UNIVERSITY OF ALBERTA

A Comparative Analysis of Manual and Vector Semantic Organisation using a Bilingual Dictionary of Plains Cree

BY

Daniel Benedict Dacanay

A THESIS

SUBMITTED TO THE FACULTY OF ARTS

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF

BACHELOR OF ARTS

DEPARTMENT OF LINGUISTICS

EDMONTON, ALBERTA

May, 2022

UNIVERSITY OF ALBERTA

FACULTY OF ARTS

UNIVERSITY OF ALBERTA
FACULTY OF ARTS

The undersigned certify that they have read and recommend to the Faculty of Arts for acceptance, a thesis entitled A Comparative Analysis of Manual and Vector Semantic Organisation using a Bilingual Dictionary of Plains Cree submitted by Daniel Benedict Dacanay in partial fulfilment of the requirements for the degree of Bachelor of Arts.

…….
Dr. Antti Arppe

…….
For the Department

“Слова … преломляются во всем, кроме самих слов.
Слова ничего не значат — слова — это вода”

“Words … are refracted in everything apart from the words themselves.
The words mean nothing - words are water”

- Andrei Tarkovsky, February 4, 1974,
Time Within Time: The Diaries 1970-1986

ABSTRACT

This thesis shall address the nature of, and various possible approaches to, semantic classification in a bilingual dictionary setting, in this instance, in that of a low-resource language (Plains Cree/*nêhiyawêwin*, ISO:crk). In doing this, we shall discuss the distinct, yet partially overlapping advantages of manual and computationally-generated semantic classifications, the methodologies and resources involved in implementing each, and the general quality of results to be expected in either instance. Through this, we will also outline a variety of possible related approaches to ontology-based vector semantic classification (or, organisation), as well as outline the possible uses of current vector semantic classification results from these methodologies. Finally, the nature of semantic classification using linguistically 'neutral' semantic classification ontologies, and the various advantages and disadvantages thereof, are to be discussed throughout.

ACKNOWLEDGEMENTS

Foremost of all, I must thank my supervisor, Dr. Antti Arppe, to whom I am most indebted for the conceptual inspiration and methodological execution of this thesis. His continual counsel has served as the foundation upon which my research (present, past, and future) has sought to sculpt itself, its essence ever in facsimile of his skill and wisdom. If not for him, these pages would be blank, and this project a vague and unrealised mirage.

I must also lend my most gracious acknowledgements to Dr. Arok Wolvengrey, not only for his continual allowances in my use of his painstakingly-created dictionary database, the construction of which alone has spanned longer than my existence, but also for his peerless expertise in Cree culture and language, and his deigning to assist me in my comparatively humble endeavours therewith.

My gratitude must also be expressed to the wise and boundlessly commendable Rose Makinaw, for her tireless devotion to the betterment of the Cree language, and for all that I have learned from her.

Katie Schmirler, whom I have oft consulted for her insights in the Cree lexicon, must be commended here for her great patience in assisting my research, as must Atticus Harrigan, whose work in Cree vector semantics laid the basis for my own.

Finally, I express my gratitude to the Social Sciences and Humanities Research Council, whose grant funding was instrumental in the undertaking of this project.

TABLE OF CONTENTS

TABLES AND FIGURES

CHAPTER 1. INTRODUCTION

1.1 Semantically Organised Lexica and Ontological Semantic Classification in Theory

The classification of large-scale lexical resources such as dictionaries on semantic lines is not a novel concept; as early as the second century CE, Philo of Byblos had written a basic thesaurus of Greek in *On Synonyms*, and by the mid 19th century, the influential Roget's Thesaurus, compiled by the eponymous Peter Mark Roget, had brought semantic classification, or rather, semantic organisation according to set classes, to the mainstream (Hüllen 2009). From a practical standpoint, semantic organisation, as opposed, for example, to alphabetical organisation, makes a great deal of sense for a lexical resource such as a dictionary; the general consensus of modern psycholinguistics is that the mental lexicon, at least in many capacities, is broadly grouped along semantic (and not orthographic) lines (Collins & Loftus 1975; Anderson 1996; Miller et al. 1993; Fellbaum 2000; Marslen-Wilson et al. 2008; Lucas 2001). For example, Marslen-Wilson & Zwitserlood (1989) found that, in Dutch, priming participants with the word *honing* ('honey') improved their recognition time for the word *bij* ('bee'), despite the two forms having no morphological, phonological, or orthographic relation. In addition to this psycholinguistic justification, the existence of large-scale semantic classifications can also facilitate the undertaking of various academic pursuits involving lexical semantics, such as facilitating vocabulary retrieval to study cross-linguistic lexicalisation patterns and lexical density (e.g. Talmy 1985), as well as being of use in a pedagogical context to better allow learners and instructors to collect domain-relevant vocabulary. Additionally, such classifications can facilitate the creation of various natural language processing applications, for example, using semantic classifications, one can create digital dictionary search methods which can return semantically

related vocabulary to the target word, even for search queries which are not in the dictionary at all (Arppe et al. in prep.); resources such as these can also serve as early developmental stages for machine translation, as well as being of use in improving the accuracy of spellcheckers (King & Dickinson 2014).

Beneficial though they may be, the typical process of acquiring semantic classifications on a large scale (that is, manual semantic annotation) is an arduous and long-winded task, often requiring months or even years to complete for a sizable lexicon (Bosch & Griesel 2017; Dacanay et al. 2021a). Although not necessarily difficult, the simple time-commitment of attempting a fully-manual semantic classification of any dictionary is sufficient to make such classifications a non-trivial, if not untenable, challenge for many low-resource language communities, a fact which, if adhering to strictly traditional methods, would bar them access to the development of the aforementioned language tools. For this reason, it is proposed here that the task of semantic classification, or at least substantial portions of it, may be effectively undertaken not as a manual endeavour, but as a computational one, leveraging freely available recent NLP technologies for majority languages, namely, vector semantics, to expedite the task to the degree of feasibility (in terms both of the necessary material and temporal resources) for reliable application on low-resource languages. As a practical demonstration of this, full-scale semantic classifications will be carried out on a lexical resource in Plains Cree (*nêhiyawêwin*, ISO: crk), a low-resource Indigenous language of Western Canada, using both traditional manual classification methods and various computational methods employing vector semantics, allowing for an informed comparison between both the results and the practical implementations of both approaches.

1.2 Plains Cree

Plains Cree (known endonymically as *nêhiyawêwin*, or in some circles as Cree y-dialect) is an Algonquian language spoken throughout Alberta and Saskatchewan, as well as in parts of the Northwest Territories and Montana. Plains Cree is the most widely-spoken member language of the Cree dialect continuum, both geographically and demographically, having a speakerbase estimated somewhere between 3070 and 33 975 (Ethnologue 2015; Statistics Canada 2017). This population makes Plains Cree one of the most populous Indigenous languages in Canada, with Plains Cree speakers making up as much as a third of the total ~116 000 Cree speakers in the country, and nearly a sixth of the total 228 000 Aboriginal language speakers (Statistics Canada 2017). Typologically, Cree is a highly polysynthetic language, with most inflectional and derivational morphology centred around verbs, and adjectival and adverbial meanings encoded either using lexical affixes on nouns and verbs or through intransitive, stative verbs, rather than as distinct, independent parts of speech. Using this morphology, Cree verbs often convey complex meanings only expressible in more isolating languages (such as English) through full clauses or sentences; for example, the Cree word *akwanâhkwêsin*, which can be defined as ‘s/he lies with his/her face covered’, or *mâci-ayamihcikêw*, defined as ‘s/he starts reading’ (Wolvengrey 2011).

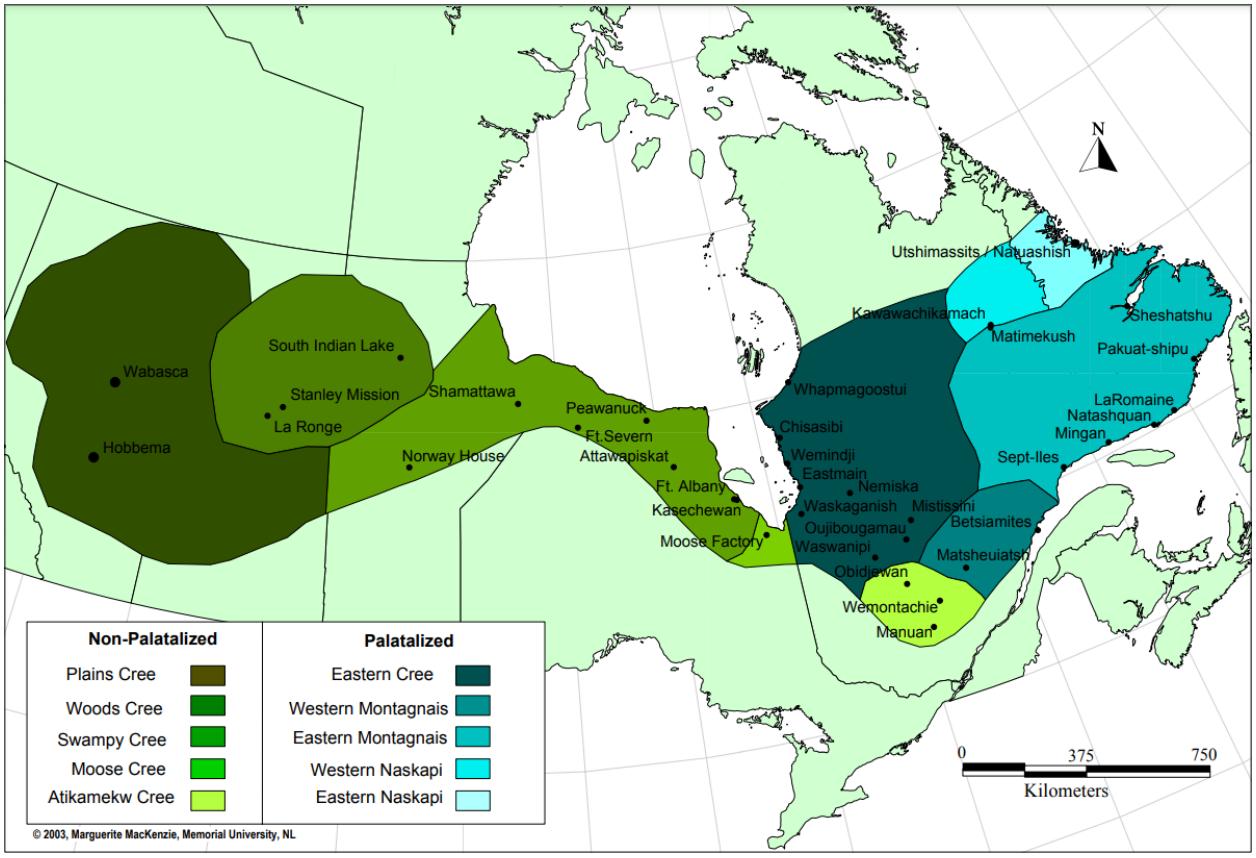

Figure 1, a map of the Cree dialect continuum, with Plains Cree representing the westernmost significant dialect group (Junker 2018).

Despite its relatively wide speakerbase, Plains Cree remains faced with largely the same threats facing Indigenous languages across North America in general; namely, a relatively low level of intergenerational transmission and a notable scarcity of representation in contemporary forms of linguistic expression and communication, such as digital media (although this is slowly changing (Arppe et al. 2016)). Despite being relatively well-documented lexicographically, with eight published dictionaries ranging in length from ~6000 to ~21 000 entries (Tremblay 2005), contemporary lexical resources for Plains Cree remain primarily analogue, and online representation of the language is exceedingly sparse; for example, Plains Cree has only 460 total entries on Wiktionary, compared, for instance, to 563 for Toki Pona, an experimental minimalist

conlang created in 2001 with as few as 165 fluent speakers, and 2463 for Klingon, another conlang with as few as 30 fluent speakers (Wiktionary 2022).

This minimal digital presence poses a substantial threat to the continued longevity of Plains Cree as a 21st century language, relegating access to critical language resources such as dictionaries to only those able to travel to read physical copies, and serving as a major obstacle to those wishing to engage in self-study of the language without access to native speakers. The utility of natural language processing technologies in linguistic revitalisation efforts and in making endangered languages "viable in the web and digital world" (Jokinen et al. 2016) has been extensively noted (Meighan 2021); for example, in the case of Hawaiian, a language whose revitalisation efforts since the 1980s have been broadly categorised as successful (Cowell 2012; Eisenlohr 2004), the development of digital language archives, bilingual and monolingual online forums (Warschauer 1998), custom native search engines (Donaghy 1998), and other web infrastructure has consistently been given high priority as means to "engage youth to learn their language" (Galla 2009), with Cowell (2012) noting the effect of this digital representation in creating "yet another new domain for Hawaiian language usage, and effectively ideologiz[ing] the language as modern and youth-oriented". Similar digitally-focused revitalisation efforts have been undertaken by linguists working with Sami languages in Northern Europe (Outakoski et al. 2018), Maori in New Zealand (Keegan & Manuirirangi 2011; Solano et al. 2018), and Inuktitut in the Canadian Arctic (Tan & Sadat 2020). Following in these successful examples, the necessity of an expanded online presence for Plains Cree, supported by a robust network of digital language tools, becomes apparent as a matter of foremost concern for the continued survival and daily usage of

the language in the 21st century (Arppe et al. 2016; Littell et al. 2018), with basic resources such as semantic organisations of the lexicon being critical in the development of such tools.

1.3 *Cree: Words / nêhiyawêwin: itwêwina*

The largest current dictionary of Plains Cree is *Cree: Words / nêhiyawêwin: itwêwina* (abbr. CW), a bilingual Cree-English dictionary compiled throughout the late 1990s and early 2000s by Dr. Arok Wolvengrey (2011), and available presently in both print and digital forms. The underlying digital database for the dictionary is continually updated, consisting at the time of writing of 21 345 entries (5212 nouns, 13 669 verbs, and 2464 affixes, particles, etc.), with each Cree entry having an English definition and part-of-speech code, as well as various morphological notes.

Entry Word (SRO)	Entry Word (Syllabics)	Part-of-Speech	English Definition	Stem	Derivation
amisk	ᐊᒥᖽᐟ	NA-3	beaver	amiskw-	amiskw
amiskwayân	ᐊᒥᖽᑲᔭᐢ	NA-1	beaver-pelt	amiskw-ayân	amiskw- + /-wayân/
wâpamêw	ᐋᑊᐸᒣᐤ	VTA-1	s/he sees s.o., s/he witnesses s.o.	wâpam-	/wâp-/ + /-am/

Table 1, a demonstration of the structure of several entries from CW. Of relevance to our investigation are the Standard Roman Orthography (SRO) representation of the Cree word, the English definition, and the part-of-speech code (which also indicates inflectional subcategories such as animacy).

Although CW is the largest currently available Plains Cree dictionary, various others, including some with existing implementations of semantic classification, can be found; namely, the *Maskwacis Cree Dictionary* (2009), containing 8986 entries, had its lexical contents semantically

classified into domains using the SIL Rapid Word Collection Methodology (see Section 1.6) by a group of undergraduate students at the University of Alberta in 2014[1]. By contrast, CW remains organised primarily alphabetically; although CW does indeed have some existing *ad hoc* semantic categorisations grouping its entries, the extent of these classifications is extremely limited, with only about 6.1% of entries (1303 in total) being categorised in this way. These classifications were largely idiosyncratic, and not made in accordance with any specific classification scheme (Wolvengrey, personal correspondence, 2020); as such, for the purposes of our investigation, they were ignored.

1.4 Semantic Classification in Practice

When semantically classifying a language's vocabulary using an existing, pre-compiled lexical resource (such as a dictionary), the traditional method is a fairly simple one, involving obtaining the resource in a usable form, identifying all of the target language vocabulary within that resource, and then either clustering semantically related vocabulary into *ad hoc* groups or assigning each of them to preset semantic categories in a semantic classification scheme or ontology of some kind. In the latter approach, the resultant semantic categories are often arranged hierarchically based on some semblance of hypernymy and hyponymy, with the end result often resembling a semantic 'tree', beginning with one or several extremely general nodes (often representing lexical instantiations of semantic primes such as MOVE or THING (Bundy & Wallen 1984)), and radiating outwards into increasingly specific categories as one descends the tree. The exact semantic categories which are used as nodes generally varies depending both on the intent of the linguist and on the perceived nature of the content of the target language's

[1] These students being Megan Bontogon, Sarah Lamarche, and Elizabeth Pankratz

lexicon; as such, in creating *ad hoc* semantic classifications for a language, no one linguist's classifications will ever exactly match another's, even if classifying an identical lexical resource, let alone when classifying the lexica of entirely different languages. This simple fact has often resulted in different languages and language communities creating semantic classification schemes, or semantic ontologies, with entirely different structural principles, complicating the process of cross-linguistic comparisons of semantic content and dissuading many linguists from attempting such classifications in the first place (Stutzman & Warfel 2022). As such, there have been a number of attempts to create language-neutral, 'universal' semantic classification structures, allowing both for linguists to be able to semantically organise a given language's lexicon without the need to construct an entirely new system of semantic classifications, as well as facilitating a greater degree of ease in comparing the semantic content of different, often unrelated languages or lexica. We discuss here two such 'language-neutral' ontologies, namely, the Princeton WordNet and the SIL Rapid Word Collection Methodology.

1.5 WordNet

Perhaps the most successful 'language-neutral' classification ontology[2] has been the Princeton WordNet (abbr. WN), a hierarchical semantic classification system first used to semantically organise the English lexicon in the early 1990s (Miller et al. 1993; Fellbaum 2000), which has since been adopted as something of an international standard for semantic classification, with WordNets of various sizes existing for hundreds of languages, ranging from major international languages such as Arabic (Black et al. 2006) and German (Hamp & Feldweg 1997) to regional

[2] Although WordNet is not an ontology in the strictest philosophical sense, rather being a "description of lexical knowledge" which, by virtue of its semantic breadth, takes on "many similarities" with one (Miller & Hristea 2006), for reasons of parsimony, we shall refer to it in this paper as a 'semantic classification ontology', a term which we will also use to describe RW in Section 1.6

minority languages such as Scottish Gaelic (Bella et al. 2020) and Mansi (Horváth et al. 2016), as well as including multilingual, comparative databases such as EuroWordNet (Vossen 1998; Vossen 2004). Correspondingly, WordNets differ widely in terms of their size, from the 155 327 entry English WordNet, to the 8412 entry Northern Sotho WordNet (Bosch & Griesel 2017). However, one relative constant in creating a WordNet is the complexity of the relationships represented within it; a fully elaborated WordNet, for example, models hypernymy, hyponymy, synonymy, antonymy, meronymy, homonymy, gradation, and entailment, among other relationships between its entries (Miller et al. 1993).

The basic entry classification unit within WordNet is the *synset* (or, synonym set), a set 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 et al. 1993). An individual synset consists internally of all synset members, a definition, and optionally one or several example sentences (see Figure 2). Among themselves, these synsets are then divided according to their part-of-speech (the four in English WN being nouns, verbs, adjectives, and adverbs), with the structure of and represented relationships between different synsets differing according to this part-of-speech. For example, all nouns (of which there are a total of 117 097 across WN) are contained on a single hierarchical tree originating with the synset *(n) entity#1* and radiating downwards with hyponyms of increasing specificity (*(n) physical entity#1*, *(n) abstraction#6*, etc.). Verbs (of which there are 11 488) are represented similarly, but are spread out across several hundred, smaller hierarchical trees, rather than being consolidated entirely on one. Adjectives (of which there are 22 141) and adverbs (of which there are 4601), by contrast, are represented by various dipolar sets of antonyms clustered by similarity. Function words such

as determiners are excluded from WN's structure on the basis that, in the mental lexicon, they are "probably stored separately as part of the syntactic component of language" (Miller et al. 1993).

Based on these organisational principles, a structurally complete WordNet for all nominal and verbal synsets can be constructed using only the relationships of synonymy, hypernymy, and hyponymy, described as "the central organizing principle" of WordNet as a whole (Miller et al. 1993), as all such synsets are bound to at least one other synset of the same part-of-speech by minimally one of these relationships. All other relationships modelled in a full WordNet may thus be considered secondary to its core structure.

In order to represent polysemy, all WordNet synsets (whether ambiguous or otherwise) have 'sense numbers', denoting the particular word sense indicated by a synset's contents. The order of these numbers for the senses of an ambiguous wordform is entirely arbitrary, with lower numbers not necessarily representing more common senses. Additionally, these numbers reset for homographic synset heads across parts of speech. For example:

- *(n) punch#1, clout#4, poke#5, lick#3, biff#1, slug#8* ((boxing) a blow with the fist) "I gave him a clout on his nose"
- *(n) punch#2* (an iced mixed drink usually containing alcohol and prepared for multiple servings; normally served in a punch bowl)
- *(n) punch#3, puncher#3* (a tool for making holes or indentations)

- *(v) punch#1*, *plug#3* (deliver a quick blow to) "he punched me in the stomach"

- *(v) punch#2* (drive forcibly as if by a punch) "the nail punched through the wall"
- *(v) punch#3*, *perforate#1* (make a hole into or between, as for ease of separation) "perforate the sheets of paper"

One final factor of note on WordNet is the nature of its semantic content. WordNets are intended to be more-or-less exhaustive semantically-organised compilations of a language's vocabulary, with no explicit minimal requirements of frequency of use in order for a synset to be added; as such, WordNet synsets contain many low-frequency, often highly specific words, as well as synsets for proper nouns which are often more encyclopaedic than lexicographic in nature, describing historical events, figures, and locations (see Section 4.2.2).

1.6 Rapid Words

Although WordNet provides a rigorous and nuanced framework for large-scale semantic classification, WordNet's structure is also marked with a great deal of internal complexity, often resulting in full WordNets taking years to construct, even when assisted by dozens of trained linguists and native speakers (Bosch & Griesl 2017). As such, WordNet is often unsuitable for use in the creation of first-pass semantic classifications, such as those performed by linguists gathering and organising data in the field, and there exist many smaller-scale, reduced semantic ontologies whose designs are tailored towards covering the breadth of a language's lexicon while still remaining simple enough to feasibly be applied in a matter of months or weeks. It was precisely these design constraints which governed the creation of SIL's Rapid Word Collection Methodology, or Rapid Words (abbr. RW) (Moe 2003). Containing 1789 general semantic domains, hierarchically organised under nine high-level semantic categories, such as *1. Universe,*

creation and *4. Social Behavior*, Rapid Words' design as an aid for dictionary vocabulary elicitation lends itself to an additional, retroactive utility as a means of semantically organising the eventual dictionary itself, a purpose for which it has been used numerous times in the past. Within SIL's own documentary archive (webonary.org), for example, many of the languages documented have their lexica organised using Rapid Words domains (to give two examples, Buli (Kröger 2021) and Marwari (Dewra & Dailey 2015)).

As previously mentioned, Rapid Words has also been applied as a classification scheme to existing lexica of Plains Cree, having been used by a group of undergraduate students at the University of Alberta in 2014 to classify the contents of the *Maskwacis Cree Dictionary* (2009). Rather than being created to serve as an organisational framework for a semantic dictionary however, these classifications were primarily compiled to aid in partitioning vocabulary for a series of elicitation sessions intended to gather audio recordings of novel and existing lexical items from Cree native speakers for a web-based spoken dictionary (Littlechild et al. 2018; Reule 2018); correspondingly, although the resultant audio from these elicitation sessions has been published (being available, for example, though the University of Alberta's https://itwewina.altlab.app/), the semantically classified version of the Maskwacis Dictionary itself has not. Nonetheless, the decision to apply RW as a means of semantic classification for CW was partially motivated by this earlier, successful use of the ontology, and should additionally permit for later comparative studies of the semantic contents of both dictionaries.

Unlike WN, RW domains are not bound by part-of-speech, nor is their hierarchical organisation explicitly defined by any particular semantic relationship(s), instead being arranged

pragmatically according to topics which are most likely to yield novel vocabulary. Each RW domain is subdivided into elicitation questions pertaining to that domain (see Figure 2), with sample answers provided for each such question in English. On account of the ontology's smaller overall size, RW domains are typically much more semantically general than their correspondent WN synsets; a comparison of the internal structure and degree of specificity of both ontologies is demonstrated on the following page:

(n) entity#1 (that which is perceived or known or inferred to have its own distinct existence (living or nonliving))
- ***(n) physical entity#1*** (an entity that has a physical existence)
  - ***(n) thing#12*** (a separate and self-contained entity)
    - ***(n) part#3***, ***(n) piece#3*** (a portion of a natural object) "they analyzed the river into three parts", "he needed a piece of granite"
      - ***(n) body_part#1*** (any part of an organism such as an organ or extremity)
        - ***(n) external_body_part#1*** (any body part visible externally)
          - ***(n) extremity#1***, ***(n) appendage#1***, ***(n) member#3*** (an external body part that projects from the body) "it is important to keep the extremities warm"
            - ***(n) limb#1*** (one of the jointed appendages of an animal used for locomotion or grasping: arm; leg; wing; flipper)
              - ***(n) thigh#1*** (the part of the leg between the hip and the knee)

- ***2. Person***
  - ***2.1 Body***
    - ***2.1.3 Limb***
      - ***2.1.3.2 Leg***

Use this domain for parts of the leg and foot

What general words refer to the entire leg?
- leg

What are the parts of the leg?
- upper leg, groin, thigh, knee, kneecap, lower leg, calf, shin …

What words refer to a part of a leg when it is in a particular position?
- lap

What words describe a person's legs?
- pigeon toed, knock-kneed, bow-legged, flat-footed

Figure 2, a diagram demonstrating some basic structural differences between WordNet (top) and Rapid Words (bottom) through their encoding of the English word 'thigh'; in WordNet, *(n) thigh#1* is considered a hyponym of *(n) limb#1*, which is itself a hyponym of *(n) extremity#1*, and so on, whereas in Rapid Words, 'thigh' fits into the semantic domain of *2.1.3.2 Leg* (specifically into the elicitation question 'What are the parts of the leg?').

CHAPTER 2. USING WORDNET AND RAPID WORDS FOR MANUAL SEMANTIC CLASSIFICATION

2.1 WordNet as a Means of Semantic Classification

As previously mentioned, creating even a fairly modest WordNet which encodes all permitted semantic and morphological relationships with all of its vocabulary can easily take years; for example, the African WordNet Project, which created fully elaborated WordNets for five low-resource Bantu languages, took over 8 years to construct a set of WordNets ranging in size from ~8000 - 15 000 entries, and at that, with the aid of several teams of native-speakers with linguistic training. By contrast, in the context of Plains Cree, an eminently endangered language whose researchers often have only irregular access to native speakers, this approach is largely infeasible on any realistic timescale with currently available resources.

As such, in this investigation, we elected to exploit the aforementioned fact that a structurally coherent WordNet can be constructed for nouns and verbs using only hypernymy, hyponymy, and synonymy. This resulted in the creation of a simplified, ‘skeletal’ WordNet, a set of all WordNet synsets connected only by the ‘central’ hypernymy hierarchy and their synset-internal synonymy. To apply this as a classification scheme, since WordNet’s basic structure is intended to be language neutral, we simply co-opted the hypernymy hierarchy for the existing entries in English WordNet and used these synsets as categories for semantic classification, with the new, Cree language skeletal WordNet being automatically populated through classifying entries in the target language source into their appropriate corresponding places in the co-opted WordNet hierarchy. This process of using target language vocabulary to populate the structural backbone of an existing WordNet, rather than constructing a new hypernymy hierarchy specifically for the

target language within WordNet constraints, is referred to as the ‘merge approach’ (Bosch & Griesel 2017; Vetulani et al. 2010), and its utility in reducing the time necessary to construct a usable WordNet is well attested (Vincze & Almási 2014).

The ramifications of using the underlying structure of a semantic classification ontology of English (language-neutral though it may claim to be) to semantically classify Cree vocabulary are non-trivial, and shall be discussed throughout Chapters 4 and 6, but as both a pragmatic alternative to spending years constructing a fully Cree WordNet structure from the ground up and a means of enabling semantic classification of Cree vocabulary without being (or having consistent access to) a fluent speaker (given that the semantic categories of WordNet are already labelled in English), the merge approach proved the most feasible solution to applying WordNet as a means of semantic organisation for the vocabulary in CW.

2.2 Rapid Words as a Means of Classification

On account of its smaller size and simpler overall structure, it was not necessary to make any modifications to Rapid Words for the manual or vector classifications. Instead, much like with WN, each RW domain was taken as a classification category in and of itself, with sufficiently similar Cree words, regardless of part-of-speech, grouped together and classified into their respective domain(s). This basic classification strategy is identical to the method used to semantically classify the *Maskwacis Cree Dictionary* using RW, as previously mentioned.

2.3 Alternative Ontologies

As a final note, it must be included here that there do, in fact, exist semantic classification ontologies designed specifically to cater to languages in the Cree dialect continuum. Namely, the *Eastern James Bay Cree Thematic Dictionary* (Visitor et al. 2013) was compiled with its entries organised according to a purpose-built structure of semantic relations, created with input from Cree native speakers, intending to reflect the actual relative lexical density of words within various semantic domains in typical spoken Cree, rather than to have a more-or-less evenly distributed structure, such as that of WN or RW. Correspondingly, this classification scheme, at 241 categories and only three levels of quasi-hypernymic depth, is also much smaller than either of the aforementioned 'general-purpose' ontologies. Although this Cree-specific semantic classification ontology may have represented a more accurate language-internal perspective as to how semantic categories are perceived by Cree speakers, we elected not to make use of it in this investigation due to its potential limits on transferability. Since the classification ontology of Visitor et al. (2013) is explicitly designed for, and has only ever been used to classify, vocabulary of Cree, it would be more difficult to accurately compare classifications made in this system to semantically classified resources of other, typologically different languages, reducing the use of the classifications for cross-linguistic semantic study. Additionally, our manual and vector classification methods (as will be shown) were designed to not be explicitly bound to Plains Cree, and to be theoretically applicable with minimal modification to bilingual dictionaries of any language, provided they have majority-language glosses. The use of the East Cree ontology, however, would limit the potential cross-linguistic applicability of the method to within the Algonquian family, and in a practical sense, to comparison only with the content of the existing *Eastern James Bay Cree Thematic Dictionary*. As such, although this East Cree ontology is

certainly a useful resource to the Algonquian semanticist, it was not used in the present investigation.

2.4 Basic Method of Classification

The actual process of classification for the CW content was as follows; for each entry in the CW dictionary, the one or several English WordNet synsets which were the most closely semantically related to the meaning of the given Cree word were chosen as classification(s). The part-of-speech of the English synsets was ignored, and classifications were made purely on the basis of lexical semantic proximity. If the full meaning of any given CW entry could not be satisfactorily expressed with a single synset, as many were used as were necessary to cover a fuller breadth of the original Cree word's meaning. However, if the meaning of the Cree word could be reasonably approximated with only a single synset, only a single synset was used. For the Rapid Words classifications, largely the same method was applied. For example:

Cree Entry	Code	English Definition	WN Manual Classification(s)	RW Manual Classification(s)
kinosêw	NA-2	fish	*(n) fish#1*	*1.6.1.5, Fish*
kinosêskâw	VII-1v	there is an abundance of fish	*(n) fish#1* and *(adj) abundant#1*	*1.6.1.5, Fish*
kinosêwêw	VAI-1	s/he fishes, s/he catches fish	*(v) fish#2*	*6.4.5, Fishing*
kinosêwiw	VAI-1	s/he is a fish	*(n) fish#1*	*1.6.1.5, Fish*
kinosêwimâkosiw	VAI-1	s/he smells fishy	*(adj) fishy#1* and *(v) smell#2*	*1.6.2.3, Parts of a Fish*
iýinito-kinosêw	NA-2	ordinary fish; pike, jackfish	*(n) fish#1* and *(n) pike#2*	*1.6.1.5, Fish*

osîhikinosêwêw	VAI-1	s/he prepares his/her own fish, s/he processes his/her own fish	*(v) process#1* and *(v) prepare#1* and *(n) fish#2*	*5.2.1.2, Steps in food preparation*

Table 2, select WN and RW classifications of CW entries

This basic classification method, with minimal alterations, was used for both the manual classifications and the later vector classifications of the CW vocabulary. Only Cree nouns and verbs in the CW database were used, all particles and affixes were ignored. In total, this meant that 18 881 out of 21 345 entries were manually classified, comprising 5212 nouns and 13 669 verbs.

2.5 Requirements for Manual Classification

As might be expected for a general purpose dictionary, on the whole, most vocabulary in CW was fairly pedestrian in nature, and could be easily associated with at least one semantic category in WN or RW. For example, a word such as *apiwinis* (‘seat, chair’) can be unproblematically classified in WN as *(n) chair#1* (‘a seat for one person, with a support for the back’) and in RW under the domain of *5.1.1.2 Chair*. Although even relatively simple classifications such as these (particularly the lexeme-to-lexeme WN classifications) may not be entirely accurate (the object described in English by the word ‘chair’, for instance, may possess highly divergent connotational meanings between cultures and language groups), they are nonetheless able to consistently correlate basic, denotational meanings. Although more culturally particular terms, such as *wîsahkêcâhk* (‘Wisahkecahk; Cree culture hero, legendary figure’), did occasionally occur in the dictionary source, in instances such as these, simply using a more general, superordinate WordNet synset such as *(n) hero#5* was generally seen by Cree speakers and experts as appropriate (Wolvengrey, personal correspondence, 2020).

As previously mentioned and demonstrated, multiple manual classifications in either WN or RW were given to any CW entry whose meaning spanned several classification categories. In WN, 11 092 CW entries had more than one manual classification, while in RW, 5336 had more than one classification. The discrepancy between these figures largely reflects the difference in generality between WN and RW; since RW categories are typically much broader than WN synsets, it is less often the case that several are needed to cover the meaning of any given Cree word.

Number of Manual Classifications for any Given CW Entry	Number of Entries (WN Classifications)	Number of Entries (RW Classifications)
7	2	0
6	13	1
5	71	10
4	431	69
3	2442	610
2	8131	4726
1	8180	13 936

Table 3, the distribution of the number of manual classifications necessary to classify entries in CW

The general classificatory mundanity observed in most CW entries served to verify that the principal challenge of manual classification, rather than the difficulty of the task itself, is its scale. The *Cree: Words* dictionary, at the time of its manual classification, consisted of 21 345 entries, of which 18 881 were to be classified. The total process of semantically classifying these entries in both ontologies took roughly 3 months, during which a single manual annotator spent three to four hours per day almost solely on the task of manual semantic classification, yielding

an average of ~167 classifications per hour. In doing this, classifying by WordNet was found to be slightly slower than classifying by Rapid Words (at an average of 143 vs. 192 entries per hour), in addition to requiring a greater degree of prior familiarisation, given its larger and more linguistically complex structure. The noted rate of classification with RW was comparable to that recorded for the manual classification of the *Maskwacis Cree Dictionary* in 2014, in which three annotators (all of whom had linguistic training) required roughly two weeks to semantically classify 8986 entries in the same ontology. In either case, as expected, manually classifying the full contents of a dictionary of this size requires between weeks and months of dedicated labour and at least some degree of metalinguistic competence to accomplish, constituting a commitment which may range from impossible to infeasible for documentation settings in which either time or available resources are limited.

CHAPTER 3. COMPUTATIONAL SEMANTIC CLASSIFICATION

3.1 An Overview of Computational Semantics

Although it may initially seem unusual to assign a task so seemingly rooted in lived-experience, inference, and genuine linguistic understanding as semantic classification to a computer, an unthinking, linguistically incompetent automaton, the combined potential of modern computational processing power and the sheer size of contextual data provided by the internet poses, in actuality, a wide variety of means to partially, or even fully, digitise the process of semantic classification. Part of the reason for this is one which has already been briefly mentioned; in the majority of cases, semantic classification is a straightforward, repetitive, and critically, predictable task which does not so much rely on nuanced cultural understanding as it does a simple awareness of conceptual relatedness. To return again to the example of the English word *chair* and the Cree word *apiwinis*, in the context of semantic classification, one does not need an in-depth understanding of the craftsmanship, cultural significance, or historical role of chairs in English culture, nor of *apiwinis* in Cree culture, nor indeed must one even understand what a chair or *apiwinis* is; rather, one must only be aware that the English word ‘chair’ and the Cree word *apiwinis* refer to the same tangible objects, and that one can reliably expect these words (that is, these arbitrary Saussurean ‘signifiers’) to refer to the same ‘signified’ entities, in order to classify one as the equivalent or nearest equivalent of the other. With the inevitable exception of figurative or culturally salient terms, the task of semantic classification (in the vast majority of instances) thus only truly requires the ability to recognise that two words or lexical units (in the context of our task, a dictionary entry and a WordNet synset or Rapid Words domain) denotatively refer to the same signified object or concept.

One critical linguistic theory here is the Distributional Hypothesis, a theory first conceptualised in the 1950s by figures such as Zellig Harris (1954) and often summarised through John Firth's popular maxim that "a word is characterised by the company it keeps" (Firth 1957). The Distributional Hypothesis states that synonymous or near-synonymous words tend to occupy almost identical contextual environments, and that the degree of semantic difference between such words roughly corresponds to the degree of difference in their average environments. For example, the words 'lawyer' and 'barrister', two almost completely synonymous words, tend on average to occur in almost completely identical contexts (near words such as, for example, 'court', 'judge', or 'defendant'). Meanwhile, 'paralegal', a term which is also semantically related to 'lawyer', but not as closely as 'barrister', occurs in similar, but less overlapping, contexts (near 'court' and 'defendant', but also 'secretary' and 'clerk'). The Distributional Hypothesis can also be applied predictively, stating that different words which occur in similar lexical contexts on average can be presumed to be semantically related, with closer overlap in average context indicating closer semantic relation. For example, even if one does not know the meaning of the word 'ongchoi', the fact that it can be seen to occur in contexts such as "Ongchoi is delicious sauteed with garlic", "Ongchoi is superb over rice", and "...ongchoi leaves with salty sauces…" should be sufficient to suggest to a reader who has previously seen sentences such as "spinach sauteed with garlic over rice", "chard stems and leaves are delicious", and "collard greens and other salty leafy greens" that *ongchoi* is some form of leafy green, similar to spinach, chard, or collard greens (Jurafsky & Martin 2021). In this way, even without knowing the meaning of a word, a semantic profile of it can be constructed entirely based on contextual distribution, with this profile being more accurate the greater number of contexts are factored into consideration.

The Distributional Hypothesis is a fundamental concept in the field of statistical semantics; that is, the idea that word meanings can be derived entirely automatically from the analysis of statistical patterns in their distribution in corpora and the frequency of their co-occurrence with other context words (Weaver 1955). In practice, statistical semantic methods rely on the use of a much greater variety and quantity of contextual distributions across corpora than is typically considered feasible for human annotators to process; for this reason, their practical implementation is near universally through computational means.

One such practical application of the Distributional Hypothesis is vector semantics, a means of representing the average co-occurrence context of any given lexical unit (such as a word) as a set of numerical values (or, dimensions), with each value representing some abstract aspect of the word's average context, and then using these values (collectively referred to as an embedding) to define a vector in some multidimensional space. The vectors of different words can thus be compared in this space by means such as cosine or Euclidean distance to determine how similar the embeddings (and thus, the average occurrence contexts) of the two words are, and thus by the Distributional Hypothesis, how similar the two words are in meaning. The internal mechanics of vector semantics are discussed in much more detail in Jurafsky & Martin (2021); for the purposes of this thesis, we used a pre-built, off-the-shelf vector semantic program (word2vec) with minimal alteration. The exact internal workings of word2vec are described in Mikolov et al. (2013), however, in brief, it is a neural network model which draws lexical contexts from the multi-billion word English Google News Corpus and generates embeddings composed of up to several hundred dimensions (in our investigation, three hundred) for words by taking the content

of their contexts as a bag-of-words, ignoring word order and function words, and weighing nearby context words more heavily than more distant ones. As a result of it generating vectors by taking a word's context purely as a bag-of-words and ignoring syntax, word2vec cannot effectively model polysemy or homography, unlike more recent, sentence-based vector generation models such as BERT (Devlin et al. 2019); however, word2vec remains a pragmatic, off-the-shelf tool with minimal requirements for pre-training, and limited forays into its use on Cree vocabulary have already been made (Harrigan & Arppe 2021).

As a final point, it should be noted that vector semantics is not the only possible means of computationally generating semantic classifications for lexical items. For example, if one uses an ontology such as RW, one can categorically match all Cree vocabulary with however many RW domains contain some degree of shared lexical content with the definition for that Cree entry, and then manually remove any false positives, providing a set of all domains for which the given Cree word would be a plausible member. However, for the purposes of this investigation, we have chosen instead to focus our efforts on vector semantic classification, due in no small part to the infeasibility of such a matching method for use with larger ontologies such as WN, which may easily produce hundreds of false positives for any given CW entry.

3.2 Generating Vectors

One (apparently) immediate hurdle in the application of vector semantics to Cree vocabulary would be the scarcity of Cree-language corpora. As mentioned, statistical semantic methods such as vector semantics are eminently reliant on large corpora in order to have a sufficient variety of distributional contexts for representative word vectors for most vocabulary to be generated.

Currently available Cree corpora, although relatively large by Canadian Indigenous standards, are miniscule when compared to those of majority languages such as English or French; the largest current corpus of Cree literature, a morphosyntactically tagged compilation of nine existing Cree texts, consists of 152 405 word tokens of 34 115 types (Schmirler forthcoming; Arppe et al. 2022), compared to the 1 billion tokens of the Corpus of Contemporary American English (Davies 2008), or the 14 billion token iWeb Corpus (Davies 2018). Correspondingly, attempts to generate Cree word vectors using co-occurrence data from purely Cree sources have been largely unsuccessful (Harrigan & Arppe 2021).

However, the formatting of CW, a bilingual dictionary with glosses in English, enabled us to sidestep this lack of corpus data. Given that the definitions for each entry in the dictionary are, by nature, meant to convey as close of a meaning as possible to the Cree word which they describe, rather than generating vectors based off of the Cree words themselves, one can instead generate vectors based on the English words in the definitions, averaged out as a bag-of-words using word2vec (Harrigan & Arppe 2021). In this way, one is able to leverage the enormous size of existing English corpora and years of advancements in English vector semantic technology, in essence transforming each Cree headword into a simple label for a group of English words describing it:

CW Entry	PoS	English Gloss	Words Included in the Embedding
maskoskâw	VII-1v	there are many bears around	many + bears + around
maskowiyâs	NI-1	bear meat	bear + meat
maskosimow	VAI-1	s/he dances the bear dance	dances + bear + dance

Table 4, demonstration of the lexical content in CW entries used to construct entry vectors with word2vec.

Although pragmatically useful (in that it enables vector semantics to be performed on the data in the first place), this method is not without disadvantage. Not only does it assume that each Cree word's complete meaning is (or even can be) communicated fully through a dictionary gloss in a foreign language that is rarely longer than a sentence, it also operates on the assumption that the individual English words used as translations carry identical meanings to their Cree counterparts, both having no additional shades of meaning not communicated through the Cree word and having all additional senses, meanings, and connotations associated with that word. On account of both of these factors, our 'Cree' word vectors are, in reality, English word vectors which describe Cree words. Although, theoretically, the use of endemically Cree word vectors may have provided more 'genuine' statistical representations of the CW content, this method of simply generating word vectors based on majority language glosses does have the advantage of being broadly applicable cross-linguistically; as long as the entry definitions are given in English (or another language with existent vector semantic resources), this method can theoretically be applied to bilingual dictionaries with headwords in any language. Additionally, the use of vectors from Cree corpora would complicate the process of comparing Cree vocabulary with English-based semantic classification ontologies such as RW or WN using vector semantics, as it would require the comparison of usage contexts for Cree words in Cree text with those of

English words in English text; given the fundamental typological differences between the languages, even with large Cree corpora, finding common contexts between the two would be problematic. Thus, as with our choice to use WN and RW as classification schemes, our vector generation method is, if not as strictly adherent to target language-internal semantics as possible, designed with pragmatism and ease of application as foremost concerns.

To compare with the CW vectors, vectors were also generated for every synset in WordNet and every domain in Rapid Words, with all synset or domain internal content (minus function words) being averaged as a bag of words:

WN Synset Head	PoS	WN Synset in Database	Material used for Embedding
(n) bear#1	Noun	02131653 05 n 01 **bear** 0 009 @ 02075296 n 0000 #m 02131418 n 0000 ~ 01322983 n 0000 ~ 02132136 n 0000 ~ 02132320 n 0000 ~ 02133161 n 0000 ~ 02133704 n 0000 ~ 02134084 n 0000 ~ 02134418 n 0000 | **massive plantigrade carnivorous ~~or~~ omnivorous mammals with long shaggy coats ~~and~~ strong claws**	bear + massive + plantigrade + carnivorous + omnivorous + mammals + long + shaggy + coats + claws
(v) forage#2	Verb	01179996 34 v 01 **forage** 0 003 @ 01182162 v 0000 + 07817067 n 0102 ~ 01206120 v 0000 02 + 01 00 + 04 00 | **wander ~~and~~ feed; "~~The~~ animals forage ~~in the~~ woods"**	forage + wander + feed + animals + forage + woods
(adj) predatory#2	Adjective	00084491 00 s 06 **predatory** 0 **rapacious** 0 **raptorial** 0 **ravening** 0 **vulturine** 0 **vulturous** 0 003 & 00082711 a	predatory + rapacious + raptorial + ravening + vulturine + vulturous + living + preying + other + animals +

		0000 + 01618959 n 0601 + 01606971 n 0302 | **living ~~by~~ preying ~~on~~ other animals especially ~~by~~ catching living prey; "~~a~~ predatory bird"; "~~the~~ rapacious wolf"; "raptorial birds"; "ravening wolves"; "~~a~~ vulturine taste ~~for~~ offal"**	especially + catching + living + prey + predatory + bird+ rapacious + wolf + raptorial + birds + ravening + wolves + vulturine + taste + offal

Table 5, demonstration of the lexical content in WN synsets used to construct synset vectors with word2vec.

For Rapid Words domains, the domain description, all elicitation questions, and all example answers were used as vector generation material.

RW Domain Code	Full Domain Content	Material used for Embedding
1.6.1.1, Mammal	1.6.1.1 **Mammal** **Use ~~this~~ domain ~~for~~ general words referring ~~to~~ mammals (phylum Chordata, class Mammalia).** **~~What~~ general words refer ~~to~~ mammals?** ***mammal, mammalian, animal***	Mammal + use + domain + general + words + referring + mammals + phylum + Chordata + class + Mammalia + general + words + refer + mammals + mammal + mammalian + animal

Table 6, demonstration of the lexical content in RW domains used to construct synset vectors with word2vec

3.3 Applying Vectors for Classification

Once these semantic vectors were created, the only remaining step was to compare the vectors of each WordNet synset and each Rapid Words domain with each entry in *Cree: Words*. This comparison was made by finding the cosine of the angle of any two given vectors, this being

widely-considered a “standard way” to compute semantic similarity in the context of vector semantics (Jurafsky & Martin 2021). The closer the cosine value of the angle of any given two vectors is to 1, the more closely associated these two vectors are in multidimensional vector space, and thus the more often the words which they represent appear in overlapping distributions, indicating according to the Distributional Hypothesis that these represented words are more closely semantically related. Thus, for every CW entry, a list of all WN synsets and all RW domains ranked in order of similarity to the given entry was generated, with the ‘accuracy’ of vector classifications for any given CW entry for our purposes being measured by the position of the synset or domain used as the manual classification on the ranked list of vector classifications; the higher the rank of the manual classification, the more ‘accurate’ the vector classifications for that entry. For example, for the Cree word *tawikaham* (‘s/he slashes, s/he slashes s.t., s/he chops s.t.’), which has the manual classifications *(v) slash#1* and *(v) chop#4* in WN, all 155 327 WN synset vectors are compared against the vector of the Cree word and ranked in order of similarity, generating a list such as this:

1. *(v) slash#3*: (Cosine Similarity 0.68576694),
2. *(v) gash#1*: (Cosine Similarity 0.68576694),
3. *(n) slasher#1*: (Cosine Similarity 0.63708573),
4. ***(v) slash#1*: (Cosine Similarity 0.58829528),**
5. *(v) cut_down#2*: (Cosine Similarity 0.58829528),
6. *(v) slash#4*: (Cosine Similarity 0.57162038),
7. *(v) cut#24*: (Cosine Similarity 0.56211452),
8. *(v) slit#1*: (Cosine Similarity 0.54561763),

9. *(v) slice#1*: (Cosine Similarity 0.54561763),
10. *(adj) knifelike#1*: (Cosine Similarity 0.54486923),

…

On this list, the manual classification in the highest rank, *(v) slash#1*, occurs in the 4th position, making it the 4th 'most similar' synset in WN to the Cree word *tawikaham* according to word2vec. For RW, in which the manual classification for *tawikaham* is *7.8.3 Cut*, the manual classification is ranked 3rd among the vector classifications; thus, at least in absolute terms, the RW vector classification for *tawikaham* is more accurate than its counterpart in WN:

1. *8.1.4.3 Decrease*: (Cosine Similarity 0.44402625),
2. *5.4.3.5 Cut hair*: (Cosine Similarity 0.44291632),
3. ***7.8.3 Cut*: (Cosine Similarity 0.43714895),**
4. *6.7.1 Cutting tool*: (Cosine Similarity 0.41539036),
5. *5.4.7 Care for the fingernails*: (Cosine Similarity 0.41230196),
6. *9.3.1.4 To a smaller degree*: (Cosine Similarity 0.41123268),
7. *2.6.3.4 Labor and birth pains*: (Cosine Similarity 0.40292210),
8. *3.5.6.5 Cry, tear*: (Cosine Similarity 0.40179453),
9. *6.2.4.4 Trim plants*: (Cosine Similarity 0.39176951),
10. *7.2.2.5.1 Fall*: (Cosine Similarity 0.39012915),

…

Comparing all 155 327 WordNet synsets to all 18 881 noun and verb entries in the *Cree: Words* dictionary using cosine took roughly five days, while comparing all 1789 Rapid Words domains to all of these entries took less than 16 hours using a mid-range, 2-core laptop with 8GB of RAM. This process of comparison, however, may be described as ‘embarrassingly parallelisable’, being easily separated into a large number of parallel tasks with little to no codependency. As such, more powerful computers can perform this task in temporal ease; for example, when run on Compute Canada’s Cedar high-performance computing cluster using 64 cores in parallel, each with 4-8 GB of RAM, comparing all 155 327 WordNet synsets to all 18 881 CW entries took only 90 minutes.

CHAPTER 4. INITIAL VECTOR CLASSIFICATION RESULTS

4.1 Comparison of WN and RW Vector Classifications

	Verbs, RW top	Verbs, RW median	Nouns RW top	Nouns RW median	Verbs, WN top	Verbs, WN median	Nouns, WN top	Nouns, WN median
0%	1	1	1	1	1	1	1	1
10%	1	1	1	1	5	11	1	2
20%	1	2.5	1	1	18	51.7	2	4
30%	3	6.5	1	1	51.6	166.3	4	8
40%	7	18	1	2	136.8	448.8	7	16.1
50%	**18**	**42**	**2**	**3.5**	**333**	**1045**	**15**	**30.5**
60%	42	80	4	7	762.2	2057.3	28	60
70%	85	140	10	16	1633.8	4096.4	59	139
80%	167	236	25	34	3553.8	8036.9	164	375.4
90%	335	373	75.3	108.3	9553.8	17488.6	864	1670.4
100%	983	983	915	915	137352	137352	121883	121883

Table 7, the vector assigned ranks of manual WN and RW classifications in percentiles, for both the top-ranked manual classification and the median of all manual classifications if there were several. For example, row 5 column 8 indicates that 30% of the time, the highest ranked manual classification among the vector classifications was within the top 4 for CW nouns using WN. The row of medians for each ontology-PoS combination is bolded. (Dacanay et al. 2021b).

If the intended ideal of vector semantic classification is to be considered as the exact replication of semantic classifications made by human beings, the results of our initial vector classifications with both WordNet and Rapid Words may be said to have been mixed successes, with a strong divide in accuracy between Cree parts-of-speech. The median position of the top manual classification for Cree nouns among the vector classifications was 15 in WordNet and 2 in Rapid Words, while the median position of the top manual classification for Cree verbs was 333 in WordNet and 18 in Rapid Words. Instances in which the top-ranking vector classification was an

exact match for the manual classification, our theoretical ‘best-case’ scenario, were relatively rare, occurring only 315 times in WordNet for the 13 669 Cree verbs (2.3% of total cases), and 726 times in WordNet for the 5212 Cree nouns (13.9%). For Rapid Words, the top vector classification was an exact match for the manual classification for 2345 entries (45%) for Cree nouns and for 2733 entries (20%) for Cree verbs. The manual classification appeared within the top 5 vector classifications ~1720 times (33%) for nouns in WN, ~1337 times (10%) for verbs in WN, ~3231 times (62%) for nouns in RW, and ~4679 times (35%) for verbs in RW. The manual classification appeared in the top 10 vector classifications ~2293 times (44%) for nouns in WN, ~1913 times (14%) for verbs in WN, ~3648 times (70%) for nouns in RW, and ~5877 times (43%) for verbs in RW. Although exact matches were uncommon (particularly among verbs), for most CW entries, the top-ranking sets of vector classifications were, if not precisely manual-like, at least semantically relevant to the given CW entry. For example, for a Cree word such as *mostos* (‘cow, cattle, buffalo’), even though the manual classification (*(n) cattle#1*) is only ranked 53rd among the vector classifications, the top vector classification is *(n) cow_pasture#1*, a domain-relevant, if not manual-like, synset, and the top ten classifications are all related to cattle-rearing and bovines. This phenomenon is discussed in more detail in Section 4.2.1.

Broadly speaking, across both ontologies, Cree noun classifications were more accurate (that is, the median position of the top manual classification was ranked higher among the vector classifications on average) than those of Cree verbs. This discrepancy is likely the product of lexicalisation pattern differences between Cree and English, rather than a shortcoming of vector semantic classification as a whole. As mentioned, Cree is a highly polysynthetic language which makes use of verbal affixes and stative verbs to express meanings which, in English, would be

expressed through adjectives or adverbs; in fact, Cree lacks adjectives and adverbs as a distinct lexical class altogether (Wolfart 1973). Verbs make up a substantial majority of the Cree lexicon (as much as 79% in existing corpora (Harrigan et al. 2017)), with verbs being used to express a variety of meanings which are either inexpressible through a single English word or express a concept which, in English, would not be lexicalised verbally. For example, although the best possible WordNet classification the Cree word *sîhkaciw* (‘s/he is very thin, s/he is lean’) might be a synset such as *(adj) lean#1*, the precise meaning of *sîhkaciw* is not accurately expressed through this classification, being that it is of a different part-of-speech entirely. Although one could add *(v) be#1* as a second classification to remedy this, such an addition would, to a degree, reduce the overall accuracy of the manual classification in that, while *sîhkaciw* and *(adj) lean#1* are closely semantically related, *sîhkaciw* and *(v) be#1* are only tangentially related in isolation. This issue is partially resolved through using Rapid Words domains, which are broader and not tied to specific lexical items or parts-of-speech, but the vector embeddings for these domains are still generated from English words, which are subject to the same discrepancies in precise meaning as previously mentioned. Additionally, some Cree verbs simply have no lexicalised equivalent in English; for example, for *kwêskahêw* (‘s/he changes s.o.'s position in lying or sitting’), it is unclear how any single English word could possibly serve as a ‘best’ classification, with the meaning only being expressible by dividing it more or less equally among several synsets (e.g. *(v) lie#2*, *(v) sit#1*, and *(v) reposition#1*), none of which are entirely accurate classifications in isolation.

By comparison, Cree noun derivation and inflection are much more restricted, and Cree and English nouns fill largely the same syntactic and semantic niches; as such, nominal lexicalisation

patterns are fairly similar between the two languages, with concepts lexicalised as nouns in Cree thus being more likely to have a near exactly matching lexicalisation in English. Thus, we believe the part-of-speech mediated accuracy discrepancy seen in vector classifications with both ontologies to be a cross-linguistic problem concerning English and Cree, rather than a fault of the vector method; vector semantics cannot choose a reasonable 'best' classification if no such individual classification exists.

One consistent difference in accuracy which does appear to be caused by aspects of vector classification as a method is that between the two ontologies, Rapid Words and WordNet. Superficially, across both Cree parts-of-speech, Rapid Words outperforms WordNet in absolute terms, with manual classifications having a median vector position of 2 and 18 for Cree nouns and verbs respectively, compared to 15 and 333 for WordNet. However, WN vector classifications are still more accurate than RW relative to the total number of possible classifications (for example, in WN the manual classifications have a median position in the top 0.0096% (15th out of 155 327) for Cree nouns and 0.214% (333rd out of 155 327) for verbs compared to the top 0.112% (2nd out of 1789) and 1.00% (18th out of 1789) for Cree nouns and verbs respectively in RW). Despite this, in absolute terms, RW classifications still appear, by our metric, reliably more accurate or 'human-like'; that is, manual classifications appear on average at a higher rank among the vector classifications.

4.2 Initial Problems with WN Vector Classifications

In order to understand the potential reasons behind the seemingly superior performance of RW compared to WN in vector classifications, one must first be familiar with some of the most

pervasive issues encountered in those vector classifications. These chief issues may be divided into three basic categories, being the issue of excessive degrees of specificity, the issue of irrelevant, 'regift words' (see Section 4.2.2) appearing in top-ranking vector classifications, and the issue of semantically irrelevant synset content and WN-internal 'distractors'. Finally, the issue of erroneously-represented English polysemy among Cree vector embeddings, a problem equally prevalent among WN and RW vector classifications, must also be discussed.

4.2.1 Specificity

Perhaps the most frequently noted and systematic of four aforementioned issues is that of specificity, or more broadly of lexical precision. The vector method, particularly when used with WordNet, has consistently struggled to produce top-ranking classifications which match the specific English synsets used as the manual classifications, even when these high-ranking classifications are within the same basic semantic domain. For example, for the Cree word *apihkêsîs* ('spider'), even though the top WordNet classifications consist entirely of types of spider, spider products, and (as a function of word2vec's inability to reliably model polysemy) spider monkeys, the exact manual classification for this entry, (*(n) spider#1*) is only ranked 185th:

1. *(n) spider_web#1* (Cosine Similarity: 0.79724957),
2. *(n) spider's_web#1* (Cosine Similarity: 0.79724957),
3. *(n) spider_web#2* (Cosine Similarity: 0.79230372),
4. *(n) spider's_web#2* (Cosine Similarity: 0.79230372),
5. *(n) genus_Ateles#1* (Cosine Similarity: 0.76677546),

6. *(n) orb_web#1* (Cosine Similarity: 0.75656461),
7. *(n) family_Majidae#1* (Cosine Similarity: 0.75422829),
8. *(n) funnel_web#1* (Cosine Similarity: 0.67104201),
9. *(n) spider_monkey#1* (Cosine Similarity: 0.66930821),
10. *(n) Ateles_geoffroyi#1* (Cosine Similarity: 0.66930821),

Similarly, for the Cree verb *kîmwêw* ('s/he whispers'), the top manual classification (*(v) whisper#1*) is only ranked 199th, despite the top 10 classifications all, in one way or another, relating to either actions or people associated with quiet, concealed speech:

1. *(v) murmur#1* (0.69050016),
2. *(adj) voiceless#3* (0.64343067),
3. *(adj) breathed#1* (0.64343067),
4. *(n) yenta#2* (0.62532448),
5. *(n) cat#3* (0.61659776),
6. *(adv) girlishly#1* (0.61572368),
7. *(n) sweet_nothings#1* (0.60610735),
8. *(n) honeyed_words#1* (0.60610735),
9. *(v) mutter#2* (0.60246510),
10. *(v) murmur#2* (0.60246510),

This general trend of top-ranking vector classifications being highly semantically related, but overly specific and thus not precise lexical matches, is not unique to our study; for example,

when using similar methods to classify vocabulary in Choctaw, Brixley et al. (2020) found that, when seeking vector semantic classifications for nominal and adjectival forms of the word ‘female’, the top-ranking returns were all specific female names.

To determine the extent of this problem, one may assess the specificity of any given vector classification relative to its manual classification (if both share a common ancestor synset on the hypernymy hierarchy) by measuring how many levels farther down on the hierarchy both classifications are compared to this common ancestor; if the vector classification is lower on the hierarchy, it is ‘more hyponymic’ and thus can be judged ‘more specific’ relative to the manual classification. With this as criteria for specificity, in cases in which both the top vector and manual classification shared a common ancestor synset (which constituted 60.7% of cases for nouns and only 15.5% of cases for verbs), the prevalence of overspecificity among vector classifications is, although statistically significant, relatively limited. For CW entries with noun classifications, the top vector classification was more specific than the manual classification 43% of the time, compared to the manual classification being more specific than the top vector classification 26.1% of the time, the two classifications being on the same relative level of specificity 18.5% of the time, and the two being exact matches 9.8% of the time. For CW entries with verb classifications, the top vector classification was more specific 34.3% of the time, the manual classification was more specific 14.5% of the time, the two were on the same relative level of specificity 28.2% of the time, and the two were exact matches 24% of the time.

Hypernym Level Difference of Vector Classifications Relative to Manual Classification	CW Entries with WN Noun Classifications	%	CW Entries with WN Verb Classifications	%
-6 … -N	53	0.9	1	0.1
-5	70	1.2	3	0.2
-4	138	2.4	8	0.5
-3	239	4.1	22	1.4
-2	424	7.3	43	2.7
-1	595	10.2	154	9.6
+0	1080	18.5	451	28.2
=0	571	9.8	384	24
+1	1139	19.5	360	22.5
+2	612	10.5	125	7.8
+3	364	6.2	33	2.1
+4	228	3.9	15	0.9
+5	112	1.9	2	0.1
+6 … -N	60	1.0	0	0

Table 8, positions of vector classifications on the hypernymy hierarchy relative to manual classifications for CW entries whose manual classifications are WN nouns or verbs. For example, row 10 (+2) indicates that the vector classification is two levels farther down in the hypernym hierarchy (and thus, is two ‘levels’ more specific) than the manual classification 612 times for CW entries with noun classifications and 125 times for CW entries with verb classifications. Row 7 (+0) indicates the number of instances in which the manual and top vector classifications are on the same relative level of the hypernym hierarchy, but are not matches, while row 8 (=0) indicates instances where the vector and manual classifications were exact matches.

Thus, though vector overspecificity is more prevalent than vector underspecificity across parts of speech, it is also frequently the case that top vector classifications were merely semantically imprecise, scattershot categories from within the correct level of the hypernymy hierarchy and a relevant general semantic region. For both nouns and verbs, although vector overspecificity was the most common individual scenario, the majority of top vector classifications (at least, when both vector and manual classifications shared a common ancestor) were either on the same level as, were exact matches with, or were less specific than, their corresponding manual classification (57% of the time for nouns and 65.7% of the time for verbs). Thus, although overspecificity compared to manual classifications is attested and statistically significant, the issue appears more broadly to constitute a general trend of semantic imprecision among high ranking vector classifications.

This imprecision seems likely to be attributable to a fundamental methodological difference between human and statistical semantic interpretation, in that while a human being mentally represents a semantic concept, such as, for example, ‘duck’, as a fuzzy and variable region, consisting of a central prototype (‘medium-sized freshwater bird’) surrounded by relevant exemplars (‘mallard’, ‘canvasback’, etc.) (Taylor 2008), a statistical semantic model such as vector semantics interprets ‘duck’ as a precisely defined vector interacting with other vectors at exact points in multidimensional space, without the capacity for human semantic ‘fuzziness’ by default. As such, while a human might think a generic, umbrella classification such as *(n) duck#1* would be best for the Cree word for duck (*sîsîp*), a vector model would instead view whatever specific WordNet vector happens to be the closest to that of the bag-of-words embedding based on the CW definition for *sîsîp*, which may be any word with a closely associated distributional

relationship to ducks (in our case, the top vector classification for *sîsîp* was *(n) duck_hunter#1*), rather than necessarily the broader term which best describes them as a semantic category. Considering this, it is unsurprising that category or synset-level vector classifications based on fuzzily-defined mental semantic classes differ in specificity from those based on precise statistical correspondences.

As mentioned, this vector overspecificity was much more prevalent among WordNet vector classifications than among those with Rapid Words. The principal reason for this appears to be the fact that Rapid Words is a much smaller ontology (at 1789 domains rather than WordNet's 155 327 synsets), and thus that there are simply fewer highly specific classification categories to choose from, limiting the selection available to the vector model of relevant 'wrong' choices. As such, if the vector model is able to identify the basic semantic region of a Cree word (which, as discussed, it typically was), the number of choices for classifications within that semantic region is not only much smaller, but each individual choice is also necessarily more general, reducing the possibility of over-specific or over-precise classifications crowding the top vector ranks in RW.

4.2.2 'Regift' Words and Proper Nouns

A second phenomenon of note among our initial vector classifications was that of a small number of low-frequency English WordNet synsets appearing in the top-ranking classifications of a disproportionate number of Cree words. We nicknamed this phenomenon the 'regift' problem after an emblematic example, namely, *(v) regift#1*, an English verb which occurs only 16 times in the 1.9 billion word Corpus of Global Web-Based English (or GLoWbE) (Davies

2013), which occurs within the top 1000 vector classifications for Cree verbs 7324 times, placing it within the top 1% of (vector-based) relatedness for over 65% of all of *Cree: Words*' verbs. A number of other extremely low-frequency English WordNet synsets of this nature appear with disproportionate frequency in the high-ranking classifications of Cree entries, including *(n) dingbat#1* (149 occurrences in GLoWbE compared 8082 occurrences (71%) in the top 1000 vector classifications of a set of 11 236 Cree verbs) and the most common 'regift' synset, *(n) Rumpelstiltskin#1* (with 93 occurrences in GLoWbE compared to 8094 occurrences (72%) among Cree verb classifications). Overall, this phenomenon was much more common among verb classifications than it was among nouns; for comparison, *(v) regift#1* only appears within the top 1000 classifications of 183 (3.5%) out of 5212 Cree nouns, and *(n) dingbat#1* and *(n) Rumpelstiltskin#1* only occur 356 times (6.8%) and 299 times (5.7%) respectively. The most prevalent 'regift' synset among Cree nouns (*(n) smock#1*) only occurred 564 times (10.8%).

In addition to these WordNet synsets, there were also some 'regift'-like domains in the Rapid Words classifications; for example, subdomains to the domain *4.1.9 Kinship* (such as *4.1.9.1 Related by birth*, *4.1.9.2 Related by marriage*, etc.) occurred an average of 12 times in the top 1000 vector classifications of each Cree verb and noun, and occurred within the top 10 RW vector classifications 35.7% and 33.9% of the time Cree verbs and nouns respectively. However, the 'regift' problem as a whole was much less prevalent among the Rapid Words classifications than among those with WordNet.

Invariably of little to no semantic relevance to the Cree word(s) in whose classifications they are found, these 'regift' words do not appear to follow any overt semantic pattern; the only apparent

common thread between them is their low-frequency in corpora. It is precisely this low frequency which we suspect to be responsible for their unusual behaviour; since vector embeddings are created based on contexts in corpora, the fewer times that an individual word appears in corpora, the more disproportionately its vector is affected by individually unusual usage contexts. With ‘regift’ for example, which occurs only 16 times in the GLoWbE corpus, if even a single text was to use ‘regift’ in an uncharacteristic fashion, it would be enough to significantly impact the average context of the word in the corpus as a whole, and thus enough to skew the vector embedding. This may be another reason why ‘regift’ categories are less prevalent in Rapid Words; given that Rapid Words domains necessarily feature fewer highly infrequent words (on account of its intended use as an elicitation tool for basic vocabulary), and the fact that our Rapid Words domain vectors were defined based on the lexical contents of the entire domain, providing more content on average with which to create each vector, the potential impact of individual words with unusual vector representations was, compared to WN, greatly reduced.

One final factor of note regarding these ‘regift’ classifications is the disproportionate selection of (semantically irrelevant) proper nouns as WordNet vector classifications for both Cree nouns and verbs. In addition to the aforementioned *(n) Rumpelstiltskin#1*, other uncharacteristically common proper nouns include *(n) Godiva#1* (5775 among verbs), *(n) Ariadne#1* (6391 among verbs), and *(n) Brunnhilde#1* (3079 among verbs). Although these proper nouns only make up a minority of the total ‘regift’ synsets present in WordNet, they are uniquely positioned among such synsets in that, as a result of their being almost wholly irrelevant, they can, in theory, be categorically removed as a set. As mentioned, a significant portion of WordNet’s synsets (8221

or ~7% of WN nouns) are of a more encyclopaedic than lexicographic nature, listing historical figures, myths, wars, toponyms, and so on. The exact nature of these proper nouns is reflective of the cultural context in which the Princeton WordNet came into being, namely, the Northeastern United States of the late 20th century (Lindén & Carlson 2010); for example, there is a synset for Portsmouth, New Hampshire (a town of ~21 000), but none for Chengdu, Sichuan (a major Chinese city of ~21 million inhabitants). In addition to their skewed focus, these proper noun synsets are largely irrelevant to the type of first pass semantic classification which vector semantics provides, and with very few exceptions, almost no entries in the *Cree: Words* dictionary (or, in all likelihood, in most bilingual dictionaries of endangered languages outside of the United States) would correspond to any proper nouns present in WordNet in the first place. For example, although CW does indeed contain 196 Cree proper nouns, most of these refer to historical figures, mythic characters, or place names which have no direct correspondence in WordNet, but which may reasonably be represented by a more generic synset:

CW Entry	Gloss	WN Manual Class	RW Manual Class
acâhkosa kâ-otakohpit	Starblanket; literally: "One who has Stars as a Blanket"; name of Cree chief, signatory to Treaty 4	*(n) Indian chief#1*	*9.7.1.1 Personal Names* and *3.5.4.5 History*
kâ-ohpawakâstahk	Flying Dust, SK; Cree reserve; literally: "Where the Dust Flies Up"	*(n) Indian reservation#1*	*9.7.2.3 Names of Cities*
paskwâwiýinînâhk	in Plains Cree country	*(n) prairie#1*	*9.7.2.2 Names of Regions*
pimicâskwêyâsihk	Lloydminster, SK	*(n) city#1*	*9.7.2.3 Names of Cities*

Table 9, example manual classifications of various Cree proper nouns, all of which can be easily represented by more general, 'class' categories, such as *(n) city#1* or *(n) Indian reservation#1*.

Given that these specific, proper nouns can all be represented in a reasonably grounded fashion with more generic terms, which would in any case lead to them being more easily incorporated into semantic domains with their more general Cree counterparts, there appears little compelling reason to include proper nouns at all within WordNet when using it as a vector semantic classification ontology; removing them not only does not affect the pool of reasonable, human-like classifications available to the model for the vast majority of entries, but also reduces the prevalence of 'regift' classifications. Although WordNet's internal database formatting does not explicitly mark proper nouns in general, it does mark proper nouns which are specific instances of a class (e.g. 'Rosa Parks' being an instance of the class 'woman' or 'Berlin' being an instance of the class 'city') rather than classes themselves with the tag @i (Miller & Hristrea 2006). Thus, all such proper nouns (which would include *(n) Rumpelstiltskin#1*, *(n) Ariadne#1*, and others) could be systematically removed with a simple find-and-delete on WordNet's noun

database file. Concerning the other regift synsets which are not proper nouns (such as, for example, *(v) regift#1* itself), one could consider removing these by requiring some semi-arbitrary minimum corpus frequency for WN synsets to be used as vector classification categories; however, this method may prove disproportionately exclusionary to synsets of certain semantic regions which, although infrequently referenced in most English corpora, may be of great relevance to Cree speakers, such as, for instance, leatherworking or archery (traditional practices of cultural significance in Cree society).

4.2.3 Semantically Irrelevant Synset Content

One final factor negatively impacting the quality of WordNet vector classifications appears to be the internal lexical makeup of WordNet's synsets. As has been demonstrated, a 'full' WordNet synset consists of all of the synset members, a definition for those members, and one or several example sentences. The latter of these, the example sentence(s), is of particular note, as these sentences often contain semantically irrelevant material which, although perhaps more representative of genuine, non-dictionary lexical contexts, nonetheless serve to provide distracting elements to the vector model when creating embeddings using a synset's content. For example, for the synset *(v) drive#2*, the internal composition is as follows:

(v) drive#2, *(v) motor#1* (travel or be transported in a vehicle) "We drove to the university every morning"; "They motored to London for the theater"

Out of the 12 content words present in this synset's full gloss, four of them ('university', 'morning', 'London', and 'theater') are largely irrelevant to the meaning of 'drive' beyond their circumstantial relation in the context of the two example sentences, essentially making a third of

the source material for the vector of *(v) drive#2* only tangentially relevant to the desired meaning of the synset, reducing the vector's overall accuracy for the purposes of semantic classification.

The problem of lexical 'distractors' is of less note in Rapid Words, as domains consist only of semantically relevant vocabulary and formulaic questions concerning this vocabulary; as such, although some terms may be more peripherally related to the core domain meaning than others, the presence of broadly unrelated vocabulary within domains (such as is seen in WordNet synset example sentences) is much less pervasive.

5.2.3 Types of Food
Use this domain for words related to types of food.

What words refer to major types of food?
cereal, staple, meat, fruit, vegetable, salad, raw food, cooked food, condiment, main dish, side dish

What words describe whether something can be eaten or not?

edible, inedible

What words describe food?

tender, tough, crisp, crispy, raw, stale, fresh

Figure 3, a demonstration of 'distractors' in a Rapid Words domain; although terms such as 'staple', 'tender', 'tough', and 'fresh' are not necessarily paradigmatic of the central meaning of the domain 'Types of food', they are still domain relevant, unlike many distractors in WordNet synsets

4.2.4 Polysemy

One final consistent inaccuracy of note among vector classifications in both WN and RW concerns polysemy. By principle, word2vec as a vector generation tool is unable to model

polysemy on account of the fact that it treats all context words as a bag-of-words and assigns one single vector to represent all senses of any given written word type. As such, by using English definition words to generate Cree vectors, top-ranking vector classifications for CW entries are occasionally entirely semantically irrelevant on account of their vector representing an alternate sense of the English word(s) in the definition with little or no semantic relation to the target Cree word. For example, for the Cree word *pîminikanis* (‘gimlet’), which refers to the handheld tool used for boring holes, all of the top ten WN vector classifications referred to beverages due to the English word ‘gimlet’ having an additional sense referring to a type of drink, which is the more frequently used sense in corpora. As such, the vector classifications for this entry reflect polysemy present in the English definition, but not in the Cree word, which only refers to the type of tool:

1. *(n) gimlet#1* (0.73154529),
2. *(n) martini#1* (0.69754860),
3. *(n) manhattan#2* (0.69237388),
4. *(n) gin_and_it#1* (0.68206313),
5. *(n) ratafia_biscuit#1* (0.64722153),
6. *(n) ratafia#2* (0.64722153),
7. *(n) Drambuie#1* (0.64671911),
8. *(n) pink_lady#1* (0.64549165),
9. *(n) planter's_punch#1* (0.64489724),
10. *(n) claret_cup#1* (0.64333994),

The top ten RW vector classifications for *pîminikanis* behaved similarly, strongly favouring domains relating to food and flavour:

1. *2.3.3 Taste* (0.58532236),
2. *5.2.2.8 Eating utensil* (0.57606717),
3. *5.2.3.1.2 Food from fruit* (0.56078271),
4. *5.2.3.4 Prepared food* (0.56060140),
5. *8.3.3.3.4 Colors the spectrum* (0.56014050),
6. *5.2.1.5 Serve food* (0.55369277),
7. *5.2.3.3.3 Spice* (0.54757216),
8. *1.5.3 Grass herb vine* (0.53842635),
9. *1.5.1 Tree* (0.52774830),
10. *5.2.3.7 Alcoholic beverage* (0.52507053),

This class of error is a direct consequence of using English definition words as the basis for the Cree entry vectors instead of *in-situ* Cree vocabulary in corpora, coupled with word2vec's inability to model polysemy. However, even if a vector model such as BERT, which can model polysemy by assigning unique vectors to individual sentential instances of words based on their specific usage context, was to be used, the nature (particularly, the length) of CW's definitions is such that confidently disambiguating the meaning of an English definition without cross-referencing it with the derivational makeup of the Cree word itself is often impossible. For example, for *pîminikanis*, the fact that the definition is only a single word, 'gimlet', with no corroborating context to indicate wordsense, would leave any vector method with only the option

to select whichever sense of the English word ‘gimlet’ is most statistically likely from corpus usage, which in this instance, was the incorrect interpretation of the word as a beverage. It is for precisely this reason that this particular error is equally common across WN and RW vector classifications; it is not ontologically motivated, rather being a consequence of our vector method’s reliance on English lexemes to represent Cree head words. As such, the only thorough solution to this problem would be the use of Cree corpora to generate embeddings; however, as has been outlined, this is impossible with current corpus sizes, and would impede vector-based comparison to English WN and RW semantic classes.

CHAPTER 5. REFINEMENTS TO VECTOR CLASSIFICATION

5.1 Outline of Various Potential Improvements

As has been outlined, vector semantics as a methodology is capable of producing broadly relevant first-pass results when compared to manual semantic classifications, however, these results remain noticeably non-humanlike (and thus unable to reliably replace manual classification without extensive human post-processing) due to their inability to capture precise semantic correspondences on the lexical level, instead only being consistent in their ability to capture relevant semantic domains. Correspondingly, smaller, more general semantic classification ontologies which employ only such domains (such as Rapid Words) produce more accurate semantic classifications than large, highly specific ontologies which operate on the level of individual lexemes (such as WordNet). As such, in order to improve vector semantic classification results with such ontologies, two evident approaches would be the reduction of their size and an increase in their generality.

With WordNet, we have conceptualised three potential approaches to this basic objective; firstly, a method which exploits WN's linear hierarchical structure to use hypernymic classifications (the Hypernymy Method), secondly, a method which uses only WN synsets at a certain preset level in the hierarchy as vector classifications (the Root Synset Method), and thirdly, a method which uses the hypernymic synset(s) of a large number of top-ranking vector classifications to determine a single top classification by plurality vote (the Voting Method). Only the first of these methods (the Hypernymy Method) has been carried out in full, and will be outlined later in this chapter (Section 5.2); the other two are only described here, and are to be attempted in future investigations.

5.1.1 The Hypernymy Method

The problem of generality and specificity in WordNet can, to a great degree, be represented through the ontology's hypernymy-hyponymy hierarchy, in which all of WordNet's nominal and verbal synsets are organised. Broadly speaking, the more lexically specific any given synset is, the lower down it is found in the hypernymy hierarchy. As previously mentioned, separate, distinct hierarchies exist for each of WordNet's four parts of speech; with nouns and verbs, these are linear, tree-like structures divided into discrete levels, wherein specific lexemes at a lower level are subordinated as hyponyms of more general lexemes at a higher level until a highly general root node is reached. Among adjectives and adverbs, a system of bipolar pairs is used to organise synsets by antonymy, being not so much hierarchical as simply pluricentric. Among nouns and verbs, however, the hypernymy hierarchy is purely linear, and one may follow any synset through its lineage in the hierarchy to access hypernymic, and thus more general, terms within the same semantic domain.

The exploitation of this structure forms the basis of one avenue of improvement for current WN vector classifications; namely, finding synsets at various hypernym levels above the existing vector classifications, and using these more general, hypernymic classifications instead. This 'Hypernymy Method' (Dacanay et al. 2022) also serves to broaden the number of synsets which may be considered matches to the manual classifications by means of the fact that a large number of specific synsets may be hyponymic to a single, more general one, reducing the precision necessary from the vector method to match an exact human classification, and in effect reducing both WordNet's total size and specificity by merging hyponymic synsets into their more general

shared hypernyms, albeit also relying on the vector and manual classifications of any given entry having a shared, hypernymic common-ancestor synset somewhere on the hypernymy hierarchy to begin with, as well as relying on the manual classification to be either a noun or verb (as adjectives and adverbs in WordNet lack hypernymy hierarchies), both of which are frequently not the case (see Section 5.2).

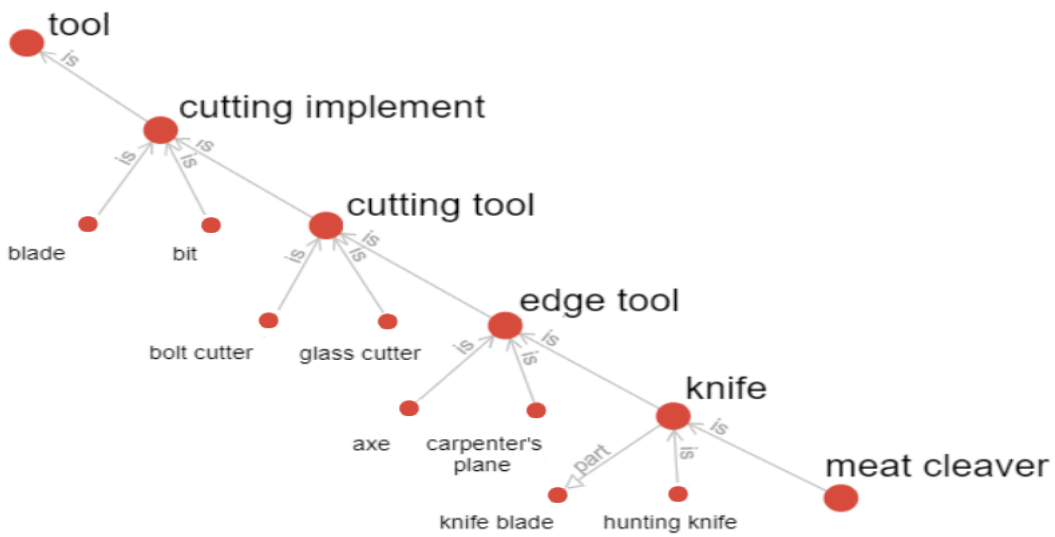

Figure 4, an example of the Hypernymy Method. For the Cree word *môhkomân* (‘knife’), one of the top vector classifications is *(n) meat_cleaver#1*, a semantically related, but overly specific synset. However, if one instead uses the hypernym of *(n) meat_cleaver#1* as the classification, it becomes *(n) knife#1*, a synset which more appropriately reflects the generality of the target Cree word. (Diagram via wordvis.com (Vercruysse 2010))

5.1.2 The Root Synset Method

Alternatively, one can exploit the hypernymy hierarchy of WordNet's nominal and verbal synsets by beginning at the root node(s) of the hierarchy and using only synsets which are a certain number of hyponymy levels below these node(s) as classification categories.

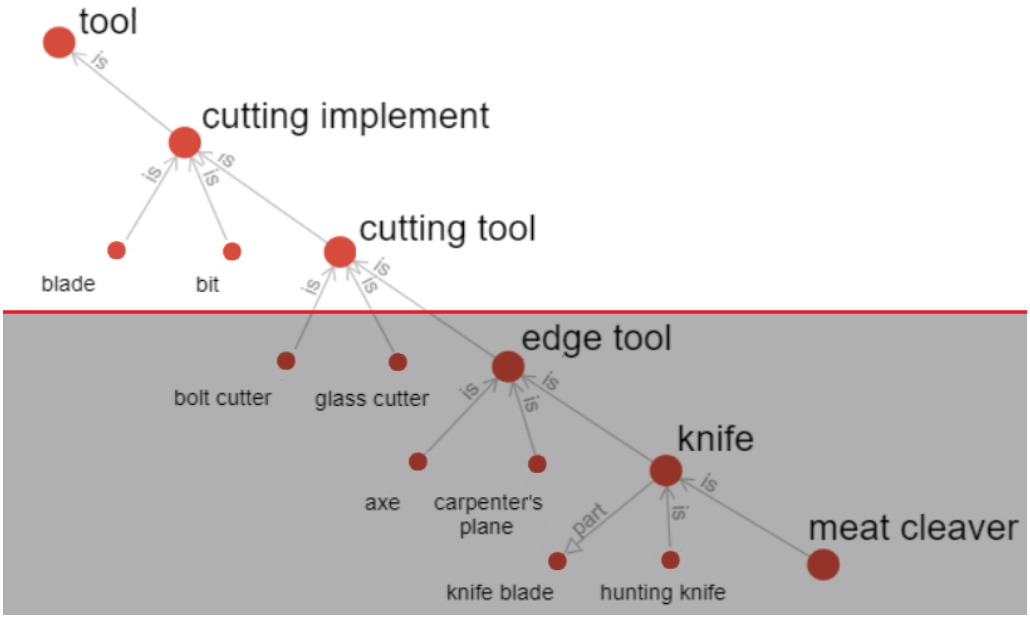

Figure 5, a demonstration of the Root Synset Method for vector classifications of *môhkomân*. If one forbids the vector method from selecting any categories lower on the hierarchy than *(n) cutting_tool#1*, then overspecific classifications such as *(n) meat_cleaver#1* will be impossible, with vector classifications instead being guided towards more generic synsets such as *(n) cutting_tool#1* or *(n) cutting_implement#1*, superficially emulating the more general domain structure of RW. (Diagram via wordvis.com (Vercruysse 2010))

This method has the advantage of both drastically reducing WordNet's apparent size through using only a reduced set of synsets while keeping its semantic breadth largely intact (in theory, reducing the semantic precision necessary of the vector method to 'match' manual classifications), as well as ensuring that all vector classifications are of a single, generic level of specificity, given that they are all at the same consistent level of the hypernymy hierarchy. Thus,

similar to the Hypernymy Method, this method largely functions to improve the accuracy of vector semantic classifications by reducing the precision necessary to obtain a match between these vector classifications and their manual counterparts. However, it also relies, to a degree, on manual classifications being uniformly at a similarly general level of the hypernymy hierarchy to begin with, which is not necessarily always the case, as well as relying on the manual classification to be a noun or verb, for the same reasons as the Hypernymy Method.

5.1.3 The Voting Method

Finally, rather than altering the level of precision or specificity of the individual vector classifications, one can instead exploit the fact that existing high-ranking vector classifications tend to be disparately semantically related to various aspects of the overall meaning of the Cree word by using a kind of vote, whereby one takes the hypernymic synsets for some number of top-ranking vector classifications (say, the top 100), identifies the most common hypernym among this set, and uses this hypernymic synset as the classification for the Cree word as a whole. In essence, this 'Voting Method' relies on a statistical implementation of the 'wisdom of the crowd' whereby, the collective average of a large number of imprecise, but feasibly semantically related vector classifications is thought to produce an end result which is as good as, or superior to, any one of its individual constituent classifications. A truncated example of this is given in Table 10; if one moves each of the top 15 vector classifications for the Cree word *masinahikan* ('book, letter, mail, written document, report, paper, magazine, will') up by three hypernymy levels, the most common classification becomes *(n) writing#2* (defined in WordNet as "the work of a writer; anything expressed in letters of the alphabet (especially when

considered from the point of view of style and effect)"), a more fittingly generic classification than, for example, *(n) letter_paper#1,* the initial top vector classification for *masinahikan*.

masinahikan ('book, letter, mail, written document, report, paper, magazine, will')			
Initial Vector Classification	Vector Classification, 1 Level up	Vector Classification, 2 Levels up	Vector Classification 3 Levels up
1. *(n) letter_paper#1*	*(n) writing_paper#1*	*(n) paper#1*	*(n) material#1*
2. *(n) order_paper#1*	*(n) writing_paper#1*	*(n) paper#1*	*(n) material#1*
3. *(n) order_book#1*	*(n) writing_paper#1*	*(n) paper#1*	*(n) material#1*
4. *(n) page#1*	*(n) leaf#2*	*(n) sheet#2*	*(n) paper#1*
5. *(n) review_copy#1*	*(n) book#1*	*(n) publication#1*	*(n) work#2*
6. *(n) manuscript#2*	*(n) autograph#1*	*(n) writing#2*	*(n) written_language#1*
7. *(n) holograph#1*	*(n) autograph#1*	*(n) writing#2*	*(n) written_language#1*
8. *(n) white_paper#1*	*(n) report#1*	*(n) document#1*	*(n) writing#2*
9. *(n) white_book#1*	*(n) report#1*	*(n) document#1*	*(n) writing#2*
10. *(n) missive#1*	*(n) text#1*	*(n) matter#6*	*(n) writing#2*
11. *(n) letter#1*	*(n) text#1*	*(n) matter#6*	*(n) writing#2*
12. *(n) document#1*	*(n) writing#2*	*(n) written_language#1*	*(n) communication#2*
13. *(n) written_document#1*	*(n) writing#2*	*(n) written_language#1*	*(n) communication#2*
14. *(n) papers#1*	*(n) writing#2*	*(n) written_language#1*	*(n) communication#2*
15. *(n) pamphlet#1*	*(n) book#1*	*(n) publication#1*	*(n) work#2*
MODE	***(n) writing_paper#1, (n) writing#2***	***(n) paper#1, (n) written_language#1***	***(n) writing#2***

Table 10, a demonstration of the Voting Method at 1, 2, and 3 levels of hypernymy with the first 15 vector classifications for the CW entry *masinahikan*.

5.2 Applying the Hypernymy Method

		Initial Vector Top 10000	Vector Top 10000 - 1 up	Vector Top 10000 - 2 up	Vector Top 10000 - 3 up
Manual	median	**18**	19	42.5	118
	count	3656	2364	1414	727
Manual - 1 up	median	420	**16**	20	54
	count	3191	3344	3181	2625
Manual - 2 up	median	1410.5	131	**13**	17.5
	count	2310	4114	3334	3446
Manual - 3 up	median	2733	395	64	**11**
	count	1525	3854	4281	3323

Table 11, results of the Hypernymy Method on CW nouns, showing the median rank of the manual classification(s) among the vector classifications (in the rows marked 'median') after moving either classification type 1, 2, and 3 levels higher in the hierarchy. For example, the median position of the manual classification among the vector classifications for a Cree noun is 20 when the manual classification is moved one level up and the vector classification is moved two levels up (column 5, row 4). The rows marked 'count' indicate the number of CW entries for which the indicated permutation is possible; to use the previous example, it was only possible to move the manual classification up one level and the vector classification two levels in 3181 instances (out of a total 5212 CW nouns). This method thus excludes any CW entries for which the only manual classifications were WN adjectival or adverbial synsets, as such synsets have no hypernyms. Note also that this table only shows results for instances in which the manual classification was found somewhere within the top 10 000 vector classifications; otherwise, the result was not counted (hence the differing numbers of instances for various permutations).

		Auto Top 10000	Auto top 10000 - one up	Auto top 10000 - two up	Auto top 10000 - three up
Manual	median	**210**	250	384	343
	count	7577	5831	3307	1975
Manual - one up	median	1205	**93**	94	102.5
	count	5627	8525	7778	5916
Manual - two up	median	1475.5	149	**49**	44
	count	2310	8708	8718	8334
Manual - three up	median	1624	156	62	**34**
	count	5372	8753	8953	8697

Table 12, as above, for CW verbs.

Although moving all WN classification types higher in the hypernymy hierarchy did indeed improve the median accuracy of vector classifications (from 15 to 11 and 333 to 34 for nouns and verbs respectively (column 6, row 8 on Tables 11 and 12)), the nature of these improvements differed in several respects from our initial expectations. Firstly, the most accurate results were consistently the product of moving both the manual classification and the vector classification higher in the hierarchy by the same number of levels; for example, the aforementioned most accurate results for both nouns and verbs resulted from both manual and vector classifications being moved 3 levels higher in the hierarchy. This seems to indicate that the ‘overspecificity’ of WN vector classifications was of relatively little influence compared with simple lack of precision, as if overspecificity of the vector classifications relative the manual classification was the leading cause of their inaccuracy, then moving the vector classifications higher in the hierarchy while retaining the manual classifications in their place would provide the most accurate results. However, in practice, doing this only worsens the accuracy of vector

classifications the more levels up they are moved without also moving the manual classification (see row 1 of Tables 11 and 12). Rather, the fact that it is only when moving both classification types higher in the hypernymy hierarchy by the same number of levels that consistent improvements are seen indicates that it is more often the case that both classifications are on more or less the same level of the hierarchy, and simply share a common ancestor synset higher on the hierarchy than either of them.

To provide an example of this phenomenon, one may examine the case of the Cree word *asikan* (‘sock, stocking’), which has the manual classification *(n) sock#1* and a top-ranking vector classification of *(n) toboggan_cap#1*. If one moves only the vector classification higher in the hypernymy hierarchy, the classifications become *(n) cap#1* at 1 level up, *(n) headdress#1* at 2 levels up, and *(n) clothing#1* at 3 levels up, none of which are matches for the manual classification. However, if one moves the manual classification higher as well (to *(n) hosiery#1* at 1 level, *(n) footwear#1* at 2, and (*n) clothing#1* at 3), the two classifications reach a match at three levels higher in the hierarchy, even though no such match was possible if only the vector or manual classification was moved in the hierarchy individually. Although moving both classifications higher in the hierarchy may be a slight deviation from the initial intentions of the Hypernymy Method, it ultimately achieves a similar effect, being a reduction in overall classificatory specificity.

Useful though it may be, the application of the Hypernymy Method is reliant on the assumption that the manual and vector WN classifications of any given entry have, somewhere on the hierarchy, a common ancestor synset, be that the manual classification itself or (more often)

some hypernymic synset to both the manual and vector classifications. However, even in instances in which the top-ranking vector classifications are highly semantically relevant, this is not always the case. For example, for the Cree word *kotikonikan* (‘breech-loading gun’), the top-ranking vector classification is *(n) breech#1*, which, although highly semantically related to the Cree word, is hyponymous in WordNet to the synset *(n) opening#10*, rather than any synset to do with firearms. As such, even if one moves this classification higher in the hypernymy hierarchy, it will not converge with the manual classification (*(n) breechloader#1*) or any of its hypernyms until an extremely general parent synset such as *(n) artefact#1*, at which point the classification is far too generic to be practically applicable for most purposes. However, it should be noted that, in this case, the Hypernymy Method can still improve *kotikonikan*’s classification accuracy overall, as moving both manual and vector classifications up by one level causes the 25th vector classification (*(n) cannon#4*) to converge with the manual classification at their shared hypernym of *(n) gun#1*, whereas in the original vector classifications, *(n) breechloader#1* does not occur anywhere within the top 1000.

As previously mentioned, the number of instances in which the top-ranked vector classification and the manual classification share a common ancestor synset are relatively few. For nouns, this was the case only 60.7% of the time, and for verbs, only 15.5% of the time, with the particular scarcity of verbal common ancestor cases likely being the result of the fact that WN verb synsets are arranged in hundreds of mutually disconnected hypernymy hierarchies rather than a single one such as with the nouns, reducing the likelihood that any common ancestor between two verb synsets exists. In any case, on account this relative infrequency of common-ancestor cases, particularly among verbs, which constitute the majority of entries in CW, the Hypernymy

Method in this form is a largely situational means of improving the accuracy of some vector classifications, while being either inapplicable or not a viable means of improvement without further annotation for the majority of CW entries.

CHAPTER 6. DISCUSSION AND CONCLUSION

6.1 Summary of General Observations

To summarise, even with modifications to limit the size and lexical specificity of WordNet, the use of a natively smaller, simpler classification ontology such as Rapid Words still seems to produce more accurate vector classifications overall, although in either case, the vector method is only able to select the 'most human-like' possible classification (that is, the manual classification) as the top-ranking match a minority of the time. As such, it appears that ontologically-based semantic classification tasks such as this are best served by more general, domain-level classification ontologies such as RW, providing superficially more accurate vector results, faster rates of manual classification, and avoiding the complications of requiring direct lexical matches within parts of speech, even though in proportional terms, WN still outperforms RW in vector trials.

6.2 Validity of Comparison with Manual Classifications

These claims that RW produces 'more accurate' vector classifications are predicated on the notion that accuracy may be defined in terms of strict correspondence to the exact manual classification(s). While this has proven thus far a parsimonious assumption for our various attempts to improve the accuracy of vector classifications, it is nonetheless a useful exercise to more critically examine the nature of these manual classifications as targets for vector classification in the first place. As mentioned throughout Section 2, CW entries were provided with as many manual classifications in WN and RW as was thought necessary to fully represent their meanings; however, in both ontologies, a substantial majority of entries were classified using only one or two WN synsets or RW domains (see Table 3 in Section 2.5). Despite this, for

many entries, even though only a single synset or domain was necessary to provide a reasonably full semantic classification, a great many more synsets or domains would still be (to varying degrees) fitting as 'human-like' classifications. However, because only the one or two most obvious manual classifications may have been given, these alternate, but feasible classifications are still considered by our criteria to not be 'human-like' matches, simply because they were not selected during the manual classification process. For example, for the Cree word *ayisinam* ('s/he mimics s.t.'), the single manual classification given in WN was *(v) mimic#1*, a synset which covers the breadth of meaning of the Cree word fairly comprehensively. However, there are a number of other WN synsets which would make equally reasonable classifications; *(v) imitate#1*, *(v) emulate#1*, and *(v) take_after#2*, to name a few, none of which were selected as manual classifications on account of the full meaning of *ayisinam* being covered by *(v) mimic#1*. This is less eminent of an issue with the RW classifications; virtually every suitable classification for *ayisinam*, for example, would fit within the RW domain *8.3.5.5 Imitate*.

The reasoning behind only a small number of manual classifications being given for each CW entry was a simple one, namely, that the more classifications are provided for any given entry, the longer the process of manual classification takes. However, it must be noted that this fact does influence the purported 'accuracy' of our vector classifications, being that it frequently narrows the possible 'human-like' selections to only a single classification, rather than a broader set of classifications which would also be humanly acceptable. As such, one may expect the 'accuracy' of vector classifications relative to a manual benchmark to vary depending on the verbosity of the manual annotator, with the numerically-quantified accuracy of vector

classifications improving to at least some degree the more manual classifications are listed for any given target-language entry.

6.3 Current Practical Usages of Vector Classifications

Given the accuracy of current vector classifications using the methods outlined in this investigation, full-scale replacement of manual annotators for semantic classification tasks of this kind remains unfeasible. Even in the best case scenario, with Cree nouns being classified according to Rapid Words domains, the top vector classification was only a match for the top manual classification(s) in 45% of cases, and with Cree verbs, which constitute a majority of the lexicon, these numbers are much lower, at approximately 20%. Given these figures, the full use of vector semantic classifications without manual post-processing would result in the majority of target-language vocabulary being classified in a non-human-like fashion, even with ontological modifications.

However, it should not be discounted that top-ranking vector semantic classifications in both WN and RW did still tend to bear semantic relevance to their target Cree entry, nor that the manual semantic classifications often did occur somewhere within the top several dozen vector classifications, if not as the exact top match. As such, present vector semantic results may still be of some direct use in increasing the feasibility of large-scale semantic classifications of lexical resources, not by replacing manual semantic annotation, but rather by serving as an accessory tool to these annotators.

At present, if one is to take the vector classifications using RW, which remain the most accurate, even compared to the various modified WN variants, the median position of the manual classification among the vector classifications is 2 for Cree nouns and 18 for Cree verbs. As such, by listing only the top 20 RW vector classifications for any given Cree word (noun or verb), one would have a greater than 50% chance that the manual classification would be within that list; for Cree nouns, it would be a greater than 70% chance. Thus, if prior to performing a manual semantic classification of a dictionary, one generated RW vector classifications for each entry (an operation which, as mentioned, would take only a few hours), and presented the manual annotator with a selection of the top 20 vector classifications for each entry, for the majority of entries, the annotator would be able to select the best possible classification from the entries on that list, rather than needing to search through the entire ontology. This should, in turn, both save time for the manual annotator and provide them with the ability to more quickly and systematically provide multiple classifications for any given entry if necessary, which should additionally increase the semantic richness of each entry's classification(s).

Finally, although not a direct use of the ontologically-structured classification methods outlined here, an additional use of word2vec vector generation on bilingual dictionary entries is the implementation of these vectors into online dictionary searches. By comparing the vector of a user's search query on an online dictionary with the vectors of the entries in that dictionary, one can return semantically related entries to a user's search query even if those entries do not contain any of the words present in the user's search, or indeed even if the user's search was for a word not present in the dictionary source at all. This can, in turn, serve to significantly improve searchability for dictionaries with small underlying databases, which is frequently the case for

low-resource languages. Based in part on the findings of the research underlying this thesis, semantic search capabilities of this kind have already been implemented into one online dictionary of Cree (the University of Alberta's https://itwewina.altlab.app/) as well as to dictionaries of Arapaho, Haida, Woods Cree, and Tsuut'ina; the exact process involved in these implementations is to be outlined in an upcoming paper (Arppe et al. in prep.). Alternatively, using the existing WN and RW classifications for CW, one can present the entries in CW in semantic groups to begin with, allowing users to browse vocabulary by domain without needing to specify a search query at all; this may be of additional utility in the case of Cree, as many older native speakers, having never received an education in the language, may be unfamiliar with the Standard Roman Orthography, and thus more comfortable searching words by semantic domains.

6.4 Future Research

In addition to resulting in the successful creation of not one, but several semantically classified versions of a large lexical resource for Plains Cree, as well as a theoretically cross-linguistically sound method of computational semantic classification for bilingual dictionaries of any language with English glosses, this investigation has also provided insight into a variety of other fruitful avenues of research into computational semantic classification as a methodology. Firstly, two of the aforementioned refinements to the WN classifications, the 'Root Synset' Method and the 'Voting' Method (Sections 5.1.2 and 5.1.3), have yet to be attempted, nor have vector classifications based on the Cree-specific semantic classification scheme employed by Visitor et al. (2013) (Section 2.3). Additionally, we have yet to attempt the generation of vectors for CW, WN, or RW using more complex vector generation tools which are able to model polysemy, such

as BERT; a change which should hopefully reduce the impact of English-specific polysemy and homography on the creation of Cree vectors, in lieu of generating them from Cree corpus sources. Finally, although the ontology-based vector and manual classifications methods outlined in this thesis have been thoroughly applied to CW, they have yet to see full use on bilingual dictionary sources of other languages; through observation of their use outside of Cree, a more nuanced understanding could be established of the degree to which these classification methods, and the semantic ontologies underlying them, are genuinely language-neutral.

6.5 Conclusion

The semantic classification of bilingual lexical resources such as dictionaries can serve a number of practical and academic purposes in the pursuit of resource creation for language revitalisation. While obtaining such classifications through purely manual annotation can be expedited through the use of semantic classification ontologies such as WordNet or Rapid Words, such manual methods remain limited in their efficiency by means of their reliance on human annotators, and can take months of dedicated labour, even for relatively terse, first-pass classifications. As an alternative, vector semantic models can be used to automatically classify dictionary entries with classification categories in an ontology or semantic classification scheme, even without large target language corpora, albeit at the cost of reduced accuracy to manual classifications and of varying degrees of semantic precision in vector classifications. As such, the use of more general, domain-based classification ontologies to this end, rather than highly lexically specific ontologies, appears to reliably produce the most accurate vector classifications, at least relative to their corresponding manual classes. However, even with ontological modifications designed to facilitate the obligatory usage of semantically general classification categories, vector

classifications still remain insufficiently ‘human-like’, that is, insufficiently accurate to manual classifications, to be employed wholesale as a replacement for manual annotation. However, current results are reliably accurate enough to be used as an accessory to manual annotation, providing manual annotators with lists of potential classifications for individual entries, but still allowing the final decision on which classification is most suitable to be made by a human. In this way, although vector semantic classification is a methodology of definite interest and has the potential to exponentially increase the accessibility of semantic classifications in the context of low-resource languages, it remains, for the time being, not capable of fully replacing manual classification as a means of productively grouping lexical items along semantic lines.

REFERENCES

Anderson, John R. 1996. ACT: A simple theory of complex cognition. *American Psychologist*, 51(4):355-365. doi:10.1037/0003-066X.51.4.355

Arppe, Antti, Jordan Lachler, Trond Trosterud, Lene Antonsen & Sjur N. Moshagen. 2016. Basic language resource kits for endangered languages: a case study of Plains Cree. *In* Claudia Soria, Laurette Pretorius, Thierry Declerck, Joseph Mariani, Kevin Scannell & Eveline Wandl-Vogt (eds.) *Proceedings of the Collaboration and Computing for Under-Resourced Languages: Towards an Alliance for Digital Language Diversity (CCURL 2016) Workshop, LREC*, pp. 1-8

Arppe, Antti, Andrew Neitsch, Jolene Poulin, Daniel W. Hieber, Atticus G. Harrigan & Daniel B. Dacanay. Forthcoming. Finding words that aren't there: Using word embeddings to improve dictionary search for low-resource languages. [*Manuscript in preparation*]

Bella, Gábor, Fiona McNeill, Rody Gorman, Caoimhin O Donnaile, Kirsty MacDonald, Yamini Chandrashekar, Abed Alhakim Freihat & Fausto Giunchiglia. 2020. A major Wordnet for a minority language: Scottish Gaelic. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 2812–2818, Marseille, France. European Language Resources Association.

Black, William, Sabri Elkateb & Piek Vossen. 2006. Introducing the Arabic wordnet project. In *Proceedings of the third International WordNet Conference* (GWC-06).

Bosch, Sonja E. & Marissa Griesel. 2017. Strategies for building wordnets for under-resourced languages: The case of african languages. *Literator: Journal of Literary Criticism, Comparative Linguistics and Literary Studies*, 38(1):8. doi:10.4102/lit.v38i1.1351

Brixey, Jacqueline, David Sides, Timothy Vizthum, David Traum & Khalil Iskarous. 2020. Exploring a Choctaw language corpus with word vectors and minimum distance length. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pp. 2746–2753, Marseille, France. European Language Resources Association.

Bundy, Alan & Lincoln Wallen. 1984. Semantic Primitives. *In* Alan Bundy & Lincoln Wallen (Eds) *Catalogue of Artificial Intelligence Tools*. *Symbolic Computation*. Springer, p. 120. doi:10.1007/978-3-642-96868-6_228

Collins, Allan M. & Elizabeth F. Loftus. 1975. A spreading-activation theory of semantic processing. *Psychological Review*, 82(6):407–428. doi:10.1037/0033-295X.82.6.407

Cowell, Andrew. 2012. The Hawaiian model of language revitalization: Problems of extension to mainland native America. *International Journal of the Sociology of Language,* 218:172. doi:10.1515/ijsl-2012-0063

Dacanay, Daniel, Atticus Harrigan & Antti Arppe. 2021a. Computational analysis versus human intuition: a critical comparison of vector semantics with manual semantic classification in the context of Plains Cree. *In* Antti Arppe, Jeff Good, Atticus Harrigan, Mans Hulden, Jordan Lachler, Sarah Moeller, Alexis Palmer, Miikka Silfverberg, and Lane Schwartz. (eds.) *Proceedings of the 4th Workshop on Computational Methods for Endangered Languages*, 1, pp. 33-43. doi:10.33011/computel.v1i.971

Dacanay, Daniel, Atticus Harrigan, Arok Wolvengrey & Antti Arppe. 2021b. The more detail, the better? – Investigating the effects of semantic ontology specificity on vector semantic classification with a Plains Cree / nêhiyawêwin dictionary. *In* Manuel Mager, Arturo Oncevay, Annette Rios, Ivan Vladimir Meza Ruiz, Alexis Palmer, Graham Neubig & Katharina Kann (eds.) *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas,* 1, pp. 143-152. doi:10.18653/v1/2021.americasnlp-1.15

Dacanay, Daniel, Jolene Poulin & Antti Arppe (forthcoming, 2022). kwêyask kotahâskwâtam: the effects of altering specificity in WordNet on the accuracy of computational semantic classifications of Plains Cree (nêhiyawêwin). *In* Monica Macaulay & Margaret Noodin (eds.), *Papers of the Fifty-Third Algonquian Conference (PAC53)*, 53. East Lansing, Michigan: MSU Press.

Davies, Mark. 2008. The Corpus of Contemporary American English (COCA). Available online at https://www.english-corpora.org/coca/.

Davies, Mark. 2018. The iWeb Corpus. Available online at https://www.english-corpora.org/iWeb/.

Davies, Mark. 2013. Corpus of Global Web-Based English. Available online at https://www.english-corpora.org/glowbe/.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee & Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. *In* Jill Burstein, Christy Doran & Thamar Solorio (eds.) *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* 1 *(Long and Short Papers)*, pp. 4171–86. doi:10.18653/v1/N19-1423

Dewra, Vijay Raj & Jonathan Dailey. 2015. *Marwari – English dictionar*y. Project LEARN, Webonary. URL:https://www.webonary.org/marwari/

Donaghy, Keola. 1998. Hawaiian language web browser released. Indigenous Language and Technology. http://listserv.linguistlist.org/cgi-bin/ wa?A2=ind0305&L=ilat&P=1734. Accessed 22 April 2022.

Eisenlohr, Patrick. 2004. Language Revitalization and New Technologies: Cultures of Electronic Mediation and the Refiguring of Communities. *Annual Review of Anthropology*, 33:21-45. doi:10.1146/annurev.anthro.33.070203.143900.

Ethnologue. 2015. Plains Cree (crk), https://www.ethnologue.com/18/language/crk/

Fellbaum, Christiane. 2000. Wordnet : an electronic lexical database. *Language*, 76(3):706. doi:10.2307/417141

Fellbaum, Christiane. 1990. English verbs as a semantic net. *International Journal of Lexicography*, 3(4):278–301. doi:10.1093/ijl/3.4.278

Firth, John R. 1957. A synopsis of linguistic theory 1930-55. In Frank R. Palmer (ed.) (1968). *Selected Papers of J. R. Firth 1952-59*, pp. 168-205

Galla, Candace K. 2009. Indigenous language revitalization and technology from traditional to contemporary domains. *In* Jon Reyhner & Louise Lockard (eds.) *Indigenous language revitalization: Encouragement, guidance & lessons learned*. pp. 167-182 Northern Arizona University Press, Flagstaff, AZ.

Hamp, Birgit & Helmut Feldweg. 1997. GermaNet - a lexical-semantic net for German. In *Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*

Harrigan, Atticus, Katherine Schmirler, Antti Arppe, Lene Antonsen, Sjur N. Moshagen, Trond Trosterud & Arok Wolvengrey. 2017. Learning from the computational modeling of Plains Cree verbs. *Morphology*, 27(4):565–598. doi:10.1007/s11525-017-9315-x

Harrigan, Atticus & Antti Arppe. 2021. Leveraging English Word Embeddings for Semi-Automatic Semantic Classification in Nêhiyawêwin (Plains Cree). *In* Manuel Mager, Arturo Oncevay, Annette Rios, Ivan Vladimir Meza Ruiz, Alexis Palmer, Graham Neubig & Katharina Kann (eds.) *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas.* pp. 113-121. doi:10.18653/v1/2021.americasnlp-1.12

Harris, Zellig S. 1954. Distributional structure, *Word*, 10:2-3, 146-162, doi: 10.1080/00437956.1954.11659520

Horváth, Csilla, Ágoston Nagy, Norbert Szilágyi & Veronika Vincze. 2016. Where Bears Have the Eyes of Currant: Towards a Mansi WordNet. In *Proceedings of the 8th Global WordNet Conference (GWC)*, pp. 131–135, Bucharest, Romania. Global Wordnet Association.

Hüllen, Werner. 2009. *Networks and knowledge in Roget's Thesaurus*. Oxford University Press, Oxford, UK.

Jokinen, Kristiina, Katri Hiovain, Niklas Laxström, Ilona Rauhala & Graham Wilcock. 2016. DigiSami and digital natives: Interaction technology for the North Sami language. *In* Kristiina Jokinen & Graham Wilcock (eds.) *Dialogues with social robots.* Springer. pp. 3-19. doi:10.1007/978-981-10-2585-3_1

Junker, Marie-Odile. 2018. *Grammar*. East Cree Language Resources FiveThirtyEight. https://www.eastcree.org/cree/en/grammar/. Accessed 11 November 2021.

Jurafsky, Daniel & James H. Martin. 2016. *Speech and Language Processing*, volume 3. Prentice-Hall, Hoboken, NJ. pp. 102-33

Keegan, Te Taka & Hōri Manuirirangi. 2011. Minority languages & translation technologies case study: te reo Māori & Google Translator Toolkit. *Proceedings of Translating and the Computer*, 33.

King, Levi & Markus Dickinson. 2014. Leveraging known Semantics for Spelling Correction. *In* Elena Volodina, Lars Borin, Ildikó Pilán (eds.) *Proceedings of the third workshop on NLP for computer-assisted language learning*, pp. 43–58.

Kröger, Franz. 2021. *Buli - English Dictionary*. Webonary.org. SIL International. URL:https://www.webonary.org/buli/

Lindén, Krister & Lauri Carlson. 2010. FinnWordNet–WordNet på finska via översättning. *LexicoNordica*, 17:121-3.

Littell, Patrick, Anna Kazantseva, Roland Kuhn, Aidan Pine, Antti Arppe, Christopher Cox & Marie-Odile Junker. 2018. Indigenous language technologies in Canada: Assessment, challenges, and successes. *In* Emily M. Bender, Leon Derczynski, Pierre Isabelle (eds.) *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 2620-32.

Littlechild, Mary Jean, Louise Wildcat, Jerry Roasting, Harley Simon, Annette Lee, Arlene Makinaw, Rosie Rowan, Rose Makinaw, Kisikaw, Betty Simon, Brian Lightning, Brian Lee, Linda Oldpan, Miriam Buffalo, Debora Young, Ivy Raine, Paula Mackinaw, Norma Linda Saddleback, Renee Makinaw, Atticus Harrigan, Katherine Schmirler, Dustin Bowers, Megan Bontogon, Sarah Giesbrecht, Patricia Johnson, Timothy Mills, Jordan Lachler & Antti Arppe. 2018. Towards a spoken dictionary of Maskwacîs Cree. Presentation conducted at the meeting of *Stabilizing Indigenous Languages Symposium (SILS)*, University of Lethbridge, Lethbridge, Alberta.

Lucas, Margery. 2001. Semantic priming without association: A meta-analytic review. *Psychonomic bulletin review*, 7:618–30. doi:10.3758/BF03212999

Marslen-Wilson, William D. & Pienie Zwitserlood. 1989. Accessing spoken words: The importance of word onsets. *Journal of Experimental Psychology: Human Perception and Performance*, 15:576–585. doi:10.1037/0096-1523.15.3.576

Marslen-Wilson, William D., Miljana Bozic & Billi Randall. 2008. Early decomposition in visual word recognition: Dissociating morphology, form, and meaning. *Language and Cognitive Processes,* 23(3):394-421. doi:10.1080/01690960701588004

Meighan, Paul J. 2021. Decolonizing the digital landscape: the role of technology in Indigenous language revitalization. *AlterNative: An International Journal of Indigenous Peoples*. 17(3):397-405. doi:10.1177/11771801211037672

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado & Jeffrey Dean. 1990. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 26(2):3111–3119. doi:10.48550/arXiv.1310.4546

Miller, George A. & Florentina Hristea. 2006. WordNet nouns: classes and instances. *Computational Linguistics*, 32:1-3. doi:10.1162/coli.2006.32.1.1.

Miller, George A., Richard Beckwith, Christiane Fellbaum, Derek Gross & Katherine J. Miller. 1993. Introduction to WordNet: an on-line lexical database. *International Journal of Lexicography,* 3(4):235–244. doi:10.1093/ijl/3.4.235

Moe, Ronald. 2003. Compiling dictionaries using semantic domains. *Lexikos*, 13:215-223, doi:10.5788/13-0-731

Outakoski, Hanna, Coppélie Cocq & Peter Steggo. 2018. Strengthening Indigenous languages in the digital age: social media–supported learning in Sápmi. *Media International Australia*, 169(1):21-31. doi:10.1177/1329878X18803700

Tan Ngoc Le & Fatiha Sadat. 2020. Revitalization of indigenous languages through pre-processing and neural machine translation: the case of Inuktitut. *In* Donia Scott, Nuria Bel, Chengqing Zong (eds.) *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 4661–6. doi:10.18653/v1/2020.coling-main.410

Talmy, Leonard. 1985. Lexicalisation patterns: semantic structure in lexical forms. *In* Timothy Shopen (Ed.) *Language typology and syntactic description*, Cambridge University Press, pp.57-149

Taylor, John R. 2008. Prototypes in cognitive linguistics. In Peter Robinson & Nick C. Ellis (eds.) *Handbook of Cognitive Linguistics and Second Language Acquisition*. Routledge, New York, NY.

Tremblay, Manon. 2005. *Les dictionnaires de la langue crie : histoire et regard critique*. Thèses et mémoires électroniques de l'Université de Montréal, Montréal, QC. 17-24

Reule, Tanzi. 2018. Elicitation and speech acts in the Maskwacîs Spoken Cree Dictionary Project. [*Honours Thesis*], University of Alberta.

Saddleback, Linda (ed.). 2009. *Maskwacîs Dictionary of Cree Words / Nêhiyaw Pîkiskwêwinisa*, Maskwachees Cultural College Maskwacîs, AB.

Schmirler, Katherine. Forthcoming. Syntactic features and text types in 20th century spoken Plains Cree: A constraint grammar approach. [*Unpublished Doctoral Dissertation*]. University of Alberta.

Solano, Rolando C., Sally A. Nicholas & Samantha Wray. 2018. Development of natural language processing tools for Cook Islands Māori. *In* Sunghwan Mac Kim, Xiuzhen (Jenny) Zhang (eds.) *Proceedings of the Australasian Language Technology Association Workshop 2018*, pp. 26–33, Dunedin, New Zealand.

Stanners, Robert F., James J. Neiser, William P. Hernon & Roger Hall. 1979. Memory representation for morphologically related words. *Journal of Verbal Learning & Verbal Behavior*, 18(4), 399–412. doi:10.1016/S0022-5371(79)90219-6

Stutzman, Verna & Kevin Warfel. 2022. Compiling dictionaries for minority and endangered languages. *In* Howard Jackson (Ed.) *The Bloomsbury Handbook of Lexicography*, Bloomsbury Publishing, pp.285-308

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

Vercruysse, Steven. 2010. WordVis, the visual dictionary. https://wordvis.com/. Accessed. 10 March 2022.

Vetulani, Zygmunt, Marek Kubis & Tomasz Obrębski. 2010. PolNet — Polish WordNet: Data and Tools. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA), pp. 3793-7. doi:10.1.1.676.1425

Vincze, Veronika & Attila Almási. 2014. Non-lexicalized concepts in Wordnets: a case study of English and Hungarian. *In* Heili Orav, Christiane Fellbaum & Piek Vossen (eds.) *Proceedings of the 7th Global WordNet Conference 2014 (GWC2014), Demonstration Session*, Tartu, Estonia. pp. 118–126..

Visitor, Linda, Marie-Odile Junker & Mimie Neacappo. 2013. Eastern James Bay Cree Thematic Dictionary (Northern Dialect). Cree School Board, Chisasibi, QC.

Vossen, Piek. 2004. EuroWordNet: a multilingual database of autonomous and language-specific WordNets connected via an inter-lingual index. *International Journal of Lexicography*, 17(2):161–173, doi:10.1093/ijl/17.2.161

Vossen, Piek. 1998. Introduction to EuroWordNet. *Computers and the Humanities*, 32:73–89. doi:10.1007/978-94-017-1491-4_1

Warschauer, Mark. 1998. Technology and Indigenous language revitalization: Analyzing the experience of Hawai'i. *Canadian Modern Language Review*, 55(1):140–161. doi:10.3138/cmlr.55.1.139

Weaver, Warren. 1955. Translation. *In* William N. Locke & Donald A. Booth (eds.) *Machine Translation of Languages*. MIT Press, Cambridge, MA.

Wiktionary: Statistics. 2022. In *Wiktionary*. https://en.wiktionary.org/wiki/Wiktionary:Statistics

Wolfart, H. Christoph. 1973. Plains Cree: a grammatical study. *Transactions of the American Philosophical Society*, 63(5):1-90. doi:10.2307/1006246

Wolvengrey, Arok. 2011. *Cree: Words*. University of Regina Press, Regina, CA.