# Low hanging fruit and the Boasian trilogy in digital lexicography of morphologically rich languages: Lessons from a survey of Indigenous language resources in Canada

Elizabeth Pankratz, Antti Arppe, and Jordan Lachler
*University of Alberta*

**Abstract**

Online lexicographical resources for the morphologically rich Indigenous languages in Canada use a wide range of strategies for conveying their language's morphological system, i.e. how words are inflected and derived, which this paper illustrates in a survey of seventeen bilingual online resources. The strategies these resources employ boil down to two basic approaches to the underlying structure of the resource: 1) a lexical database, or 2) a computational model. Most resources we surveyed are constructed around lexical databases. These assume the word(form) as the basic unit, an assumption that makes it difficult to incorporate the language's sub-word, morphological structure in full detail. However, one resource uses a computational morphological model to bring the language's morphology into the core of the lexicon – this proved to be a "low-hanging fruit" in the application of language technology that had been accomplished within a reasonable time-frame, as has been advocated by Trond Trosterud. We discuss the value created and questions raised by this approach and argue that it successfully overcomes the traditional Boasian three-way partition of dictionary, grammar, and text, creating integrated language resources that meet the modern needs of low-resource endangered languages and their communities.

Keywords: Indigenous languages, electronic dictionaries, lexicography, morphology, finite-state modeling, Plains Cree, Canada

## 1 Introduction

Modern learners of majority languages have many high-quality digital resources at their fingertips, like online dictionaries, linguistically analyzed collections of written and spoken language, spell-checkers, grammar-checkers, and language-learning applications. It is easy to take access to these for granted. However, learners of smaller, lower-resourced, and often endangered languages face a different situation. Such languages are much less likely to have high-quality digital resources, even though these are the languages whose need for them is greatest. Strong digital resources make the language more accessible to current speakers and learners, facilitating language revitalization, and they are also valuable for documentation and preserving the language for the future (cf. Trosterud 2006; Arppe et al. 2016: 6).

So that any of these goals can be achieved, resources must accurately reflect the language they encompass. A challenge to this is that the Indigenous languages in Canada are morphologically very rich – some lemmas may have hundreds of different inflections – and yet the resources that are used to convey them are not always structured in such a way that makes incorporating complex morphological systems straightforward.

The *de facto* standard in bilingual lexicography is to compile lists in which words or phrases in the Indigenous language and the majority language are mapped to each other one-to-one (cf. Prinsloo 2012, Lew 2012, Granger 2012). Once digitized, these word or phrase lists are stored as lexical databases, which assume the entire word as the basic unit, informed as they are by the Indo-European majority language tradition, in which this assumption is unproblematic.

However, taking the phonological or orthographic word as a basic unit when working with languages with very complex morphological structure – such as the Indigenous languages of North and South America

© 2022 Elizabeth Pankratz, Antti Arppe and Jordan Lachler. *Nordlyd* 46.1: 193–204, *Morfologi, målstrev og maskinar*: *Trond Trosterud {fyller | täyttää | deavdá | turns} 60!*, edited by Lene Antonsen, Sjur Nørstebø Moshagen and Øystein A. Vangsnes. Published at UiT The Arctic University of Norway.
http://septentrio.uit.no/index.php/nordlyd https://doi.org/10.7557/12.6441

– means that lexicographers need to employ more creative strategies to still convey the language's morphology in a clear way. Complex inflectional morphology is usually captured using inflectional paradigms: tables that map certain combinations of morphosyntactic properties (that appear in the rows and columns) onto certain inflected forms (that appear in the cells; cf. Spencer 2016: 27; Stump 2016: 7–10). A full paradigm contains the entire set of inflected forms that are associated with (and are possible for) a particular lemma, i.e. the citation form, or base form, which has been selected to represent the entire lexeme, which is manifested by all these possible inflected forms. However, without great technical effort it is difficult to incorporate inflectional paradigms into a word-based resource – what counts as the head word? How are the various inflected forms linked together? How are multimorphemic inflected forms assembled in the first place? The ways in which online resources for morphologically complex Indigenous languages in Canada approach these and further questions will be explored in what follows.

The current contribution shares the results of a metalexicographical survey of existing online bilingual resources for Indigenous languages in Canada with a focus on how the resources reflect the morphological richness of their languages (Section 2). Based on the survey's results, we suggest that moving away from the traditional bilingual dictionary structure and moving toward using computational models of the languages' rich morphological systems is a desirable direction for lexicography of low-resource, morphologically-complex languages, and is a "low-hanging fruit" in the application of language technology (Section 3), the pursuit of which has been passionately advocated and recommended by Trond Trosterud, and which has been a great inspiration for the authors.

## 2 The metalexicographical survey

Our goal in conducting this survey was to determine how online resources for Indigenous languages in Canada convey information about the complex morphology and inflectional paradigm-formation of these languages. This question is interesting because, as we mentioned above, resources whose data is stored in lexical databases are inherently hard-pressed to do justice to a morphologically rich language, and ethnographic linguists and lexicographers must come up with alternative strategies to convey the language's morphology.

For illustration, consider the internal complexity of the Plains Cree word *niwâpamâw* in (1) (from Harrigan et al. 2017: 570):

(1) ni-wâpam-â-w
1SG.SBJ.INDP-see.VTA-THM-1SG.SBJ.3SG.OBJ[1]
'I see him/her (animate)'

A lexical database resource essentially has two choices when representing a wordform like this. It could take the unit *niwâpamâw* as the headword and equate it with the phrase 'I see him/her', relying on the user to deconstruct the inflectional morphology, or it could take the root *wâpam-* 'see' as the headword and rely on the speaker to build up the surrounding morphology. In either case, the user needs further information about how to deal with the morphemes required to use this word accurately. How do lexical database resources for Indigenous languages in Canada convey this information? And what other lexicographical infrastructures exist that sidestep this issue entirely? These questions will be explored in the rest of this section.

[1] Abbreviations from Harrigan et al. (2017: 568): 1SG.SBJ: first person singular subject; INDP: Independent order; VTA: Verb Animate Transitive; THM: Theme sign; 3SG.OBJ: third person singular object.

### *2.1 Method*

We began by taking the list of Indigenous languages recognized by Statistics Canada (2011) and searching in Google "[name of language] online dictionary".[2] In this way, we discovered the seventeen resources listed in Table 1.

These resources are all bilingual between the Indigenous language and English (and in a few cases, Canada's other official language, French). They range from simple HTML/XML-structured word lists to intricate websites that rival majority languages online dictionaries. Three of the websites – First Voices, Mother Tongues Dictionaries, and Ohwejagekha: Ha'degaenage – are multi-database interfaces that contain data from many Indigenous languages. This reduces the overhead of maintaining a website, making it simple for a community initiative to publish their language data. In the survey, we included the largest resource from each of these collections.

| **Language(s)** | **Resource name** |
|---|---|
| Algonquin | The Algonquin Way Dictionary |
| Dakota | Dakota Dictionary Online |
| East Cree | Eastern James Bay Cree Dictionary |
| Gitksan | Gitksan/English Online Dictionary (Beta) |
| Híłzaqv | Mother Tongues Dictionaries: Híłzaqv |
| Mi'kmaq | Mi'kmaq Online |
| Mohawk | Ohwejagekha: Ha'degaenage: Mohawk |
| Moose Cree, Swampy Cree | Spoken Cree |
| Northern St̓át̓imcets | First Voices: Northern St̓át̓imcets |
| Ojibwe | The Ojibwe People's Dictionary |
| Passamaquoddy-Maliseet | Passmaquoddy-Maliseet Dictionary |
| Plains Cree | itwêwina |
| Plains Cree | Online Cree Dictionary |
| SENĆOŦEN | SENĆOŦEN Classified Word List |
| Sm'algyax | Sm'algyax Living Legacy Talking Dictionary |
| Tłı̨chǫ Yatıì | Tłı̨chǫ Yatıì Multimedia Dictionary |
| Tlingit | Online Tlingit Verb Dictionary |

Table 1. The online resources for Canadian Indigenous languages included in the 2019 survey.

We originally evaluated each dictionary according to a broad set of criteria about how electronic lexicography can improve on the print dictionary tradition (primarily developed from Lew 2012, Granger 2012, and Prinsloo 2012). Many of these criteria exceed the scope of the current paper. For example, we evaluated whether the dictionaries incorporate multimedia elements and what sorts of extralexicographical (e.g. cultural, pedagogical) information is also available, and we recorded metrics like how many clicks it takes to get from the resource's front page to a page with information about the desired word (see Lachler and Pankratz 2017 for some of these results). Concerning our current goal, the portrayal of morphological structure, the primary desideratum from the literature was that online resources take advantage of technology to incorporate "more and better information" (Granger 2012: 2). In the next section, we will discuss the wide range in how more and better morphological information is represented in online resources for Indigenous languages in Canada.

[2] This method will only deliver web pages that have been indexed by Google, so some smaller and less-accessed resources may have eluded us for this reason. However, we assume that the websites that are indexed are the ones oriented toward interested users who might search for information about the language, and these are the resources that we want to explore.

### 2.2 *Results: The emergence of two lexicographical approaches*

In the seventeen resources we surveyed, we observed three main strategies for conveying morphological information, summarized in Table 2. The resources can be grouped according to which of these (sub)strategies they use. The number of resources using each strategy is shown in Table 2, and in Table 3, we have grouped the resources according to the strategies from Table 2 that they employ.

The surveyed resources have two possible structures, as shown at the right of Table 2: sixteen resources appear to be built around lexical databases, and one uses a computational model of the language's morphology (which is nevertheless based on a lexical database). These two approaches will be discussed more deeply in Section 3, but for now, we will go into more detail about the three groups of strategies. A common thread connecting most of these resources is that individual orthographic word-forms have their own page on the website, which will be referred to as the "lexical entry" or simply "entry" for that word.[3]

| | Strategy | Number | Structure |
|---|---|---|---|
| 1. | The resource contains no morphological information. | 4 | lexical database |
| 2. | Each inflected wordform has its own entry. | 2 | |
| 3. | Each paradigm or root has its own entry. | 11 | |
| | 3a. Morphemes also have their own entry. | 1 | |
| | 3b. Multiple complex wordforms provided, no gloss. | 2 | |
| | 3c. Multiple complex wordforms provided, glossed. | 3 | |
| | 3d. Generalized morphological slot templates provided. | 1 | |
| | 3e. Individual morphological slot templates in each entry. | 1 | |
| | 3f. Generalized full paradigms provided. | 2 | |
| | 3g. Individual full individual paradigms in each entry. | 1 | computational model |

Table 2. The strategies for portraying morphological information and the number of resources we surveyed that used each of them.

| Group | Resource name |
|---|---|
| 1 | Gitksan/English Online Dictionary (Beta)<br>Ohwejagekha: Ha'degaenage: Mohawk<br>Online Cree Dictionary<br>The Algonquin Way Dictionary |
| 2 | SENĆOŦEN Classified Word List<br>Tłı̨chǫ Yatıı̀ Multimedia Dictionary |
| 3a | First Voices: Northern St̓át̓imcets |
| 3b | Dakota Dictionary Online (public version)[4]<br>Spoken Cree |
| 3c | Mother Tongues Dictionaries: Híɫzaqv<br>Mi'kmaq Online<br>The Ojibwe People's Dictionary |
| 3d | Sm'algyax Living Legacy Talking Dictionary |
| 3e | Online Tlingit Verb Dictionary |
| 3f | Eastern James Bay Cree Dictionary<br>Passamaquoddy-Maliseet Dictionary |
| 3g | itwêwina |

Table 3. The surveyed resources listed according to how they express the morphology in their language (group numbers from the strategies listed in Table 2).

[3] The term "dictionary entry" often used for the same purpose, but since we are looking at resources that might not be considered dictionaries in the strictest sense, we have opted for "lexical entry" here.

[4] Lemma linking based on linguistic analysis is available for registered users of this resource, but we have evaluated the freely, publicly available version for the current survey.

### *2.2.1 Group 1: The resource contains no morphological information*

Resources were grouped together here if only a headword in the Indigenous language and an English or French translation are provided, with no information about any morphological structure. As Table 3 shows, four of the seventeen surveyed resources fall into this group. That said, two of these only advertise themselves as word lists or phrase lists, so holding them to a higher lexicographical standard may be unfair. In any case, such lists are the bones of a lexicographical resource and a valuable first step in language documentation (cf. Bowern 2015).

### *2.2.2 Group 2: Each inflected wordform has its own entry*

The two resources belonging to this group contain morphological information to the extent that the resource does contain multiple documented wordforms from the same lemma and sometimes an entire paradigm's worth of wordforms, but they are not conceptually grouped together. Rather, they are all listed at the same level as everything else, each as a single entry in the resource. This is a step forward over Group 1 because the resource reflects some of the wordforms possible in the language, though there is room for more abstraction away from the surface and toward the underlying paradigms.

### *2.1.3 Group 3: Each paradigm or root has its own entry*

Like Group 2 resources, the ones in this group successfully include several wordforms, but beyond Group 2, these resources abstract away from the multiple orthographic words to their belonging to the same lemma/lexeme, which is what ties them all together. This step of abstraction is performed to different degrees of extensiveness and generalization, which is why this group has been divided into seven subgroups.

Group 3a resources recognize the morphemes as separate entities and list them parallel to the stems in the database (in one case, without information about how stems and morphemes can be combined to create new wordforms). The next two groups, 3b and 3c, have in common that a lexical entry may contain other wordforms which are morphologically related to the headword. What differentiates the groups is that in Group 3b resources, these wordforms are given without glossing or translation, while Group 3c resources also explain the morphological characteristics of the additional wordforms. For example, in Group 3c a noun's lexical entry might list another form as the plural, or a verb's lexical entry might provide and gloss that verb's first person singular form.

What Groups 3d and 3e share is one more step toward abstraction, namely that in both subgroups, morphological templates are provided (i.e. abstract schemas listing the order in which morphemes and stems appear in a morphologically-complex word). The templates in Group 3d resources are generalized versions which are valid for particular inflection classes. The templates given in Group 3e resources are not applicable to wider classes of stems, but given individually in each lexical entry, with the headword already integrated into the template itself.

At this point, as long as information about the morphemes themselves is provided, we begin to arrive at the resources which make it possible to construct a correctly-inflected word only from the information in the resource itself. However, this morphological information is not always there. When it is available, it is often in the form of lists of prefixes, verbal themes, etc., requiring quite in-depth linguistic knowledge to figure out how it should all be combined. All in all, the Group 3d and 3e resources we surveyed are linguistically very detailed, but might not be as accessible to community members some without linguistic training.

Finally, the last two groups are those which combine linguistic informativeness with layperson-oriented accessibility. Groups 3f and 3g include those resources which offer full inflectional paradigms (and enough information that the user can quickly find the paradigm for the word class they are interested in). The division between these is parallel to the division between the two groups of template-based resources above. Group 3f provides generalized paradigms which are representative of a particular word class, while every lexical entry in a Group 3g resource contains its own individual full paradigm with all core inflectional forms (though more derivational affixation have been left out).

As Table 2 shows, Group 3f is the last group where we observe a resource constructed around a lexical database. With this type of structure, it would not be easy to include full paradigms for each individual root contained in the resource, especially because so much of the stored information would be redundant, appearing over and over again in different but related paradigms. This is where the computational model of the language's morphology found in Group 3g shows its strengths and value for providing quick access to detailed linguistic information.

### *2.2.4 The take-away*

In sum, the traditional conceptual layout of a bilingual dictionary – linking a single headword in one language to an equivalent in the other – is maintained in most of the resources we surveyed. This is unsurprising, since online language resources built around lexical databases are the *de facto* standard (cf. Prinsloo 2012; Lew 2012; Granger 2012), and not without reason. Descriptive linguists often begin by collecting lists of inflected wordforms and phrases before delving more deeply into the morphology and subtler linguistic structures (cf. Bowern 2015). The software that these linguists use – Toolbox and the like – digitizes these word lists in lexical databases, enforcing the one-to-one mapping between words in each language.

Even though this database structure is limited in how it can express morphological richness, we have observed certain strategies by which the surveyed resources nevertheless convey morphological information.

In what follows, we do not wish to downplay the hard work that goes into making any lexicographical resource, whether built around a lexical database or a computational model. However, we would like to argue that one of the most promising benefits that the digital domain offers to language resources is that of complete integration, especially for low-resource or endangered languages (cf. Prinsloo 2012: 127), and that this direction should be pursued further in the lexicography of Indigenous languages in Canada and elsewhere.

## 3 Discussion: Reinterpreting the Boasian trilogy

Our metalexicographical survey has shown that language resources constructed around a lexical database do their best to overcome the limitations of this format by employing strategies like providing morphological templates or linking the user to relevant paradigms containing the core inflectional forms that the user may want. These are good measures for creating valuable linguistic resources. However, any lexical database resource keeps the grammar outside of the core of the resource by design, no matter how detailed the morphological information is that it provides. Here, we would like to follow Prinsloo (2012) and argue that a digital lexicographical resource should be an all-in-one, integrated tool with all the lexicographic information that any user group should need. This is a demanding goal that cannot be achieved without rethinking lexicographical and documentary traditions and trying something new, adapting long-held methodological cornerstones to meet modern needs.

One such cornerstone in documentary linguistics is the production of a grammar, a dictionary, and texts in a language: the 'Boasian trilogy' (cf. Rice 2011: 192–193; Woodbury 2011: 163), so called based on Boas' (1917: 1) observation that in Indigenous language research "[w]e have vocabularies; but, excepting the old missionary grammars, there is very little systematic work. Even when we have grammars, we have no bodies of aboriginal texts."

Having this set of resources is valuable not just for linguistic research, but also for language documentation and revitalization in the community. The very name makes it clear that there are three distinct parts, but the three-way partition can be integrated into one resource using digital technology. We saw this in Group 3g in our survey, which contains a resource for Plains Cree named *itwêwina* (literally the plural noun form 'words' in Plains Cree). We only discuss the morphological integration aspect of *itwêwina* here, but it is indeed outfitted with a corpus of Plains Cree texts as well, completing the 'trilogy'. We must also concede here that we are leaving out in this discussion syntactic knowledge as an important part of grammar, though one has to recognize that in such morphologically rich and complex languages the morphosyntactic component, i.e. inflected wordforms, is the one undertaking much of the "heavy lifting" in grammar.

*itwêwina* (https://itwewina.altlab.app) is built around a set of finite state transducers (FSTs, cf. Beesley and Karttunen 2003) which map between the orthographic wordform and its abstract morphological and morphosyntactic features and root. This process works in either direction, either encoding (producing an inflected form of a lemma based on morphological and morphosyntactic features) or decoding (parsing an already-inflected word; cf. Johnson et al. 2013: 62 for North and South; Harrigan et al. 2017 for Plains Cree verbs; Snoek et al. 2014 for Plains Cree nouns). In an FST-based online dictionary, the user can search with a fully-inflected orthographic word, and the computational model will deconstruct it into its stem and affixes, provide information about the morphemes, and direct the user to the appropriate entry in the dictionary. (This may be possible for lexical database resources too, but none of the ones we surveyed had this capability.)

Resources like *itwêwina* that can morphologically analyze and decompose complex words are a step toward a new state of the art in digital lexicography of Indigenous languages. Other languages in the world have resources which are also making this step. Turning to the Bantu languages of Africa, we find the website *isiZulu.net*, the most sophisticated resource available for a Bantu language (Prinsloo 2012: 135). Like *itwêwina*, the user can enter a fully-inflected orthographic word in the search bar, and the website returns the dictionary entry for the root of that word, as well as information about the original word both as linguistic glosses and as plain-speech translations, which makes the resource viable for communities without specialist linguistic knowledge. In the United States, Arapaho also has a resource built around finite state machines (Kazeminejad et al. 2017). And the same FST framework used by *itwêwina*, *Neahttadigisánit* (literally 'internet digital words' in North Saami), was first developed for the Saami languages of Northern Scandinavia by Giellatekno at UiT Arctic University of Norway (Tromsø), founded and directed by Trond Trosterud (cf. Johnson et al. 2013), and then brought to Canada for use with the morphologically complex languages there. *itwêwina* is the best-developed resource with this technology in Canada, but comparable resources for Northern Haida, Tsuut'ina, East Cree, and Odawa are currently in development (cf. Lachler et al. 2018; Arppe et al. 2017a; Arppe et al. 2017b; Bowers et al. 2017).

Indeed, this extensibility is one of the advantages of such a computational model. Adapting an existing model to new but typologically related languages is fairly straight-forward (if the linguistic documentation is there), much easier than compiling an entirely new lexical database from scratch, offering much smaller languages the same chance at excellent linguistic resources.

Another advantage of computational-model-based dictionaries is how much of the language they can encompass; they are arguably "the only way of ensuring good coverage of the language" (Trosterud 2004: 92). This is because adding one new root into an FST-based system results in the ability to decode any number of inflected wordforms built around that root. In contrast, adding one word into a lexical database dictionary results in the ability to decode exactly that one word more (Johnson et al. 2013: 69). In languages like Plains Cree and others in the Algonquian family, where the number of forms in a paradigm is in principle infinite (because one can add as many preverbs as one likes, though even without these, the paradigms of certain verb classes can be hundreds large), it is simply not feasible to try to contain all or even most words in a lexical database by adding them one by one. Assembling the computational morphological model is a great deal of work at the beginning, though surprisingly straight-forward provided that the linguistic description has been meticulous in determining morphological behavior of words, but it pays off immensely over time. If such extensive and comprehensive linguistic description already exists, as was the case for Plains Cree with the lexical database underlying the bilingual Cree-English dictionary, *nêhiyawêwin : itwêwina / Cree : Words*, diligently compiled by Wolvengrey (2001), then developing a computational model based on such a database and next expanding the database with that model turned out to be a veritable "low-hanging fruit", as Trond Trosterud has pitched it, in the application of language technology.

Above, we encountered the issue that sophisticated linguistic knowledge would sometimes be required to build correctly- inflected wordforms based only on the information given in the resource. This is another difficulty that is overcome by the computational model, because it is possible to display the information from one underlying model in many different ways. For example, *itwêwina* has linguistic descriptions of the paradigms ("first person singular actor → third person singular goal", "Present tense"), as well as layperson-oriented ones in plain English (e.g. "I → him/her", "Something is happening now"), and even in "plain" nêhiyawêwin, i.e. Cree (e.g. "niya → kiya", "ê-ispayik anohc/mêkwâc/mâna"). This helps overcome the linguistic terminology barrier and makes the resource much more accessible to the community and non-specialists in general.

Another benefit of generating full paradigms is that the language user in search of a word expressing particular morphological characteristics on a particular stem will be able to find one, even if it has not been attested in the existing documentary corpus for that language. This allows the user to express the specific meaning they intend in much the same way that native speakers do, relying on the existing morphological patterns in the language.

Also, having such a model means that one can use it not just locally in an online dictionary, but in other tools as well. For example, Plains Cree also has a spell-checker under development (cf. Arppe et al. 2016: 5–6), a computer-aided language learning resource called *nêhiyawêtân* (literally 'Let's speak Cree'; cf. Bontogon et al. 2018), and a reader plugin, which allows the user to click on a word in Plains Cree appearing on any website and see a pop-up providing the lemma resulting from the linguistic analysis of the word and the English translation of the lemma (cf. Johnson et al. 2013: 65).

Arguably one of the most crucial advantages of such a resource, especially for communities working on language revitalization, is its usefulness to learners. Encoding and decoding morphologically complex wordforms requires extensive morphophonological knowledge, which a learner is unlikely to have, at least at the beginning. A computational model that can encode and decode complex wordforms does the worst of the work for the learner, so they have more energy left over to learn. As time goes on and the motivated learner's competence increases, they will need to consult the resource less and less, so it is particularly valuable for the early stages of language acquisition that could otherwise be overwhelming.

Finally, we could imagine that Table 2's list of strategies for conveying morphological information extends even further into more complex computational domains like deep/machine learning applied on (large) text collections. However, we would argue that the FST-based method actually has an advantage over deep learning approaches, because it does not need as much data to create a successful model. As conventional wisdom has it, the amount of data necessary for tried-and-true computational learning methods would be at least ten million words.[5] This amount of data is very unlikely to be available for low-resource languages, but an FST can be created without a substantial corpus. Therefore, FSTs have been practically the only way to create an extensive computational resource for these languages (Trosterud 2004: 92). Intriguingly, it has recently been shown that a neural encoder-decoder model can learn from comprehensive collections of morphological paradigms to mimic the behaviour of a handmade FST; cf. Moeller et al. (2018).

All in all, there are many lexicographical, documentary, and pedagogical advantages to having a computational model of a low-resource language's morphology. However, some issues must also be considered before adopting this type of model, in particular the question of how important it is that all wordforms contained in the resource be attested, verified forms in the language. We will focus on this issue for the remainder of this section.

### *3.1 Documentary and computational considerations*

A generative morphological tool like the one incorporated within *itwêwina* raises a principled question about generalization and overgeneralization of paradigms: Is it better for a lexical resource to represent only words that have been verified, upholding native speakers' intuitions about possibilities and restrictions in their language, or to cover as much of the language as possible in an urgent language endangerment situation?

We may understand these two positions in parallel to the two lexicographical approaches outlined above. Lexical databases contain those words which are attested through fieldwork or consultation with the language community, while a computational model produces maximum possible coverage of the language,

[5] Mans Hulden, personal communication, for the case of German morphology.

even at the cost of overgenerating forms that speakers may not actually use, which has been described as the assumption of "lexical generality" (Bybee 1985: 84–87; Arppe et al. 2000: 5). And different sorts of linguists will prefer one approach or the other: A primarily computational linguist accepts the possibility of mistakes or overgeneralized forms creeping into the resource, while a primarily documentary linguist prefers to check and test everything, letting only gold-standard data through, although it only represents a fraction of the language.

Different communities (and subgroups within these communities) will also likely react to this question very differently. In our experience, language communities at the extremes of endangerment – those with very few elders or native speakers left, *and* those with millions of speakers – both tend to be less bothered by possible issues of morphological overgeneration. For very small endangered languages, keeping the language generally alive is the greatest priority, even if some words that nobody actually uses arise in the process, and for majority languages, native speakers and fluent learners know how the language is used anyway, so a few extra forms are not an issue. Crucially, even the potentially overgenerated word-forms are created according to common morphological processes and regular word-formation strategies observed in the language in question, and allow for as close an approximation as possible as to what a particular form could be, even if it has not been attested (rather than not being able to utter that form, or have to use a periphrastic construction, or need to use a majority language, to express some set of linguistic features).

However, toward the middle of the endangerment spectrum, priorities may be different and reactions may be more mixed. Some communities or community members might object to the fact that not every word has been checked, while others might appreciate how much of the language such a model can generate. And other community factors also come into play, such as the total number of speakers, their degree of language proficiency, whether a standardized form of the language exists, what kind of language education is available, and so on (cf. Arppe et al. 2016: 3). The decision about which of the two drawbacks – an overproductive resource or not enough coverage of the language – is the lesser of two evils should be done together with the community and in consideration of the other factors at play.

Overall, the two approaches must inform one another. A purely computational resource without actual verified data, for example, is no good to anybody. But we believe that the computational approach is particularly useful for time-sensitive endangered language situations, especially with languages like those in Canada, where it is prohibitively infeasible to check every form of every paradigm. Do we wait until we have a completely finished, perfected model which exactly reflects speaker intuitions, and until then have nothing (never mind that no lexicographical projects are ever really finished)? Or do we accept that it will be imperfect and will make mistakes around idiosyncrasies and other areas of variation? We have argued in this paper for the value of the second approach, especially given its flexibility: a computational model can generate all possible wordforms and then we can mark out those that have been attested, to distinguish them from the ones that are purely the creations of the model. This validation can be done steadily over time to increase the set of attested wordforms, allowing the two methods to meet in the middle. This approach lets us work in a time-sensitive way and simultaneously do justice to the morphological complexity of Indigenous languages in Canada.

## 4 Conclusion

In our survey of seventeen online resources for Indigenous languages in Canada, we found that these resources pursue different strategies to convey their language's complex morphology. These strategies are applied to get around the inherent limitations of resources built around a lexical database for expressing the complex morphological (sub-word) structures that are the norm in Canadian Indigenous languages. We have shown that one resource unites aspects of the grammar and the lexicon and more accurately reflects the language's morphology by incorporating a computational morphological model.

We have also argued that the computational approach represents a step forward in the evolution of lexicography, especially for low-resource languages and their communities. For these communities, the reinterpretation of the Boasian trilogy and the print dictionary tradition means the creation of an integrated resource, one that contains enough information that speakers and learners can use it to correctly construct the complex words in their language. The ability to use the language correctly and fully, and to discover these forms easily, can be a significant boon to revitalization efforts.

Furthermore, we argue that a computational morphological model allows us as linguists to give more back to the community than do standard lexical database dictionaries (cf. also Lachler and Pankratz 2017). Such a model can be used to develop diverse language technologies like parsers, spell-checkers, and language learning applications, which are all ways to use technology to place the language back into the hands of the community and bring it back into day-to-day life. Language resources that sit on shelves (or hard-drives), safely archived, may help linguistic research, but do not help language revitalization in the communities who helped create the resources in the first place. The incorporation of computational models into digital lexicography for low-resource languages is a way to take decades of work that has shaped and redefined much of linguistic theory and turn it into something useful for both linguists and the community – this often really is a low-hanging fruit in the application of language technology, ripe for the picking.

In short, we cannot forget that language documentation and revitalization is a race against time, and tools like computational models that encompass the morphological richness of these languages help us help these communities, do it well, and do it quickly.

## References

Arppe, Antti, Jordan Lachler, Trond Trosterud, Lene Antonsen, and Sjur N. Moshagen. 2016. Basic Language Resource Kits for Endangered Languages: A Case Study of Plains Cree. *Proceedings of CCURL 2016 – Collaboration and Computing for Under-Resourced Languages.* 1–8.

Arppe, Antti, Mari Voipio, and Malene Würtz. 2000. Creating Inflecting Electronic Dictionaries. Edited by Carl-Erik Lindberg and Steffen Nordahl Lund. *Proceedings of the 17th Scandinavian Conference of Linguistics, Nyborg*. 20–22.

Arppe, Antti, Christopher Cox, Mans Hulden, Jordan Lachler, Sjur N. Moshagen, Miikka Silfverberg and Trond Trosterud. 2017a. Computational Modeling of Verbs in Dene Languages: The Case of Tsuut'ina. In Alessandro Jaker (ed.), *Working Papers in Athabaskan (Dene) Languages,* 51—69. Alaska Native Language Center Working Papers 13. Alaska Native Language Center, Fairbanks.

Arppe, Antti, Marie-Odile Junker, and Delasie Torkornoo. 2017b. Converting a comprehensive lexical database into a computational model: The case of East Cree verb inflection. *Proceedings of the 2nd Workshop on Computational Methods for Endangered Languages (ComputEL-2).* 43–47. https://doi.org/10.18653/v1/W17-0108.

Beesley, Kenneth R., and Lauri Karttunen. 2003. *Finite state morphology.* Studies in Computational Linguistics. Center for the Study of Language and Information, Stanford, California.

Boas, Franz. 1917. Introductory. *International Journal of American Linguistics* 1(1): 1–8. https://doi.org/10.1086/463708.

Bontogon, Megan, Antti Arppe, Lene Antonsen, Dorothy Thunder, and Jordan Lachler. 2018. Intelligent Computer Assisted Language Learning (ICALL) for *nêhiyawêwin*: An In-Depth User-Experience Evaluation. *Canadian Modern Language Review* 74: 337–362. https://doi.org/10.3138/cmlr.4054.

Bowern, Claire. 2015. *Linguistic fieldwork: A practical guide.* Springer. https://doi.org/10.1057/9781137340801.

Bowers, Dustin, Antti Arppe, Jordan Lachler, Sjur N. Moshagen, and Trond Trosterud. 2017. A Morphological Parser for Odawa. *Proceedings of the 2nd Workshop on Computational Methods for Endangered Languages (ComputEL-2).* 1–9. https://doi.org/10.18653/v1/W17-0101.

Bybee, Joan. 1985. *Morphology: A study of the relation between meaning and form.* (Vol. 9, Typological studies in language). John Benjamins, Amsterdam. https://doi.org/10.1075/tsl.9.

Granger, Sylvaine. 2012. Introduction: Electronic lexicography – from challenge to opportunity. In *Electronic Lexicography*, edited by Sylvaine Granger and Magali Paquot, 1–12. Oxford University Press, Oxford. https://doi.org/10.1093/acprof:oso/9780199654864.003.0001.

Harrigan, Atticus G., Katherine Schmirler, Antti Arppe, Lene Antonsen, Trond Trosterud, and Arok Wolvengrey. 2017. Learning from the computational modelling of Plains Cree verbs. *Morphology* 27: 565–598. https://doi.org/10.1007/s11525-017-9315-x.

Johnson, Ryan, Lene Antonsen, and Trond Trosterud. 2013. Using Finite State Transducers for Making Efficient Reading Comprehension Dictionaries. *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013).* 59–71.

Kazeminejad, Ghazaleh, Andrew Cowell, and Mans Hulden. 2017. Creating lexical resources for polysynthetic languages – the case of Arapaho. *Proceedings of the 2nd Workshop on Computational Methods for Endangered Languages (ComputEL-2).* 10–18. https://doi.org/10.18653/v1/W17-0102.

Lachler, Jordan, and Elizabeth Pankratz. 2017. Moving toward value-added digital repatriation in lexicography for Indigenous languages in Canada. Edited by Nicholas Ostler, Vera Ferreira and Chris Moseley. *Proceedings of the 21st FEL Conference: Communities in Control: Learning tools and strategies for multilingual endangered language communities.* 107–114.

Lachler, Jordan, Lene Antonsen, Trond Trosterud, Sjur N. Moshagen, and Antti Arppe. 2018. Modeling Northern Haida Morphology. *Proceedings of the 11th Language Resources and Evaluation Conference (LREC 2018)*, Miyazaki, Japan, May 7–12, 2018, 2326–2330.

Lew, Robert. 2012. How Can We Make Electronic Dictionaries More Effective? In *Electronic Lexicography*, edited by Sylvaine Granger, and Magali Paquot, 343–361. Oxford University Press, Oxford. https://doi.org/10.1093/acprof:oso/9780199654864.003.0016.

Moeller, Sarah, Ghazaleh Kazeminejad, Andrew Cowell, and Mans Hulden. 2018. A neural morphological analyzer for Arapaho verbs learned from a finite state transducer. *Proceedings of Workshop on Polysynthetic Languages.* Association for Computational Linguistics, 12–20. https://aclanthology.org/W18-4802.

Prinsloo, D. J. 2012. Electronic lexicography for lesser-resourced languages: The South African context. In *Electronic Lexicography*, edited by Sylvaine Granger, and Magali Paquot, 119–143. Oxford University Press, Oxford. https://doi.org/10.1093/acprof:oso/9780199654864.003.0007.

Rice, Karen. 2011. Documentary linguistics and community relations. *Language Documentation & Conservation* 5: 187–207.

Statistics Canada. 2011. Aboriginal languages in Canada. *2011 Census of Population, Catalogue no. 98-314-X2011003.* https://www12.statcan.gc.ca/census-recensement/2011/as-sa/98-314-x/98-314-x2011003_3-eng.pdf (accessed April 19, 2019).

Snoek, Conor, Dorothy Thunder, Kaidi Lõo, Antti Arppe, Jordan Lachler, Sjur Moshagen, and Trond Trosterud. 2014. Modeling the Noun Morphology of Plains Cree. *Proceedings of ComputEL: Workshop on the use of computational methods in the study of endangered languages*, 52nd Annual Meeting of the Association for Computational Linguistics, Baltimore, Maryland, 26 June 2014, 34–42. ACL Anthology. https://doi.org/10.3115/v1/W14-2205.

Spencer, Andrew. 2016. Two morphologies or one? Inflection versus word formation. In *The Cambridge Handbook of Morphology*, edited by Andrew Hippisley and Gregory Stump, 27–49. Cambridge University Press, Cambridge. https://doi.org/10.1017/9781139814720.002.

Stump, Gregory. 2016. *Inflectional paradigms: Content and form at the syntax-morphology interface.* Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9781316105290.

Trosterud, Trond. 2004. Porting morphological analysis and disambiguation to new languages. *First Steps in Language Documentation for Minority Languages: Proceedings of the SALTMIL Workshop at LREC 2004.* 90–92.

Trosterud, Trond. 2006. Grammatically based language technology for minority languages. *Lesser-Known languages of South Asia: Status and policies, case studies and applications of information technology.* Mouton de Gruyter, Berlin. 293–316. https://doi.org/10.1515/9783110197785.

Wolvengrey, Arok. 2001. *nêhiyawêwin: itwêwina / Cree: Words*, bilingual edition. University of Regina Press, Regina.

Woodbury, Anthony C. 2011. Language documentation. In *The Cambridge Handbook of Endangered Languages*, edited by Peter K. Austin and Julia Sallabank, 159–186. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9780511975981.009.

## Language resource references

Eggleston, Keri. *Online Tlingit Verb Dictionary*. http://ankn.uaf.edu/~tlingitverbs/ (accessed April 19, 2019).

Ellis, Doug. *Spoken Cree: Moose and Swampy Cree Dictionary.* http://www.spokencree.org/Glossary (accessed April 19, 2019).

First Voices. *Northern St̓át̓imcets.* http://www.firstvoices.com/en/Northern-Statimcets (accessed April 8, 2019).

isiZulu.net. *isiZulu.net: Bilingual Zulu-English dictionary.* https://isizulu.net/ (accessed April 19, 2019).

Junker, Marie-Odile, Marguerite MacKenzie, Luci Bobbish-Salt, Alice Duff, Linda Visitor, Ruth Salt, Anna Blacksmith, Patricia Diamond, and Pearl Weistche. *The Eastern James Bay Cree Dictionary on the Web: English-Cree and Cree-English, French-Cree and Cree-French (Northern and Southern dialects).* http://dictionary.eastcree.org/words (accessed April 19, 2019).

Mi'kmaq Online. *Mi'kmaq Online.* https://www.mikmaqonline.org/ (accessed April 19, 2019).

Miyo Wahkohtowin Community Education Authority. *Online Cree Dictionary.* http://www.creedictionary.com/ (accessed April 19, 2019).

Mother Tongues Dictionaries. *Online Híłzaqv Dictionary.* https://mothertongues.org/heiltsuk/ (accessed April 19, 2019).

Ohwejagekhá: Ha'degaenage. *Mohawk.* http://ohwejagehka.com/mohawk/ (accessed April 19, 2019).

Omàmiwininì Pimàdjwowin/The Algonquin Way Cultural Centre. *The Algonquin Way Dictionary.* http://www.thealgonquinway.ca/English/dictionary-e.php (accessed April 19, 2019).

Passamaquoddy-Maliseet Language Portal. *Passamaquoddy-Maliseet Dictionary.* https://pmportal.org/browse-dictionary (accessed April 19, 2019).

SENĆOŦEN Classified Word List. *SENĆOŦEN Classified Word List.* https://itservices.cas.unt.edu/~montler/Saanich/WordList/ (accessed April 19, 2019).

The University of Northern British Columbia. *Sm'algyax Living Legacy Talking Dictionary.* http://web.unbc.ca/~smalgyax/ (accessed April 19, 2019).

University of Alberta ALTLab. *itwêwina.* https://itwewina.altlab.app/ (accessed April 19, 2019).

University of British Columbia Department of Linguistics. *Gitksan/English Online Dictionary (Beta).* http://gitdict.nfshost.com (accessed April 19, 2019).

University of Minnesota Department of American Indian Studies. *Dakota Dictionary Online.* https://filemaker.cla.umn.edu/dakota/ (accessed April 19, 2019).

University of Minnesota Department of American Indian Studies. *The Ojibwe People's Dictionary.* https://ojibwe.lib.umn.edu/ (accessed April 19, 2019).

University of Victoria Linguistics Department. *Tłı̨chǫ Yatıı̀ Multimedia Dictionary.* http://tlicho.ling.uvic.ca/ (accessed April 19, 2019).