Finding words that aren't there: Using word embeddings to improve dictionary search for low-resource languages

Antti Arppe, Andrew Neitsch, Daniel B. Dacanay, Jolene Poulin, Daniel W. Hieber, and **Atticus G. Harrigan**
Alberta Language Technology Lab, University of Alberta
https://altlab.ualberta.ca

Abstract

Modern machine learning techniques have produced many impressive results in language technology, but these techniques generally require an amount of training data that is many orders of magnitude greater than what exists for low-resource languages in general, and endangered languages in particular. However, dictionary definitions in a comparatively much more well-resourced majority language can provide a link between low-resource languages and machine learning models trained on massive amounts of majority-language training data. Promising results have been achieved by leveraging these embeddings in the search mechanisms of bilingual dictionaries of Plains Cree (*nêhiyawêwin*), Arapaho (*Hinóno'éitíit*), Northern Haida (*X̲aad Kíl*), and Tsuut'ina (*Tsúùt'ínà*), four Indigenous languages spoken in North America. Not only are the search results in the majority language of the definitions more relevant, but they can be semantically relevant in ways not achievable with classic information retrieval techniques: users can perform successful searches for words that do not occur at all in the dictionary. Not only this, but these techniques are directly applicable to any bilingual dictionary providing translations between a high- and low-resource language.

1 Introduction

This paper presents and evaluates an approach for improving the searchability of electronic dictionaries of low-resource languages, exemplified using bilingual dictionaries of Plains Cree (endonym: *nêhiyawêwin*; Glottocode: `plai1258`; ISO 639-3 code: `crk`), Arapaho (*Hinóno'éitíit*; Glottocode: `arap1274`; ISO 639-3 code: `arp`), Northern Haida (*X̲aad Kíl*; Glottocode: `haida1248`; ISO 639-3 code: `hdn`), and Tsuut'ina (*Tsúùt'ínà*; Glottocode: `sars1236`; ISO 639-3 code: `srs`), leveraging existing semantic embedding technology for majority languages in the novel context of low-resource minority languages.

Broadly speaking, search and information retrieval revolves around determining the means by which one may reliably find the most relevant discrete document(s) from a set of multiple such documents (or entries, in the case of dictionaries). In the case of bilingual dictionaries, presenting entry headwords in an minority Indigenous language with definitions in a majority target language, the definitions in the majority language (in our case, English) of each entry may be considered the"documents" one searches when using target language (English) search terms. The challenge, therefore, is determining how to find the most relevant Indigenous language words (which are the headwords of the entries) for these queries. This is particularly challenging when the sought-after target language definitions do not contain the exact search terms, but instead use related target language words; in these cases, even exact search word matches do not necessarily translate to the highest relevance. For instance, Indigenous languages that have a complex morphological system can store large amounts of information and meaning within a single lexeme, which in morphologically simpler languages (such as English) may need to be represented with multiple words or phrases. Consider, for example the Plains Cree words *nôtamiskwêw* for 's/he hunts beavers', and *êskêw* for 's/he makes a hole in the ice to hunt beaver; s/he breaks up a beaver lodge (i.e. in hunting)', which would both require a combination of the English search terms 'hunt' and 'beaver' to be accurately matched. In contrast, an entry such as *mâmawohkamâtowak*, meaning 'they do things together, they cooperate; they work (at it/him) together as a group; they assemble themselves to help one another.' would be matched with the search terms 'cooperate', or 'work' and 'together', but would be missed with the obvious synonym 'collaborate'.

Thus, the general problem remains determining the means to capture and represent the underlying meanings of both 1) the targets, the entries represented by their (English) definitions, and in 2) the (English) search terms, particularly when they may be more than the sum of the words in isolation. For endangered languages which are often also less-resourced ones, this challenge becomes greater as the vocabulary contents in their dictionaries (and definitions) are typically much more limited than those in majority languages, resulting in even fewer words to potentially match. For example, even many high-frequency English words, such as *national*, *administration*, and *network* (all within the top 1000 most frequent content words in large corpora such as COCA), have no matches whatsoever in the English definitions in any bilingual dictionaries of the four Indigenous languages named above. Searching with these words using typical methods would therefore result in `"No results"` – a discouraging outcome for a user – even when these dictionaries actually do contain relevant entries that could be shown.

Word-embeddings present one possible solution for this. Because they represent the underlying concepts that the individual words are pointing at, this allows us to represent concepts, or combinations of concepts, that individual words are pointing towards. In turn, this allows for comparing the concepts referred to by the search terms and the entry definitions, rather than the individual words themselves. This paper discusses the implementation and evaluation of this solution to four bilingual dictionaries between an Indigenous language and English, all of which we have made available online.[1]

2 Background and previous related work

2.1 How Indigenous lexical resources are (often) limited

The majority of endangered and Indigenous languages are extremely low-resourced, with corpora and lexical databases that are a fraction of the size of even basic learner's dictionaries in major languages such as English. Often these lexical databases are the product of fieldwork conducted by just one or a small number of linguists, in projects where financial and temporal constraints prevent the kind of extensive data collection that occurs for well-resourced languages.

For example, in a survey of 284 published dictionaries and lexical databases of lesser-resourced languages (Hieber, in progress), the mean number of entries per language is 5,772 and the median is 4,321, with only 39 sources containing more than 10,000 entries, and only five having more than 20,000. Only two sources—Mundari (Glottocode: `mund1320`; ISO: `unr`) and Marwari (Glottocode: `raja1256`; ISO: `mwr`)—reach 50,000 entries. By comparison, the Cambridge Learner's Dictionary of English (O'Shea et al., 2012)—marketed as covering only vocabulary relevant to the B1–B2 (intermediate) levels of CEFR (the Common European Framework of Reference, used for assessing language proficiency)— contains over 35,000 entries. This intermediate-level dictionary therefore contains more entries than all but two dictionaries in the history of Indigenous language documentation.

In addition to the aforementioned temporal and financial limitations limiting dictionary sizes in many Indigenous languages, many such languages also suffer from lexical attrition accompanying the process of language obsolescence (or "language death") (Sands et al., 2007). The remaining speakers may simply not remember as many words as their predecessors once did. For other languages, the number of lexemes may *in fact* be smaller than speakers of major Indo-European languages are accustomed to. Words in some languages may cover a broader semantic field, on average, than their Indo-European counterparts. Jack Martin (p.c.) notes for his lexical databases on U.S. Southeastern languages that "these numbers, while low by English standards, actually reflect a very high percentage of the words that are used".

Other languages have fewer lexemes by virtue of how their grammar operates. The Tsafiki language (a.k.a. Colorado; Glottocode: `colo1256`; ISO: `cof`), for example, has 4,000 lexical entries and only 32 true verbs, but includes another 6,000 subentries formed by adding suffixes to those 4,000 base entries to create new words (Dickinson, 2000). Inuit languages are likewise renowned for possessing thousands of lexical suffixes that can derive new words, even though the number of base roots is actually rather small. If dictionaries of these languages are based on roots rather than stems (as is often the case), lookup and search can become

[1] These on-line dictionaries are the following: *itwêwina* (Plains Cree-to-English) `https://itwewina.altlab.app`; *Nihíitono* (Arapaho-to-English) `https://nihiitono.altlab.dev`; *Gúusaaw* (Northern Haida-to-English) `https://guusaaw.altlab.dev`; and *Gūnáhà* (Tsuut'ina-to-English) `https://gunaha.altlab.dev`.

quite difficult for dictionary users, who must first locate the relevant main entry, and then the target subentry.

All this is to say that, for most documentary lexical databases, the number of entries is quite small compared to well-resourced languages. This fact creates a significant problem for potential users of these databases: because there are so few entries, it can be difficult to locate the entry most relevant to the user's search term. This problem arises in primarily two ways: 1) the language may not have a specific term for the (majority language) concept the user is searching for; and 2) the language has a term for the (majority language) search query, but no definition exactly matches that query. For instance, searching the Plains Cree-to-English dictionary (`http://creedictionary.com/search/?q=collaborate`) gives no result for the English search term *collaborate*, though this resource does provide matches for the semantically synonymous word *cooperate* and as well as synonymous multi-word expression *work together* (*mâmawatoskêwak* 'they work together´). Neither does one get a match for *procrastinate*, though the same dictionary does contain many entries concerning the semantically related concept *delay*, e.g. *otamihtwâsow* 's/he delays him/herself with work´.

In the first case, it would be useful if the dictionary could display results that are semantically related to the search term, or in a neighboring semantic field, or have some sort of semantic relationship to the search term (hypernymy, meronymy, antonymy, etc.), preferably with the results sorted by relevance. Thereby, one would hope to be given the same Plains Cree result *mâmawatoskêwak* for the search term *collaborate*, as for what is already provided for *co-operate* and *work together*. This is not how most electronic dictionaries historically have worked, and those dictionaries that do incorporate some measure of semantic association rely on massive datasets to accomplish it (see §2.3) – an approach not feasible for low-resource languages.

There are many causes for the second case, wherein a lexical database contains an entry that would be considered a correct match for the user's search term, but the user is unsuccessful in locating it. It may be the case that the language has a word for the search term, but the definition of that word does not encompass the entirety of the semantic breadth of the term. This is quite common for documentary lexicons, which are often based as much on wordlist elicitation as corpus data (usually more so), often resulting in only frequent, 'core' meanings of polysemous word entries being gathered. However, documentary lexicons are also more likely to focus on what are called *basic level* terms, that is, terms which are considered the most cognitively and linguistically salient (Taylor, 2003), to the exclusion of others. As a consequence, documentary lexicons often lack entries for terms that are either very high or low in ontological specificity; for example, they are likely to contain entries for 'arm' and 'leg' but less likely to contain entries for the more abstract 'appendage' or more specific 'paw'. In the above case, one would hope to be shown the results for *delay*, when searching with *procrastinate* (if no exact matches are to be found for this search term).

Entries that are multi-word expressions (MWEs) may also lead to less-than-ideal search results. In Plains Cree, for instance, there is a verb stem *mihcêtohk-* meaning 'to work together on something'. In a non–semantically-informed dictionary, the user must search for the exact phrase "work together" to find this entry. Searching for just "work" or "together" will likely return a host of irrelevant results such as *atoskê-* 'to work' or *miyopayin-* 'to work well' before *mihcêtohk-*, and searching with a synonym such as 'collaborate' would not yield *mihcêtohk-* among the results.

The definitional conventions of a dictionary can also significantly affect searchability. Definitions may be either *intensional* (describing the properties or necessary and sufficient conditions for a concept) or *extensional* (specifying the range or types of entities that fall within the concept (Svensén, 2009, 218–222)); for example, an intensional definition of *motor vehicle* would mention the need for a motor and use for transport/transit, etc., while an extensional definition might mention cars, motorcycles, mopeds, etc. If a user searches for one type of definitional style but the database adopts the other, lookup may fail if a semantically-informed search algorithm is not used.

Idiomatic expressions also cause difficulties for lookup, since users may search for the idiomatic meaning rather than the literal one (or vice versa). For example, the West Danish (Jutlandic) word *ræv* 'fox' also means 'sly, cunning person', and this idiomatic meaning is only sometimes included in dictionaries (Arboe, 2015, 162). In a traditional dictionary, if this sense were not included, users would not be able to find it in a search for “sly" or “cunning". For a semantically-informed dictionary, however, a search for either of these terms would very likely include *ræv* as a result.

As outlined, all of these problems may be addressed by a semantically-informed search algorithm which returns results based on semantic relevance to the search string. A more general advantage of this approach is that it allows users to search using nearly any semantic relationship (meronymy, hypernymy, etc.), and facilitates searching for less canonical types of entries, such as multi-word expressions, idioms, slang, etc. This is especially important given that the majority of lexical items sought by dictionary users tend *not* to be the canonical, single-word lexical item that dictionaries are often designed around. Research has found that users hardly ever look up common words; most searches are for idioms, encyclopedic-like information, culture-specific words, abbreviations, and slang (Svensén, 2009, 466).

As mentioned, however, implementing semantically-informed search has historically been no easy task for low-resource languages. In §3, we show how we implemented such a semantically-informed search algorithm for several low-resource languages.

2.2 Quantifying the challenge

The small size of most low resource language dictionaries inevitably results in a large number of high frequency majority language lemmata simply not occurring in any entries, resulting in a significant portion of even fairly innocuous majority language search queries returning no exact matches using traditional search methods. For Plains Cree, for example, only 88% of the top 1000 most frequent English lemmata are present within the definitions of the current dictionary, and for languages such as Tsuut’ina (Glottocode: `sars1236`; ISO 639-3 code: `srs`), this proportion is as low as 44.7% (Table 1).

The nature of the high-frequency vocabulary which tends to be missing in these four dictionaries is variable, but follows some general patterns, with common words relating to government, legislature, technology, and abstract concepts often being absent (such as ‘national’, ‘policy’, ‘data’, and ‘theory’, ranked by frequency in COCA at positions 311, 406, and 417, and 896 respectively). In total, 26 of the top 1000 most frequent English content lemmata did not occur in any of the four dictionaries mentioned in Table 1 (see Appendix A for the full list).

Top Lemmata	Plains Cree	Arapaho	Northern Haida	Tsuut’ina
100	99	100	93	91
200	194	198	174	145
300	287	295	249	197
400	374	393	315	237
500	462	486	378	280
600	554	578	441	325
700	639	668	494	348
800	719	759	543	385
900	809	853	598	418
1000	880	939	641	447

Table 1: Counts among the 1000 top most frequent English lemmata (as per COCA – the one-billion word Corpus of Contemporary American English), excluding function words (Davies, 2008)) not found in any definitions in dictionaries of Plains Cree (~23 000 entries), Arapaho (~25 000 entries, with some repeated lemmata (Cowell, 2012)), Northern Haida (~5500 entries (Lachler, 2010)), and Tsuut’ina (~12 500 entries, but primarily inflectional wordforms and paradigms, with a total lemma count in the low thousands)

As Table 1 demonstrates, lacunae such as aforementioned become markedly more prevalent as less frequent terms are used as search queries. However, even in instances where an exact match can be found, it may be useful for users (particularly learners) for semantically related terms to be returned as well. For instance, if a user searches for “yellow hat”, it may be of use for them to also receive entries such as “orange toque”. However, this strategy poses the further problem of sorting and presenting results in terms of relevance, as well as of determining the relative relevance of individual words in multi-word searches.

2.3 Previous approaches to expanding search

General search engines sort their search results using what is typically a proprietary sorting algorithm, making it difficult to build off of widely accepted forms of search relevance (Sullivan, 2002). Instead, it is best to examine other approaches to search retrieval and ranking, innovating and adapting these practices for the task at hand. This section outlines a number of prior approaches for search and ranking.

2.3.1 The Boolean model

The earliest approach to search result ranking is the Boolean retrieval model. This model creates a weight for each entry given the query terms, using the sum of all individual query term weights as the document weight. If a query term is in the entry, the model represents that with a 1, and with a 0 otherwise (Larson, 2012). The method then returns all entries marked as 1, with no ranking system for the results.

2.3.2 Machine-learning-aided search

A more complex approach to sorting relevant search results is to use deep learning or some form of matrix to determine how alike a search result is to the query entered (McDonald et al., 2018). However, these approaches require large amounts of training data, often more than exists in a given low-resource language. These models can still be leveraged by training them on the definitions in a bilingual dictionary, which are entered in a majority language, and providing a relevance ranking from majority language query terms to majority language definitions. This process is explained in further detail below.

2.3.3 Search by translation

In one previous description of multilingual information access, texts were translated from one source language into another target language for easier querying by the end user. This translation process was at first done manually and eventually automatically (Oard, 2012). This method presents some challenges, such as determining the original language of a text, that do not apply to the dictionary use case as the source and target languages of the dictionary are known. Applying this method to an online dictionary would mean translating each source language headword into its target language counterpart, or translating all query terms to match the dictionary entry language. However, since the dictionary already has definitions provided in the majority language, this work would be redundant. Thus, the information retrieval system should instead query on the definitions, as our approach does.

2.3.4 Search by synonym expansion

One successful example of improving search results through defining synonyms for entries may be seen in Shi et al. (2005). In their study, which specifically concerned biological terminology, they used pre-existing databases of similar terms for all biology-related entries to generate a network of synonyms, but relied on manual classification (using the Princeton WordNet (Miller, 1995; Fellbaum, 1998)) for general words and phrases, i.e. the non-medical information, in the database texts.

To circumvent the need for manual synonym classification, Zhang et al. (2017) derived a method for automatically determining synonymy. This approach, however, is only available for languages with large corpora, as it relies on creating a machine learning model in the source language to create a synonym web. This approach was successful in improving how results are clustered; as such, we used a tool based on the same word vector model below (namely, `word2vec`).

2.3.5 Semantic expansion

A final approach, would involve starting with a large pre-existing database, such as WordNet, and pairing it down to only the relevant terms for efficiency and ease of use, as was done by Turcato et al. (2000). However, this approach assumes that each low resource language entry has at least one direct synonym in the high resource language, and that the semantic hierarchies and relationships of a majority language WordNet would be applicable outright to the target language, two facts which are often untrue.

In the absence of pre-existing models to leverage for the creation of a synonym table, creating a synonym network for a low-resource language dictionary would require many hours of manual input while consulting a pre-existing word network database, such as WordNet. This has been done before for Plains Cree with some success (Dacanay et al., 2021a); however, in addition to being highly time-consuming, this method also relies on the aforementioned, typically incorrect assumption that majority language semantic categories can be applied uncritically to target language vocabulary.

2.3.6 Issues with previous approaches

While these approaches suffice for a variety of search-based problems, they do not tackle the problem in the context of a bilingual dictionary with minority language headwords. The last four approaches assume that users will only ever use majority language search terms, which is an unfair assumption. Furthermore, the data required to train any sort of neural network or to automatically classify entries into a word net or a group of synonyms is much larger than the data available for low resource languages, such as Plains Cree. As such, a new approach was required to adequately solve this search and ranking scenario.

3 Our approach

We will present our approach primarily with examples from *itwêwina* (`itwêwina.altlab.app`), an online intelligent bilingual dictionary application, making use of our *morphodict* platform[2] for Plains Cree – English, although we have implemented this feature also for bilingual on-line dictionaries for Arapaho, Northern Haida, and Tsuut'ina, and will present examples from the first two languages in the evaluation section further below.[3] *itwêwina* is freely accessible to the public and receives roughly 20,000 searches per month. It combines multiple dictionary sources (Wolvengrey; Maskwachees Cultural College, 2009; LeClaire and Cardinal, 1998), and has approximately 22,000 headwords[4], of which only about 10,000 appear in any Plains Cree corpus that we know of. Through modeling with finite-state transducers (FST) (Snoek et al., 2014; Harrigan et al., 2017), it can dynamically recognize wordforms and display paradigm tables for millions of additional inflected word-forms.

Searches can be entered in either English or Plains Cree. We break the search process into two phases: retrieval, and ranking. The goal of retrieval is to find potentially relevant definitions for the input query. For Cree-language searches, a spell-relaxed finite state transducer identifies potential matching headwords. For English-language searches, the application uses classical information retrieval techniques of matching stemmed keywords between queries and definitions. Ranking is necessary because an unsorted list of matches would provide a poor user experience: there may be many hundreds of potentially matching words. Therefore, results returned by the finite-state transducer analyzer or classical information retrieval methods are ranked using a combination of result features[5] (Turnbull and Berryman, 2016) such as corpus or dictionary frequency, or edit distance.

In an attempt to improve the relevance of the top search results returned by *itwêwina*, we added in spring of 2021 a new result feature to feed into the relevance ranking function: a semantic distance measure, based on word embeddings, between input queries and resultant definitions. While this did improve relevance, the most novel and surprising feature which this revealed was the ability of word embeddings to allow the retrieval of useful search results for words that are not even in the dictionary (target language definitions), in addition to improving the search results for multi-word phrases.

4 Method

4.1 Word embeddings

A word embedding is a dimensionality reduction technique that assigns a relatively low-dimensional vector to each element of a set of words, in a way that captures relationships between words. The vector typically consists of first-layer model weights learned during the training of a neural network. For example, the 2013 `word2vec` model of Mikolov et al. (2013a,b) provides for each of 3 million words and phrases, not a 3-million-dimensional vector without semantic relationships, but instead a 300-dimensional vector with semantic relationships. The word embedding model is trained on a portion of a corpus of approximately 100 billion words[6] of Google News articles. The training process attempts to minimize the errors in predicting which words are most likely to occur surrounding any given input word. This necessarily assigns similar vectors to words that frequently occur in similar contexts in the corpus; thus, words corresponding to similar vectors are semantically related. Furthermore, these semantic relations are often seemingly algebraic in nature, allowing (in some instances (Ethayarajh et al., 2019) (Rogers et al., 2017)) for the automated solving of word analogy equations.

4.2 Application to low-resource languages

We do not have 100 billion words of Plains Cree text to train an equivalent model on, or for any of the three other Indigenous languages discussed in this paper. The largest Cree corpus has some 150 thousand word tokens (Arppe et al., 2020). However, we can use the vectors for English, and their algebraic nature, to compute vectors for every English definition as an average of their constituent

[2] The codebase which implements this ranking feature for all these languages is publicly available: `https://github.com/UALbertaALTLab/morphodict`

[3] Tsuu'ina examples have been left out, as its bilingual dictionary source is only a glossary based on a small collection of texts with a relatively restricted and skewed vocabulary.

[4] This was the value in 2021 when the the quantitative study presented in this paper was done, after which this number has grown to more than 25,000 entries.

[5] `https://web.stanford.edu/class/cs276/`

[6] `https://code.google.com/archive/p/word2vec/`

individual English words, and compare those to the similarly calculated vector for the English input query. This has the additional benefit of working for bilingual dictionaries for *any* low-resource language for which there are pre-trained word embeddings for the language used in the definitions. Indeed, as mentioned, we have already implemented this search feature for bilingual on-line dictionaries of Arapaho, Northern Haida, and Tsuut'ina, all with definitions in English.

When the dictionary data is loaded into the system, we use the Google News vectors to compute a vector for the English definition of every Plains Cree entry by adding up vectors for each word of the definition (Harrigan and Arppe, 2021; Dacanay et al., 2021a,b). For example, for the definition "yellow hat" of *osâwastotin*, we compute $v(\texttt{yellow hat}) := v(\texttt{yellow}) + v(\texttt{hat})$ and save that. This yields a vector for the definition overall, $v(\texttt{osâwastotin}_1) := v(\texttt{yellow hat})$. When there are multiple definitions for a Plains Cree word, we save a vector for each one so that we can show the word as a result if any definition is a good match.

When someone searches for "yellow hat," we again use the news vectors to compute a vector for the input query, $v(\texttt{yellow}) + v(\texttt{hat})$, and measure its distance to every definition in the dictionary (as the cosine between the two vectors). This is one instance in which the small available lexicons of these language is actually an advantage, as it is much faster to compare the search query vector to each of the over 22,000 Plains Cree definitions than it would be for the much larger number of definitions in a more comprehensive dictionary. In this case, while classical information retrieval techniques would have had little difficulty finding results for 'yellow' or 'hat' individually, the word embedding-based model retrieves not only 'yellow' and 'hat', but also other combinations of colour and clothing not specified in the search:

1. *osâwastotin*: yellow hat
2. *nîpâmâyâtastotin*: purple hat
3. *astotin*: hat, cap, headgear
4. *osâwêkin*: yellow material, yellow cloth
5. *osâwasâkay*: yellow dress, coat

However, as mentioned, this search method can also return relevant results for queries entirely absent from the database. For example, despite having no definition for 'freighter' (indeed, no definition even containing that word), using the word embeddings, a search for 'freighter' turns up *nâpihkwân* "ship, large boat" as the top result. This is because the word embeddings of the definition suggest semantically related concepts: 'boat,' 'ship.' Our approach is similar to the reverse dictionary lookup for Wolastoqey (Passamaquoddy-Maliseet) evaluated by Bear and Cook (2022).

This word embedding method can also be used to automatically cluster words into semantic classes. In its most basic form, this can be done simply by making use of hierarchical agglomerative clustering based on a distance matrix of the word vectors. While this technique produces useful and valuable clusters out-of-box, further manual adjustment significantly improves results (Harrigan and Arppe, 2021).

5 Results and evaluation

5.1 Qualitative assessment

In practice, our semantic search functionality returns results for 'missing' words (in the English definitions) with varying degrees of quality; for the sake of qualitative assessment, we may divide these result qualities into three (subjective) categories: high, moderate, and poor. A high quality result describes an instance in which the top search result for a missing word is either synonymous with, or highly semantically related to, the query word in question. Examples of missing words with high quality top matches include 'policy', which returns the Cree entry *wiyasiwêwin* ("law, rule, decision, council, band council, office"), 'attorney', which returns *oyasiwêwiyiniw* ("band councillor, court judge, lawyer"), and 'pdf', which returns *masinahikan* ("book, letter, mail, written document, ..."). Among the top 26 highest frequency English lemmata from COCA which do not appear in any of the low-resource language dictionaries previously mentioned in section 1.1, Anglophone manual annotators evaluated 18 of the top results for Plains Cree and Arapaho, and 5 of the top results for Northern Haida as being of high quality.

Moderate quality results describe those in which the top match is broadly, but not precisely, semantically related to the query word; examples of this include 'international', which returns the Cree entry for *opîtatowêw* "Ukrainian, European", a related, but decidedly non-synonymous term. Three of the top 26 missing words for Plains Cree and Arapaho, and 11 for Northern Haida were evaluated as having top matches of moderate quality.

Poor quality results are those in which the top match is either entirely semantically unrelated, or sufficiently irrelevant to be of no use. Examples of poor quality results include ‘percent’, which returns the Cree entry for *nisto-sôniyâs* “three quarters, seventy-five cents”. For a term such as ‘percent’, the most appropriate match in current Plains Cree dictionaries would be one relating either to portions (such as *pahki-* “portion of”) or relating to the number “one-hundred” (*mitâtahtomitanaw*), however, neither of these results are returned. Another example is ‘career’, for which the top match is the Cree entry for *ispîhtaskîwin* “season”, rather than the more fitting *atoskêwin* “work, labour, employment, job, contract, industry”. 5 of the top 26 missing words for Plains Cree and Arapaho, and 10 for Northern Haida were evaluated as having poor quality top matches.

In total, for the two larger bilingual dictionaries (for Plains Cree and Arapaho), a substantial majority of top results for missing words were of high or moderate quality (21 out of 26 in both cases), with the smaller Haida dictionary performing more poorly outright (likely because of its reduced semantic coverage by virtue of size; if no entries corresponding even to the basic semantic domain of a search query are to be found, then even a theoretically perfect semantic search would not return semantically relevant results). However, even with the Haida dictionary, a majority (16 out of 26) of top results for missing words were of high or moderate quality. Results for the mean number of high, moderate, and poor matches in the top ten search results of the top 26 highest frequency missing English lemmata for these three dictionaries are detailed in Table 2.

When considering the position among all returned results of the single most semantically relevant match for these 26 missing word search queries (as per a manual annotator), the median position of this match was 2 for Plains Cree, 3 for Arapaho, and 4.5 for Haida. As such, even for the relatively small Haida dictionary, a user would typically only need to scroll through the top five search results when searching for a missing word to find the most relevant match.

5.2 Search terms with metaphorical meaning

As mentioned in section 1, this search method also allows for the use of (English) metaphorical terms as search queries; for example, when searching the English term ‘soapbox’ in Plains Cree, the top two results are *kakêskihkêmowinâhtik* “pulpit, lecturn” and *kîhkâwitaskiw* “s/he likes to scold, s/he is always cross and scolding in a loud voice”, and the top result for ‘snowballing’ is *asascikêwin* “piling things together”, with *kîpikin* “it grows quickly” being in 7th. However, the success of these metaphorical search queries remains inconsistent. For example, the top ten results for ‘snake’ contain only entries related to reptiles, and none related to deceitful, malicious humans. Similarly, multi-word metaphors tended to return results relating to the literal meaning of their constituent elements (for example, “cabin fever” returns only results relating to cabins, e.g. *wâskâhikanis* “small house, cabin”, and fevers, e.g. *sîkwâspinêwin* “spring fever”, rather than to loneliness or boredom).

	High	Moderate	Poor
Plains Cree	3.08	3.23	3.08
Arapaho	3.65	3.15	3.19
Northern Haida	0.54	3.23	6.23

Table 2: Mean number of *high*, *moderate*, and *poor* quality results in the top ten matches for missing English lemma search queries.

5.3 Polysemous search terms

On a related note, one of the most notable errors in general with our semantic search method concerns polysemous English search queries, this being largely a product of `word2vec` generating embeddings on the level of the individual word, rather than the sentence. In addition to affecting the accuracy of metaphorical search terms, this also has the effect of semantically grouping target language words based on irrelevant English collocations. For example, when given the search term ’administration’, all three dictionaries returned entries containing the word ‘bush’ within their top ten results (such as *hlk’awáng* “for S to clear C [land] of bush or trees” in Northern Haida) due to the frequent occurrence of the collocation ‘Bush administration’ in news corpora, and matches for the search query ‘reality’ contained the word ‘television’ (such as *wó3onikúu3o.o* “movie, television show, picture, photograph” in Arapaho) within the top ten results of all three dictionaries.

5.4 Multi-word search terms

The quality of search results for multi-word expressions tended to vary depending on the semantic transparency of the expression’s constituent elements. For example, when searching for the expression 'simmer down' (in which all of the constituent words are transparently related to the end meaning of the expression), the top result in the Northern Haida dictionary is *sahl ts'asäláng* 'for S to let C boil without stirring it [said of fish only]'. However, when searching for an expression such as 'blow up' (whose constituent words are only idiosyncratically related to the action described), the top results (shown here in the Plains Cree dictionary) relate to the more conventional meanings of both words individually (*pôtâtam* "s/he blows at s.t. ...", *matwêtahikêw* "s/he strikes blow"), rather than to the more fitting meaning of the full phrase (which might rather best be expressed through *pahkitêw* "it explodes"). Similarly, idiomatic multi-word phrases behaved in much the same way as idiomatic expressions in general, overwhelmingly returning results relating to the literal meanings of their constituent elements; for example, 'see eye to eye' returns top results relating to eyes and sight (e.g. *miskîsikos* "eye, small eye, little eye", but no results related to the phrasal meaning of understanding.

These results are perhaps unsurprising, given the means by which our word embeddings were generated; `word2vec` being a tool which creates embeddings for individual words with their context taken only as a bag-of-words, rather than for creating them for whole phrases, it is to be expected that the isolated meanings of each individual word in a multi-word expression would take precedence over the meaning of the phrase as a whole. One possible means of addressing this would be the use of a sentence-based language model such as `BERT` ((Devlin et al., 2019)), which is able to generate contextualised word embeddings based on specific sentential surroundings, possibly allowing for a better modelling of common, semantically opaque multi-word phrases.

5.5 Preliminary quantitative assessment

In addition to its ability to return useful results for entries not present in the dictionary, the use of word embeddings can also improve relevance ranking for results which *are* present. Although more rigorous analysis is needed, preliminary results indicate that use of word embedding distance increases one of our key search quality metrics from 0.61 to 0.70 (this metric being a measure of the frequency with which certain desirable results appear as a top-10 result on a test set of 549 sample core vocabulary item queries).

6 Future work and conclusion

One potentially promising research theme to explore is the improvement of multi-word vector creation methods for definitions and search strings, perhaps through the use of term-frequency—inverse-document-frequency weights when adding word vectors to form definition vectors. Similarly, investigating whether newer pre-trained word embedding models (Pennington et al., 2014; Speer et al., 2017) could produce higher-quality results. Advances in the use of word embeddings for other NLP tasks in low-resource languages (e.g. Adams et al. 2017) may also translate to improved dictionary search.

7 Limitations

When a dictionary for a low-resource language lacks a word, but has several related ones in terms of synonymy or semantic similarity, it is a definite benefit to be able to provide those to the dictionary user instead of merely saying, "No results found." However, there are some potential drawbacks here: for example, this could increase the rate at which words acquire connotations by analogy with English. 'Locomotive' and 'train' are closely related concepts in English; but that does not necessarily hold for every language, and there is some risk in implying that it does.

Language instructors will be all too familiar with students using tools like Google Translate to do their homework for them instead of doing the hard work of learning the language. On a larger scale, Google Translate itself was formerly available as a free service that software developers could use to do automated machine translation in bulk; this was abruptly discontinued in 2011. Industry rumour[7] held that the bulk service was being used to generate so much of the parallel text appearing on the internet—parallel text needed to train machine translation models—that those models could no longer improve sufficiently if they continued to inadvertently be fed primarily their own outputs. This highlights the possible risk that applying machine learning tools like word embeddings can end up distorting language. To this end, we believe that the use of word embeddings to provide analogous

[7] https://kv-emptypages.blogspot.com/2011/06/analysis-of-shutdown-announcements-of.html

words to dictionary users is beneficial, but does not and cannot replace actual lexicography.

8 Ethics Statement

The on-line dictionaries described in this manuscript have been developed in order to support the explicit objectives of the language communities in question, to support their language instruction, maintenance, and revitalization activities.

9 Acknowledgements

We thank the Social Sciences and Humanities Research Council (SSHRC) for their Partnership Grant (#895-2019-1012) that has supported this project during the last several years. Earlier stages of the software development of *morphodict*, our morphologically intelligent on-line dictionary platform that we make use of, were implemented by Eddie Antonio Santos. We would also want to recognize the extensive documentation work by Arok Wolvengrey (Plains Cree), Andy Cowell (Arapaho), Jordan Lachler (Northern Haida), Chris Cox (Tsuut'ina), and Bruce Starlight (Tsuut'ina), not to mention the significant contributions of Elders, knowledge keepers, and speakers in the various Indigenous language communities, in compiling together the lexicographical resources incorporated in the intelligent dictionaries discussed in this paper.

References

Oliver Adams, Adam Makarucha, Graham Neubig, Steven Bird, and Trevor Cohn. 2017. Cross-lingual word embeddings for low-resource language modeling. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 937–947.

Torben Arboe. 2015. Receding idioms in West Danish (Jutlandic). In Elisabeth Piirainen and Ari Sherris, editors, *Language endangerment: Disappearing metaphors and shifting conceptualizations*, pages 155–173. John Benjamins, Amsterdam.

Antti Arppe, Katherine Schmirler, Atticus G Harrigan, and Arok Wolvengrey. 2020. A morphosyntactically tagged corpus for Plains Cree. In *Papers of the 49th Algonquian Conference (PAC49)*, volume 49, pages 1–16.

Diego Bear and Paul Cook. 2022. Leveraging a bilingual dictionary to learn Wolastoqey word representations. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1159–1166, Marseille, France. European Language Resources Association.

Andy Cowell, editor. 2012. *English-Arapaho Dictionary, 4th ed.* Center for the Study of the Indigenous Languages of the West.

Daniel Dacanay, Atticus Harrigan, and 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 *Proceedings of the 4th Workshop on the Use of Computational Methods in the Study of Endangered Languages Volume 1 (Papers)*, pages 33–43.

Daniel Dacanay, Atticus Harrigan, Arok Wolvengrey, and 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 *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 143–152.

Mark Davies. 2008. The Corpus of Contemporary American English (COCA).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Connie Dickinson. 2000. Complex predicates in Tsafiki. *Proceedings of the Annual Meeting of the Berkeley Linguistics Society*, 26(2):27–37.

Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. 2019. Towards understanding linear word analogies. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3253–3262, Florence, Italy. Association for Computational Linguistics.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Atticus Harrigan and Antti Arppe. 2021. Leveraging English word embeddings for semi-automatic semantic classification in *Nêhiyawêwin* (Plains Cree). In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 113–121.

Atticus G Harrigan, Katherine Schmirler, Antti Arppe, Lene Antonsen, Trond Trosterud, and Arok Wolvenengrey. 2017. Learning from the computational modelling of Plains Cree verbs. *Morphology*, 27(4):565–598.

Daniel W. Hieber. in progress. A survey of lexical resources in low-resource languages.

Jordan Lachler, editor. 2010. *Dictionary of Alaskan Haida*. Sealaska Heritage Institute.

Ray R. Larson. 2012. *Understanding Information Retrieval Systems*, chapter Information Retrieval Systems. Auerbach Publications.

Nancy LeClaire and George Cardinal, editors. 1998. *Alberta Elders' Cree Dictionary*. Duval Publishing House Ltd.

Maskwachees Cultural College. 2009. *Maskwacîs dictionary of Cree words / nêhiyaw pîkiskwêwinisa*.

Ryan T. McDonald, Georgios-Ioannis Brokos, and Ion Androutsopoulos. 2018. Deep relevance ranking using enhanced document-query interactions. *CoRR*, abs/1809.01682.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *International Conference on Learning Representations*, Scottsdale, Arizona, USA.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

George A. Miller. 1995. Wordnet: A lexical database for English. *Communications of the ACM*, 38(11):39–41.

Douglas W. Oard. 2012. *Understanding Information Retrieval Systems*, chapter Multilingual Information Access. Auerbach Publications.

Stella O'Shea, Helen Waterhouse, et al., editors. 2012. *Cambridge learner's dictionary*, 4 edition. Cambridge University Press, Cambridge.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Anna Rogers, Aleksandr Drozd, and Bofang Li. 2017. The (too many) problems of analogical reasoning with word vectors. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, pages 135–148, Vancouver, Canada. Association for Computational Linguistics.

Bonny Sands, Amanda L. Miller, and Johanna Brugman. 2007. The lexicon in language attrition: The case of Nluu. In *Selected Proceedings of the 37th Annual Conference on African Linguistics*, pages 55–65, Somerville, MA. Cascadilla Proceedings Project.

Zhongmin Shi, Baohua Gu, Fred Popowich, and Anoop Sarkar. 2005. Synonym-based expansion and boosting-based re-ranking: A two-phase approach for genomic information retrieval. In *Text Retrieval Conference (TREC)*.

Conor Snoek, Dorothy Thunder, Kaidi Lõo, Antti Arppe, Jordan Lachler, Sjur Moshagen, and Trond Trosterud. 2014. Modeling the noun morphology of Plains Cree. In *Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 34–42.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 4444–4451. AAAI Press.

Danny Sullivan. 2002. How search engines work.

Bo Svensén. 2009. *A handbook of lexicography: The theory and practice of dictionary-making*. Cambridge University Press, Cambridge.

John R. Taylor. 2003. *Linguistic categorization*, 3 edition. Oxford Textbooks in Linguistics. Oxford University Press, Oxford.

Davide Turcato, Fred Popowich, Janine Toole, Dan Fass, Devlan Nicholson, and Gordon Tisher. 2000. Adapting a synonym database to specific domains. In *Proceedings of the ACL-2000 Workshop on Recent Advances in Natural Language Processing and Information Retrieval: Held in Conjunction with the 38th Annual Meeting of the Association for Computational Linguistics - Volume 11*, RANLPIR '00, page 1–11, USA. Association for Computational Linguistics.

Doug Turnbull and John Berryman. 2016. *Relevant Search: With Applications for Solr and Elasticsearch*. Manning.

Arok Wolvengrey, editor. *nēhiýawēwin: itwēwina / Cree: Words*. University of Regina Press.

Li Zhang, Jun Li, and Chao Wang. 2017. Automatic synonym extraction using word2vec and spectral clustering. In *2017 36th Chinese Control Conference (CCC)*, pages 5629–5632.

A Appendix

The following is a list of all 26 English lemmas within the top 1000 most common content lemmas in COCA (Davies, 2008) which are not present in the definitions of any of entries in the consulted dictionaries of Plains Cree, Arapaho, Northern Haida, and Tsuut'ina, along with their frequency rank in COCA overall (including function words).

1. percent (265)
2. national (311)

3. policy (406)
4. data (417)
5. international (616)
6. campaign (634)
7. author (680)
8. administration (744)
9. career (796)
10. candidate (830)
11. network (882)
12. district (885)
13. theory (896)
14. reality (956)
15. democratic (1020)
16. democratic (1028)
17. politics (1059)
18. user (1081)
19. attorney (1102)
20. budget (1107)
21. senator (1144)
22. Senate (1155)
23. violence (1156)
24. civil (1171)
25. institution (1190)
26. professional (1192)