusion of these terms into Wolvengrey (2001) indicates their lexicalization and distinction from the non-reflexive forms.

The actual quality of clustering varied form class to class. In general, nouns resulted in much more cohesive clusters out-of-the-box and required far less postprocessing. For example, nearly all of the HAC class NI_{14} items referred to parts of human bodies (and those that did not fit this description were terms clearly related to body parts like *aspatâskwahpisowin*, 'back rest'), NI_{13} was made up of trapping/hunting words and words for nests/animals.

The NA classes produced through HAC were similarly straightforward: NI_9 was made up of words for trees, poles, sticks, and plants; NI_8 was made up entirely of words form beasts of burden, carts, wheels, etc.; while much of NA_3 and NA_7, and nearly all of NA_2 referred to other animals. Once manually postprocessed, the NA lexemes settled into 8 classes: `NA-persons`, `NA-beast-of-burden`, `NA-food`, `NA-celestial`, `NA-body-part`, `NA-religion`, `NA-money/count`, and `NA-shield`.

The NDI and NDA classes required almost no postprocessing: NDA_1 and NDA_3 were each made up of various family and non-family based relationships, while NDA_2 was made up of words for body parts and clothing. The resulting classes for these were: `NDA-Relations`, `NDA-Body`, and `NDA-Clothing`.

The NDI lexemes basically took two classes: the vast majority of NDI forms referred to bodies and body parts while two lexemes referred to the concept of a house, resulting in only two classes: `NDI-body`, and `NDI-house`.

Verbs, on the other hand, required quite a deal more postprocessing. VIIs showed the best clustering results without postprocessing. For example, VII_6 was entirely made up of taste/smell lexemes, VII_7 was almost entirely weather-related, VII_8 contained verbs that only take plural subjects, VII_9 had only lexemes referring to sound and sight, and VII_10 had only nominal-like verbs (e.g. *mîsiyâpiskâw* '(it is) rust(y)'). Despite these well formed clusters, VII_1 through VII_5 were less cohesive and required manual clustering. In the end, distinct classes were identified: `II-natural-land`, `II-weather-time`, `II-sensory-attitude`, `II-plural`, `II-move`, `II-time`, and `II-named`.[8] Although postprocessing was required, this was not too substantial in scope or time.

The VAIs required significantly more work. Some classes were well defined, such as VAI_{23} whose members all described some sort of flight, but VAI_{12} contains verbs of expectoration, singing, dancing, and even

[8] The concepts of *weather* and *time* were combined here as many of the Nêhiyawêwin words for specific times also contain some concept of weather (e.g. the term for 'day' is *kîsikâw*, clearly related to the word for 'sky/heavens', *kîsik*; similarly, the word for 'night' is *tipiskâw*, which is the same word used for the night sky. Additionally, words like *pipon*, 'winter' and *sîkwan* 'spring' are obviously related to both time and weather.

	HAC classes	Manually Adjusted Classes	Lexemes
VII	10	6	581
VAI	25	13	5254
VTI	15	6	1825
VTA	20	7	1781
NI	15	13	3650
NDI	3	2	245
NA	10	8	1676
NDA	3	3	191

Table 1: HAC built cluster counts vs. counts after postprocessing

painting. The HAC classes were consolidated into 13 classes: `AI-state, AI-action, AI-reflexive, AI-cooking, AI-speech, AI-collective, AI-care, AI-heat/fire, AI-money/count, AI-pray, AI-childcare, AI-canine,` and `AI-cover.`

The VTIs similarly required manual postprocessing after HAC clustering. Although some classes such as VTI_{11} (entirely to do with cutting or breaking) or VTI_{14} (entirely to do with pulling) were very well formed, the majority of the classes needed further subdivision (though significantly less so than with the VAIs, resulting in the following 6 classes: `TI-action, TI-nonaction, TI-speech, TI-money/counter, TI-fit`, and `TI-food`.

Finally, the VTAs required a similar amount of postpreocessing as the VAIs. Although a few classes were well formed (such as VTA_4 which was entirely made up of verbs for 'causing' something), the vast majority of HAC classes contained two or more clear semantic groupings. Through manual postprocessing, the following set of classes were defined: `VTA_allow, VTA_alter, VTA_body-position, VTA_care-for, VTA_cause, VTA_clothes, VTA_cognition, VTA_create, VTA_deceive, VTA_do, VTA_existential, VTA_food, VTA_hunt, VTA_miss/err, VTA_money, VTA_move, VTA_play, VTA_restrain, VTA_religious, VTA_seek, VTA_sense, VTA_speech, VTA_teach, VTA_tire, VTA_treat-a-way, VTA_(un)cover`

4.1 Evaluation

In addition the above evaluation in the description of the manual scrutiny and adjustment of HAC results, which is in and of itself an evaluation of the technique presented in this paper (with single-subject experimentation proposed as a rapid path to data for less-resourced languages such as Vietnamese (Pham and Baayen, 2015)), we present a preliminary quantitative evaluation of this technique. This evaluation allows us to judge how useful these classes are in practical terms, providing an indirect measure of the informational value of the clusters. We make use of the mixed effects modelling that initially motivated automatic semantic clustering, focusing on a morphological alternation called Nêhiyawêwin Order, wherein a verb may take the form *ninipân* (the *Independent*) or *ê-nipâyân* (the *ê-Conjunct)*, both of which may be translated as 'I sleep.' The exact details of this alternation remain unclear, though there appears to be some syntactic and pragmatic motivation (Cook, 2014). Using R (R Core Team, 2017) and the `lme4` package (Bates et al., 2015), we ran a logistic regression to predict alternation using verbal semantic classes as categorical variables. In order to isolate the effect of semantic class, no other effects were used. The semantic classes were included as random effects. To assess the effectiveness of semantic class in this context, we assess the pseudo-R^2 value, a measure of Goodness-of-Fit. Unlike a regular R^2 measure, the pseudo-R^2 can not be interpreted as a direct measure of how much a model explains variance, and generally "good" pseudo-R^2 value are comparatively smaller (McFadden et al., 1973), though a higher value still represents a better fit. As a general rule, a pseudo-R^2 of 0.20 to 0.40 represents a well fit model. (McFadden,

	Manual	HAC-Only
VII	0.18	0.19
VAI	0.13	0.09
VTI	0.04	0.01
VTA	0.06	0.06

Table 2: pseudo-R^2 Values for Modelling Independent vs. ê-Conjunct Order Choice Based on Manual and Automatic Clustering Evaluation

1977)[9] Models were fit for each of the four conjugation classes for both classes produced directly from the Hierarchical Agglomerative Clustering as well those manually adjusted. We used a subset of the Ahenakew-Wolfart Corpus (Arppe et al., 2020), containing 10,764 verb tokens observed in either the Independent or ê-Conjunct forms. The resulting pseudo-R^2 scores represent the way in which automatic and semi-manual clusters can explain the Nêhiyawêwin Order alternation.

Table 2 presents the result of these analyses. the *Manual* column represents clusters that were manually adjusted, while the *HAC-Only* column represents the result of the logistic model that used only the fully automatic HAC-produced clusters. The manually adjusted and HAC-only classes performed similarly, especially for VTAs, though manual adjustment had a slightly worse fit for the VIIs, and conversely the VAI and VTI has somewhat significantly better fits using the manually adjusted classes. Although it appears that manual adjustment produced classes that were somewhat better able to explain this alternation, both manually adjusted and HAC-only clusters appear to explain a non-negligible degree of this alternation phenomenon in the above models. This is significant, because it shows that the result of the clustering techniques presented in this paper produce a tangible and useful product for linguistic analysis. Further, it suggests that, although manual classification was sometimes more useful, automatic classes more or less performed as well, allowing for researchers to determine if the added effort is worth the small increase in informational value. Nevertheless, alternative methods of evaluation, such as evaluating clusters based on speaker input, particularly through visual meas as described in Majewska et al. (2020) should be considered.[10]

5 Discussion

In general, the best clustering was seen in classes with fewer items. The VAI and NI lexemes required the most postprocessing, with each having roughly double the number of items as the next most numerous verb/noun class. Verb classes in general seemed to produce less cohesive classes through HAC. Although the exact cause of this discrepancy in unknown, it could perhaps be due to the way words are defined in Wolvengrey (2001). In this dictionary, verb definitions almost always contain more words than noun definitions. Almost every single verb definition will have at least two words, owing to the fact that Nêhiyawêwin verbs are defined by an inflected lexeme. This means that if one looks up a word like *walk*, it would appear as: `pimohtêw: s/he walks, s/he walks along; s/he goes along`. Meanwhile, nouns tend to have shorter definitions. The definition for the act of walking, a nominalized form of the verb for walk, is written as: `pimohtêwin: walk, stroll; sidewalk`. This difference is exacerbated by the fact that definitions are often translated fairly literally. Something like *pêyakwêyimisow* might be translated simply as 's/he is selfish,' but contains morphemes meaning *one*, *think*, *reflexive*, and *s/he*. A gloss of this word is seen in (1). Rather than simply defining the word as 's/he is selfish,' (Wolvengrey, 2001) has opted to provide a more nuanced definition: `pêyakwêyimisow: s/he thinks only of him/herself, s/he is selfish, s/he is self-centered`.

(1) pêyakwêyimisow
pêyakw-êyi-m-iso-w
one-think-VTA-RFLX-3SG
's/he thinks only of him/herself'

The result of this complex form of defining is that words are defined more in line with how they are understood within the Nêhiyawêwin culture, which is indeed often manifested in the derivational morphological composition of these words. This is central to the motivation for this method of semi-automatic clustering, but produces verbs with relatively long definitions. An alternative explanation for why Nêhiyawêwin lexemes with English definitions consisting of more numerous parts of speech were more difficult to classify is that these divisions simply have significantly more variation in

[9] One can also compare the results in this paper with results from a similar alternation study in Arppe et al. (2008).

[10] It is worth noting that previous attempts at such experimentation via Nêhiyawêwin communities with which we have good relationships have been poorly received by speakers.

meaning for whatever reason. Further investigation into this is needed.

Also worth noting is the relative distributions of each of the postprocessed classes mentioned above. Table 3 details each of the postprocessed noun classes sorted by their size.

Perhaps unsurprisingly, the distribution of lexemes into different classes followed a sort of Zipfian distribution. The `NA-person` and `NA-other-animals` accounted for the vast majority of noun lexemes for animate nouns. Just under half of all NI lexemes were nominalized verbs, and roughly a quarter were smaller object-like items (e.g. tools, dishes, etc.). The NDAs were almost entirely dominated by words for family, while all but three NDIs were body part lexemes. Some categories such as `NI-scent`, `NI-days`, and `NA-shield` have extremely low membership counts, but were substantially different from other categories that they were not grouped into another class. Most interestingly, there appeared to be three NI lexemes that referred to persons, something usually reserved for NAs only. These lexemes were *okitahamâkêw* 'one who forbids,' *owiyasiwêwikimâw* 'magistrate,' and *mihkokwayawêw* 'red neck.' In all three cases, the lexemes seem to be deverbal nouns (from *kitahamâkêw* 's/he forbids,' *wiyasiwêw* 's/he makes laws,' and *mihkokwayawêw* 's/he has a red neck.'

Verbs showed a similar distribution. Table 4 details the distribution of words within each of semantic classes for verbs. With the exception of VII and VAIs, verbs were dominated by classes for action, which subsumes most volitional actions (e.g. *kîskihkwêpisiwêw* 's/he rips the face off of people,' *kâsîpayiw* 's/he deletes'), and nonaction which includes most verbs of thought, emotion, judgment, or sensory action (e.g *koskowihêw*, 's/he startles someone,' *nôcîhkawêw* 's/he seduces someone'). Other classes may include action verbs, such as `AI-cooking` and `TI-speech`. Although these verbs could be classified in one of the two previously mentioned systems, their automatic classification and semantics unify them in a way that is unique to other items in these larger classes.

Overall, verb forms, especially the most numerous classes of VAI and VTA, required a large degree of manual postprocessing. Because this approach assumes no underlying ontology, but rather attempts to work bottom-up (cf. Hanks (1996)), the time taken to postprocess VAI and VTA classes is likely not too far from what it would take to manually classify these words based off a prebuilt ontology; however, the appeal of a bottom-up classification should not be overlooked, however. As an example, many ontologies place concepts like *thinking*, and *being happy* into separate classes; however, in our classification these words were combined into a single class of *cognition*. This is done because emotion words like *môcikêyihtam*, 's/he is happy (because of something)' (in addition to being verbs and not adjectives) contain a morpheme, {-êyi-}, meaning 'thought.' For these reasons, such emotion words are often translated as having to do specifically with thought and cognition: *môcikêyihtam*, 's/he thinks happily (because of something).' (Wolvengrey, 2001) uses these sorts of definitions, and so unsurprisingly the majority of such emotion words were classified in the proposed scheme together with words of thought. Where this was not the case, manual postprocessing from a bottom-up approach allows us to maintain the cultural understanding of emotions as directly related to cognition. Furthermore, from the experiential standpoint of one of the authors, the use of semi-automatic clustering produces a kick-start that greatly aids to the starting of a semantic classification task, especially for non-native speakers.

6 Conclusion

This paper describes an attempt at, for the first time, semi-automatically classifying Nêhiyawêwin verbs and nouns. The process used in this paper is easily applied to any language that makes use of a bilingual dictionary with definitions written in a more resourced language. Resulting clusters of Nêhiyawêwin words are freely available on the online. Although the technique worked quite well with nouns, which required very little manual adjustment, verbs required more directed attention. Despite this, the technique presented in this paper offers a bottom-up, data-driven approach that takes the language on its own terms, without resorting to ontologies created primarily for other languages. If, however, one wishes to use a pre-defined ontology, the basis for this work (representing word definitions using pre-trained English word vectors) could be used in conjunction with existing ontologies to expedite the classification process. For example, Dacanay et al. (2021) compare the naive definition vectors for Wolvengrey (2001) with the same for the English WordNet word senses; word senses

NI (N)	NDI (N)	NA (N)	NDA (N)
NI-nominal (1783)	NDI-body (243)	NA-persons (720)	NDA-relations (143)
NI-object (902)	NDI-house (2)	NA-beast-of-burden (512)	NDA-body (45)
NI-natural-Force (283)		NA-food (325)	NDA-clothing (4)
NI-place (228)		NA-celestial (45)	
NI-nature-plants (198)		NA-body-part (37)	
NI-body-part (78)		NA-religion (23)	
NI-hunt-trap (60)		NA-money/count (12)	
NI-animal-product (48)		NA-shield (2)	
NI-religion (36)			
NI-alteration (23)			
NI-scent (4)			
NI-days (4)			
NI-persons (3)			

Table 3: Manually Adjusted Noun Classes

VII (N)	VAI (N)	VTI (N)	VTA (N)
II-natural-land (256)	AI-state (2083)	TI-action (1409)	TA-action (1013)
II-weather-time (103)	AI-action (1982)	TI-nonaction (293)	TA-nonaction (574)
II-sensory/attitude (92)	AI-reflexive (542)	TI-speech (80)	TA-speech (103)
II-plural (73)	AI-cooking (172)	TI-money/count	TA-food (54)
II-move (35)	AI-speech (131)	TI-fit (10)	TA-money/count (23)
II-named (3)	AI-collective (97)	TI-food (8)	TA-religion (9)
	AI-care (81)		TA-allow (5)
	AI-heat/fire (55)		
	AI-money/count (34)		
	AI-pray (29)		
	AI-childcare (17)		
	AI-canine (16)		
	AI-cover (15)		

Table 4: Manually Adjusted Verb Classes

whose vectors bear a strong correlation with the Nêhiyawêwin definitions can then be assumed to be semantically similar with a Nêhiyawêwin word, and the latter can take the WordNet classification of the former. Further research should investigate more sophisticated methods of creating embeddings, especially the use of true sentence vectors. Additionally, one could consider using weights for English words in the definitions of *Nêhiyawêwin* words based on measures like tf-idf. Over all, this technique provided promising results. Regardless of the language or particular implementation, this technique of bootstrapping under-resourced language data with pre-trained majority language vectors (for which very large corpora exist), should not be restricted by the sizes of dictionaries in the under-resourced language, as the underlying vectors are trained on a 100 million word English corpus.

Acknowledgements

We would like to thank the Social Sciences and Humanities Research Council of Canada for funding this research. We would also like to thank Dr. Arok Wolvengrey for providing his dictionary source for this study.

References

Antti Arppe, Katherine Schmirler, Atticus G Harrigan, and Arok Wolvengrey. 2020. A morphosyntactically tagged corpus for plains cree. In *Papers of the Forty-*

Ninth Algonquian Conference. Michigan State University Press.

Antti Arppe et al. 2008. Univariate, bivariate, and multivariate methods in corpus-based lexicography: A study of synonymy.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Joan Bresnan, Anna Cueni, Tatiana Nikitina, and R Harald Baayen. 2007. Predicting the dative alternation. In *Cognitive foundations of interpretation*, pages 69–94. KNAW.

Clare Cook. 2014. *The clause-typing system of Plains Cree: Indexicality, anaphoricity, and contrast*, volume 2. OUP Oxford.

Daniel Dacanay, Antti Arppe, and Atticus Harrigan. 2021. Computational Analysis versus Human Intuition: A Critical Comparison of Vector Semantics with Manual Semantic Classification in the Context of Plains Cree. In *Proceedings of the 4th Workshop on the Use of Computational Methods in the Study of Endangered Languages*, volume 1, pages 33–43.

Dagmar Divjak and Stefan Th Gries. 2006. Ways of trying in russian: Clustering behavioral profiles. *Corpus linguistics and linguistic theory*, 2(1):23–60.

Birgit Hamp and Helmut Feldweg. 1997. Germanet-a lexical-semantic net for german. In *Automatic information extraction and building of lexical semantic resources for NLP applications*.

Patrick Hanks. 1996. Contextual dependency and lexical sets. *International journal of corpus linguistics*, 1(1):75–98.

Beth Levin. 1993. *English verb classes and alternations: A preliminary investigation*. University of Chicago press.

Olga Majewska, Ivan Vulić, Diana McCarthy, and Anna Korhonen. 2020. Manual clustering and spatial arrangement of verbs for multilingual evaluation and typology analysis. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4810–4824.

Daniel McFadden. 1977. Quantitative Methods for Analyzing Travel Behaviour of Individuals: Some Recent Developments. Technical report.

Daniel McFadden et al. 1973. Conditional logit analysis of qualitative choice behavior.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

George A Miller. 1998. *WordNet: An electronic lexical database*. MIT press.

Hien Pham and Harald Baayen. 2015. Vietnamese compounds show an anti-frequency effect in visual lexical decision. *Language, Cognition and Neuroscience*, 30(9):1077–1095.

R Core Team. 2017. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Linda Visitor, Marie-Odile Junker, and Mimie Neacappo. 2013. Eastern james bay cree thematic dictionary (southern dialect). *Chisasibi: Cree School Board.*

Shan Wang and Francis Bond. 2013. Building the chinese open wordnet (cow): Starting from core synsets. In *Proceedings of the 11th Workshop on Asian Language Resources*, pages 10–18.

Arok Wolvengrey. 2001. *Nēhiyawēwin: itwēwina = Cree: words.* University of Regina press.

Arok Wolvengrey. 2005. Inversion and the absence of grammatical relations in plains cree. *Morphosyntactic expression in functional grammar*, 27.

Liang-Chih Yu, Jin Wang, K Robert Lai, and Xuejie Zhang. 2017. Refining word embeddings for sentiment analysis. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 534–539.