7-2023

Identification of Dialect for Eastern and Southwestern Ojibwe Words Using a Small Corpus

Kalvin Hartwig
Independent Researcher

Evan Lucas
Michigan Technological University, eglucas@mtu.edu

Timothy C. Havens
Michigan Technological University, thavens@mtu.edu

Follow this and additional works at: https://digitalcommons.mtu.edu/michigantech-p2
Part of the Computer Sciences Commons

Identification of Dialect for Eastern and Southwestern Ojibwe Words Using a Small Corpus

Kalvin Hartwig,
Independent Researcher
Bayfield, Wisconsin, USA
kalvin.hartwig@aya.yale.edu

Evan Lucas and **Timothy C. Havens**
Michigan Technological University
Houghton, Michigan, USA
{eglucas, thavens}@mtu.edu

Abstract

The Ojibwe language has several dialects that vary to some degree in both spoken and written form. We present a method of using support vector machines to classify two different dialects (Eastern and Southwestern Ojibwe) using a very small corpus of text. Classification accuracy at the sentence level is 90% across a five-fold cross validation and 72% when the sentence-trained model is applied to a data set of individual words. Our code and the word level data set are released openly at https://github.com/evan-person/OjibweDialect.

1 Introduction

The Ojibwe language is an Indigenous language of the Great Lakes region of Turtle Island (North America) and is also known by many other names such as Chippewa, Ojibwemowin, Anishinaabe, and Anishinaabemowin. Anishinaabemowin can also refer to the closely related tongues Potawatomi, Algonquin and Odawa. An Algonquian language, Ojibwe and its many sub-dialects can be mapped geographically. Though traditionally understood to be a prestigious language spoken by several Peoples trading or living with/near the Ojibwe, currently, Ojibwe is mostly spoken by Ojibwe people. While many Ojibwe live on reservations and reserves of sovereign Ojibwe Tribes/First Nations across *Anishinaabewaki*, Anishinaabe country, many also live in towns and cities outside of reservations and reserves.

The number of native-level fluent speakers is unfortunately fast dwindling. It is estimated that there are around 50 native-level fluent speakers living today south of the Medicine Line (the American-Canadian border), virtually all of whom are Elders (Burnette, 2023). Most of these 50 older speakers are living on two reservations. There are at least 10,000 fluent speakers north of the Medicine Line, many of whom are also older (Pangowish, 2023). The approximate distribution of current Anishinaabemowin speakers is shown in Figure 1. According to the U.S. Census Bureau (2011) and Canadian Encyclopedia (Bishop, 2019), there are about 330,000 Ojibwe living in *Anishinaabewaki*, conservatively making around 3% of the Ojibwe population fluent speakers.

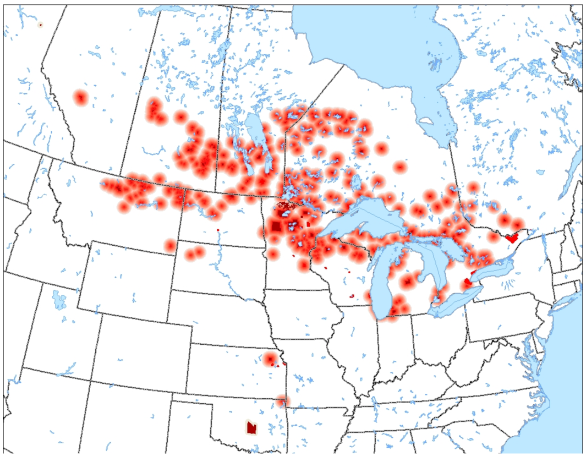

Figure 1: 2007 distribution of Anishinaabemowin speakers on Turtle Island, including Ojibwe and sister languages such as Potawatomi and Algonquin. From Lippert (2007)

There are many efforts in place to try and stem Ojibwe's decline. Passing along the knowledge and practice of speaking the Ojibwe language is an important part of maintaining Ojibwe culture; language is identified as one of the four pillars of Indigenous Peoplehood by Holm et al. (2003) along with ceremonies, land, and sacred history. Some say Ojibwe identity itself is at risk if the language is no longer spoken (Hartwig, 2012; McInnes, 2014) and therefore Ojibwemowin revitalization is one of the highest priorities for many language warriors. Language courses, as well as immersion and spending time with Elders, have been traditional tools of revitalizing Anishinaabemowin (Pitawanakwat, 2018). Over the past couple decades, there have also been efforts to use more media technology as

Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP), pages 58–66
July 14, 2023 ©2023 Association for Computational Linguistics

a tool for revitalizing the language (Hermes and King, 2013). Our work builds on this history.

The rest of this paper is organized as follows. Section 2 is split into three sub-sections: author and paper backgrounds are briefly reviewed in Section 2.1, a brief overview of Ojibwe dialects is covered in Section 2.2, and a review of related work is presented in Section 2.3. Methodology used and discussion of the written corpus used is presented in Section 3. A discussion of results is included in Section 4. We summarize our work in Section 5. We discuss the limitations of this work in the Limitations section and share some of the ethical concerns raised by this work in the Ethical Statement. Finally, we recognize and give thanks to the people who made this work possible in the Acknowledgements.

2 Background

2.1 Positionality

This work was performed as a collaboration between two of the authors (Hartwig and Lucas) as an exploratory work looking at ways that *Natural Language Processing* (NLP) tools could be used to aid Ojibwe language learners. The first discussions held between authors tried to address connections between the needs and interests of people working in Ojibwe language education and capabilities of NLP methods that work with limited text and varied dialects. To help capture background of the authors, we have included the following positionality statements:

- Mishkwaa-desi Kalvin Hartwig is a Member of the Sault Sainte Marie Tribe of Chippewa Indians. He serves as an independent filmmaker as well as the Anishinaabemowin gikinoo'amaagewin weninang / Anishinaabe language-culture coordinator for the Red Cliff Band of Lake Superior Chippewa Indians. He is not fluent in Ojibwe, but has been actively learning it.
- Evan Lucas is a White American who works as a graduate student studying NLP.
- Timothy Havens is a White American professor of computer science with research interests in challenging AI problems.

2.2 Ojibwe dialects

Across the geographic range shown in Figure 1, there are several dialects found (Valentine, 1994; Rhodes, 2006). Teachers and speakers of other dialects often consider Eastern Ojibwe and Odawa to be either one in the same or at least very similar. For this paper, we use “Eastern" to refer to Eastern Ojibwe, Odawa, or both. We decided to compare Eastern to the dialect of Southwestern Ojibwe; each having differences in spelling, some grammar rules, and sometimes morphological word construction (Nichols, 1980; Valentine, 2001a). A reader with familiarity in one dialect, but not another, may be unfamiliar with some of the word forms used in a different dialect. Having a tool to help identify dialects may be helpful to a language learner who may be reading a work in a different dialect, want to understand relationships between different dialects, and/or want to use spelling and grammar styles more aligned with a given dialect. Hartwig has witnessed learners of one dialect unwittingly use resources from another dialect, which may lead to confusion around spelling and grammar, but with the right guidance such confusion may be alleviated. Research with Indigenous Peoples should be a part of a reciprocal relationship, where work is done to benefit the People providing information by answering questions and exploring topics highlighted by the given Indigenous People (Smith, 2021).

Ojibwe is an oral language, but multiple writing systems have been developed to transcribe it (Treuer, 2010). Ojibwe have used pictographs and similar symbols to write out stories for an unknown period of time. As missionaries and others came to *Anishinaabewaki*, however, such newcomers decided to develop writing systems for the Ojibwe language. Various writing systems were created, including ones based on syllabics and others with Latin script. Roman character-based writing systems are most commonly used today, with the Fiero / double vowel / long vowel orthography being the most popularly used by Ojibwe language educators (Ningewance, 1999). For this reason, this paper will use examples written using the long vowel system.

2.3 Related work

Much of the computational language work that has been performed with Indigenous languages is rule-based (Mager et al., 2018), which often requires expert knowledge. Despite this, there have been attempts to use unsupervised learning methods to learn morphology of Indigenous languages with some success (Johnson and Martin, 2003). One notable example of a rule-based system that is designed for an Anishinaabe dialect is the construction of a morphological parser for Odawa (Bowers et al., 2017).

El Mekki et al. (2020) performs fusion between an n-gram based *support vector machine* (SVM) (Cortes and Vapnik, 1995) and a BERT (Devlin et al., 2019) model trained on Arabic to determine dialect across many countries and regions. The n-grams are computed at the word and character level and are normalized using *term-frequency inverse document frequency* (TF-IDF) before being used in the SVM.

Hämäläinen et al. (2021) also performs fusion between dialect classification models; however, their approach uses both text and audio as inputs and is focused on classifying 23 separate dialects of Finnish. A BERT model trained on Finnish is used to handle the text inputs, which are split at the sentence level.

Salameh et al. (2018) looks at the problem of Arabic dialect identification, introducing a commissioned data set that contains common phrases in dialects from different cities. They find that using character n-grams as well as individual words is a preferred method of featurizing inputs for sentence-level dialect determination with a Multinomial Naive Bayes classifier.

A deep learning approach utilizing pre-trained models was not considered for this work, due to the relatively small amount of text collected and the difficulty in transferring a deep learning model trained in one language to another. It has been noted by Singh et al. (2019) that tokenizers trained on one language do not necessarily transfer to another language efficiently. Another work (Maronikolakis et al., 2021) found that when transferring language models between languages, unless the tokenizer was re-trained, it was less efficient on the new language and required far more tokens to represent the same length of text.

3 Method

3.1 Text used

Several stories in Manitoulin Island varieties of Eastern Anishinaabemowin (Corbiere and Jones, 2012) and several stories from Volume 8 of the *Oshkaabewis Native Journal* (Treuer et al., 2012) for Southwestern Ojibwe were used for this work. Permission to use each of the works was granted from the respective editors. Additionally, dictionaries (Child and Nichols, 2012; Naokwegijig-Corbiere and Valentine, 2015) for each of the dialects were used to create a word list of common words that could also be used to evaluate the dialect classification model.

The amount of text used is quite small—611 sentences of Southwestern Ojibwe and 434 sentences of Eastern—which limits the use of deep learning methods that require large bodies of text from which to learn. For this reason, SVM's were chosen as a method appropriate for classifying small bodies of text, which are sometimes referred to as *low-resource* use-cases. Some sample statistics from both sets of text are included in Appendix B.

3.2 Text processing and model selection

Following the work of Hämäläinen et al. (2021) and El Mekki et al. (2020), the problem is formulated as a dialect prediction for an arbitrary number of sentences. Based on these works, an SVM using character n-grams is utilized with n-gram features combined between the relevant sentences. Stochastic gradient descent was used to train the model, minimizing hinge loss. The SVM implementation written by Pedregosa et al. (2011) was used for this work. N-grams were generated by splitting each word into all possible sets of n characters and were combined for varying numbers of sentences. For example, the word *aaniin* contains the unigrams of *a, n* and *i*; the bigrams of *aa, an, ni, ii* and *in*; and so on for three and four character combinations. Since each sentence could not possibly contain all of the n-grams in the entire text, and because SVM's require a consistent input feature set, all possible n-grams were found from the two combined sets of text and an n-gram dictionary with zero values for all n-grams was used to initialize each sentence set.

These n-grams were counted for each set of sentences and analyzed with the SVM. The two sets of text were randomly split into five parts, maintaining an equal proportion of each dialect in each split, and cross validation was performed; the model was trained on four folds of the data and the unseen fifth was used to validate the model.

Table 1: Model performance as a function of number of sentences used to infer dialect

Number of grouped sentences	Accuracy
1	0.90
2	0.95
3	0.98
4	0.98
5	0.97

4 Results

4.1 Sentence level model training and evaluation

The number of correct and incorrect predictions were summed across all five validation folds and the resulting average accuracy, weighted by class membership, is presented in Table 1. The computation and counting of n-grams was performed using single sentences up to groupings of five sentences. By grouping more sentences together, a wider sample of word parts is captured and allows the model to more easily predict which dialect is present, which is indicated by our results. In our tests, we were able to achieve nearly zero errors with five sentences being used to compute each set of n-grams.

4.2 Interpretability of model

One advantage of using an SVM is that the model weights—i.e., the support vector weights—can be used to understand which features are most influential. In the case of our problem, we are able to associate the n-grams with the highest weights to those that are (based on the training data) most associated with a given dialect. The presence of a given n-gram does not indicate dialect alone, but indicates that a word or sentence containing that n-gram is more likely to belong to a given dialect. The n-grams most associated with each dialect are given in Table 2 and are drawn from the full data set averaged across five folds, and considering all n-grams from single characters up to 4-grams.

Eastern began to reduce unstressed vowels in the early part of the twentieth century (Bloomfield, 1957), and Eastern speakers are often playfully joked about as being vowel droppers. Many vowelless n-grams, such as *bn* picked up by our model for Eastern, would be rarer to find in Southwestern Ojibwe. Several examples of vowel dropping can be found in the word list included in Appendix C such as the word for *otter* being *ngig* in Eastern and *nigig* in Southwestern.

Table 2: Top ten n-grams most associated with each dialect

N-grams most associated with Eastern Ojibwe	N-grams most associated with Southwestern Ojibwe
bn	ay
bm	aye
wi	in
oo	izhi
gd	iz
iinw	izh
gs	gay
booz	gaye
boo	ye
hoo	ina

It is possible that we are observing aspects other than dialect in our analysis, such as the language preferences of the authors of our given texts. For example, the discourse marker *izhi*, meaning *'and so'*, is noted by Fairbanks (2016) as being more frequently used by first language speakers of Ojibwe than second language speakers (among other discourse markers). However, it is also possible that *izhi* has, much like certain vowels, fallen out of common use in Eastern; determining the answer to this question is outside of the scope of our work and is something that could be explored in future collaborations with Anishinaabe language keepers. To address whether our model is overfitting to language preferences rather than aspects of dialect, using writings from a wider range of authors could be used.

4.3 Applying sentence level model to individual words

To further evaluate the model developed, a small dictionary of 50 common words that differ in spelling between Eastern and Southwestern Ojibwe was compiled. Applying the model from the sentence level training to evaluation on word level inputs is an interesting experiment in model transfer and has practical value; as many language learners will encounter unknown words and may want to determine what dialect they are originating from. Each individual word from this dictionary was evaluated using the model trained on groups of five sentences. The model was found to be 72% accu-

		Predicted	
		E	SW
Actual	E	37	13
	SW	14	33

Figure 2: Confusion matrix of individual word predictions using sentence-trained SVM

rate. A confusion matrix of the word predictions is presented in Figure 2, which shows that the model does not favor either dialect. Due to multiple common words being used in Eastern for some of the selected words, three words of Southwestern are repeated and not included in the confusion matrix. The full results of this study, including each word used and its predicted dialect, can be found in Appendix C. We considered repeating the sentence level experiment with the individual words, but found that our dictionary was insufficiently small.

5 Conclusions

In this work, we have proposed and evaluated a method for identifying dialects in Ojibwe given a small set of labeled examples. We showed that our method is 90% accurate at the single-sentence level and higher still at the multi-sentence scale. We also achieved a 72% accuracy when the sentence-level model was applied to a selected set of individual words. The model proposed also offers insight into how dialects are classified by the model, demonstrated by explaining the significance of some of the n-grams found to be most significant in determining dialect. This aspect of interpretability could offer language learners insight into features differentiating written dialects as well as providing a tool to help determine the dialect of unfamiliar text.

Limitations

This work focuses on using computational tools to determine dialect based on a small quantity of writings of a spoken language, using a writing system that was adapted recently, rather than one that evolved alongside the language for thousands of years. This limitation in orthography leads to differences in character usage, frequently between dialects (which is helpful for this problem), but there is also variation also within dialects depending on the author. For example, different writers will use different methods of transcribing a nasal sound; Eastern tends to use *nh* for nasal sounds in the middle of a word, although some writers will use a capital *N*. Southwestern Anishinaabemowin tends to use *ny, ns* or *nz* for these same nasal sounds within words. As noted by Valentine (2001b), there are variations in language within dialects, including age-stratified language proficiency, where older speakers tend to be more fluent than younger ones, largely due to differences in opportunities to learn the language. These differences might be detected and interpreted as dialect differences if the diversity in writers is not comparable between the two sets of texts being compared. Additionally, an individual's word choice may change depending on their gender or occupation (Valentine, 2001b), and having only a small sample of writings does not allow us to capture these differences well. To test how much our model is learning author preferences over dialects, using some writings from authors not included in the training data would provide some insight. Future work could do a better job of tracking authorship between cross validation folds as well as sourcing from a wider set of writers. Only small quantities of text were used for each dialect, which was limiting in terms of methods that could be used. The methods utilized in this paper could be easily applied to minority languages that do not have large quantities of written text available, of course, with permission from and in collaboration with Indigenous language keepers. Our future work could involve the collection of larger quantities of text, which would allow the use of a wider range of language analysis.

Ethical Statement

Some of the texts used for our samples were transcribed *aadizookaanan*, a type of traditional story highly revered by Ojibwe. These particular stories are not to be spoken out loud during non-winter months without snow on the ground. There are particular spiritual reasons for this, and unfortunate things can happen to individuals telling or hearing these stories when there is no nearby snow. Therefore, we will not write out the stories here and we strongly encourage citation followers to heed precaution. For more information, bring your tobacco and questions to a trusted Anishinaabe knowledge keeper.

Indigenous Peoples have experienced a long history of colonialism, including by well-meaning researchers. Please remember that Indigenous Peoples must maintain sovereignty over their languages, traditional stories, and other knowledge. All research involving Indigenous knowledge, including that for the development of generative AI, should be done ethically in reciprocal relationships with Indigenous Peoples. The research should also meet their needs and wants, as described by the given Indigenous Peoples (Smith, 2021).

Acknowledgements

Miigwech / thank you, Dr. Waagosh Anton Treuer, Dr. Ojiig Alan Corbiere, Dr. Migizi Michael Sullivan, Nswi Aanddegook Isadore Toulouse, Ninaatig Staats Pangowish, Gimiwan Dustin Burnette, Jared Blanche, Aandeg Muldrew, and everyone else who has supported us with the writing of this article.

References

Charles A Bishop. 2019. Ojibwa.

Leonard Bloomfield. 1957. *Eastern Ojibwa: Grammatical sketch, texts, and word list*. University of Michigan Press.

Dustin Bowers, Antti Arppe, Jordan Lachler, Sjur Moshagen, and Trond Trosterud. 2017. A morphological parser for odawa. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 1–9.

Gimiwan Dustin Burnette. 2023. Personal communication.

Brenda J Child and John Nichols. 2012. The ojibwe people's dictionary.

Alan Corbiere and Alana Jones. 2012. Baadwewdangig. University of Toronto Linguistics Department course Language Revitalization (LIN458) Project.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20:273–297.

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.

Abdellah El Mekki, Ahmed Alami, Hamza Alami, Ahmed Khoumsi, and Ismail Berrada. 2020. Weighted combination of bert and n-gram features for nuanced arabic dialect identification. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 268–274.

Brendan Fairbanks. 2016. *Ojibwe discourse markers*. U of Nebraska Press.

Mika Hämäläinen, Khalid Alnajjar, Niko Partanen, and Jack Rueter. 2021. Finnish dialect identification: The effect of audio and text. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8777–8783.

Kalvin Hartwig. 2012. *Language as an Aspect of Identity and Indigeneity*. Masters major paper, Yale University.

Mary Hermes and Kendall A King. 2013. Ojibwe language revitalization, multimedia technology, and family language learning.

Tom Holm, J Diane Pearson, and Ben Chavis. 2003. Peoplehood: A model for the extension of sovereignty in american indian studies. *Wicazo Sa Review*, 18(1):7–24.

Howard Johnson and Joel Martin. 2003. Unsupervised learning of morphology for english and inuktitut. In *Companion volume of the proceedings of HLT-NAACL 2003-short papers*, pages 43–45.

C J Lippert. 2007. Location of all anishinaabe reservations/reserves in north america, with diffusion rings about communities speaking an anishinaabe language. cities with anishinaabe population also shown.

Manuel Mager, Ximena Gutierrez-Vasques, Gerardo Sierra, and Ivan Meza. 2018. Challenges of language technologies for the indigenous languages of the americas. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 55–69.

Antonis Maronikolakis, Philipp Dufter, and Hinrich Schütze. 2021. Wine is not v i n. on the compatibility of tokenizations across languages. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2382–2399.

Brian D McInnes. 2014. Teaching and learning ojibwe as a second language: Considerations for a sustainable future. *Journal of Language Teaching and Research*, 5(4):751.

Mary Ann Naokwegijig-Corbiere and Rand Valentine. 2015. Nishnaabemwin web dictionary.

John David Nichols. 1980. *Ojibwe morphology: a thesis*. Ph.D. thesis, Harvard University.

Pat Ningewance. 1999. *Naasaab Izhi-anishinaabebii'igeng Conference Report: A Conference to Find a Common Anishinaabemowin Writing System*. Ministry of Education & Training, Workplace Preparation Branch, Literacy ....

Ninaatig Staats Pangowish. 2023. Naadimaadizang mi-inwaa naadimaading, helping one's self and help one another. Anishinaabemowin Teg.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Brock Pitawanakwat. 2018. Strategies and methods for anishinaabemowin revitalization. *Canadian Modern Language Review*, 74(3):460–482.

Richard A Rhodes. 2006. Ojibwe language shift: 1600-present. *Historical Linguistics and Hunter-Gatherer Populations in Global Perspective, MPI-EVA Leipzig*.

Mohammad Salameh, Houda Bouamor, and Nizar Habash. 2018. Fine-grained arabic dialect identification. In *Proceedings of the 27th international conference on computational linguistics*, pages 1332–1344.

Jasdeep Singh, Bryan McCann, Richard Socher, and Caiming Xiong. 2019. Bert is not an interlingua and the bias of tokenization. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 47–55.

Linda Tuhiwai Smith. 2021. *Decolonizing methodologies: Research and indigenous peoples*. Bloomsbury Publishing.

Anton Treuer. 2010. *Living our language: Ojibwe tales and oral histories*. Minnesota Historical Society Press.

Anton Treuer, Joe Chosa, David Treuer, James Clark, William Jones, Diane Amour, Edward Benton-Banai, Jessie Clark, Alex Decoteau, Michelle DeFoe, George Fairbanks, Rick Gresczyk, Charles Grolla, Jason Jones, Jeremy Kingsbury, Earl Otchingwanigan, and Vernon Whitefeather. 2012. *Oshkaabewis Native Journal (Vol. 8, No. 1)*. Lulu.com.

U.S. Census Bureau. 2011. 2010 census. U.S. Department of Commerce.

J Randolph Valentine. 2001a. Being and becoming in ojibwe. *Anthropological linguistics*, pages 431–470.

Jerry Randolph Valentine. 1994. *Ojibwe dialect relationships*. The University of Texas at Austin.

Randy Valentine. 2001b. *Nishnaabemwin reference grammar*. University of Toronto Press.

A Appendix A: Feature Scale Study

To understand the number of features being created and used by the model, a simple scale study was performed. Model features were counted for various numbers of n-grams used and model performance as a function of limited n-grams was computed. To understand how many of the most common n-grams are shared between dialects, the n-gram dictionaries for each dialect were sorted and compared for overlapping n-grams. The results of this are presented in Table 3, where it can be seen that the relative overlap in n-grams decreases with increasing dictionary sizes. Intuitively, this makes sense, as a common language would share the most common features between dialects and differences should become more apparent with larger feature sets. To quantify the performance of our proposed model with a very limited feature set, the smallest truncated dictionary was used to repeat the analysis and is presented in Table 4. Performance with n-grams derived from single sentences is substantially lower than when using the full n-gram dictionary, which shows how important the less common character combinations are to identifying the dialect present. Interestingly, when n-grams from four sentences are combined, performance between models is comparable.

Table 3: Overlap in top n-grams

Number of most common n-grams used	**Number of overlapping n-grams**	**Combined dictionary size**
100	82	118
500	348	652
1000	650	1350
All	2217	6259

Table 4: Model performance as a function of number of sentences used per example, using truncated n-gram dictionary with 118 n-grams.

Number of grouped sentences	**Accuracy**
1	0.64
2	0.73
3	0.82
4	0.98
5	0.97

B Appendix B: Text Statistics

To help illustrate the corpus used for this work, some statistics are shared in Table 3.1. To help keep the data set sizes similar, not all of the stories from Treuer et al. (2012) were used in this work.

Table 5: Information about texts used

Text	**Number of sentences**	**Average number of words per sentence**
Southwestern	611	6.7
Eastern	434	7.6

C Appendix C: Full Results of Individual Word Classification

The full table of word pairs between Eastern and Southwestern Ojibwe is presented in Table 6. Fifty word pairs, along with their approximate English translation, were selected by choosing words that a language learner might learn at an early stage in their learning process. When multiple words are commonly used for a similar meaning in one dialect but not another (for example *makwa, mkwa* and *mko*), the table repeats the word for the dialect without multiple common words found in the appropriate dictionary. This is done for visual clarity for the reader. Multliple words were not included in the statistics computed in Figure 2.

Table 6: Fifty common words that vary between Eastern and Southwestern Ojibwe and our model's classification

Ojibwe SW dictionary form	Classified by our model as	Ojibwe E/Odawa dictionary form	Classified by our model as	English (approximate)
aaniin	E	aanii	E	hello (pc interj)
daga	E	bna	E	please (pc disc)
niin	E	niinii	E	me (n)
niin	E	nii	E	me (n)
giin	E	gii	SW	you (n)
enyanh'	E	ehenh	E	yes (pc disc)
en'	E	enh	E	yes (pc disc)
gaawiin	SW	gaa	E	no (pc disc)
gaawiin	SW	kaa	E	no (pc disc)
wiindan	SW	waawiindaan	E	name (vti)
izhinikaazowin	SW	zhnikaazwin	E	name (n)
minawaanigozi	SW	mnowaan'gozi	E	is happy (vai)
izhinaagozi	SW	zhinaagzi	E	look a certain way (vta)
izhinaagwad	SW	zhinaagot	SW	look a certain way (vti)
ojiim	SW	jiimaa	SW	kiss (vta)
ojiindan	SW	jiindaan	SW	kiss (vti)
wiiisini	SW	wiisni	E	eat (vai)
amo	E	mwaa	E	eat (vta)
minikwe	SW	mnikwe	E	drink(vai)
aabitoojiin	SW	aabtoojiinaa	E	hug around middle (vta)
giziibiiga'an	SW	gziibiignaan	E	wash something (vti)
opin	SW	pin	SW	potato (n)
manoomin	SW	mnoomin	E	wild rice (n)
mishiimin	SW	mshiimin	E	apple (n)
odaabaan	SW	daabaan	SW	car (n)
makwa	SW	mkwa	E	bear(n)
makwa	SW	mko	E	bear (n)
ma'iingan	SW	m'iingan	SW	wolf (n)
nigig	SW	ngig	SW	otter (n)
mooz	E	moos	E	moose (n)
waawaashkeshi	E	waawaashkesh	E	white-tailed deer (n)
giingoo	E	giigoonh	E	fish (n)
ogaa	SW	gawaak	SW	walleye / pickerel (n)
adikameg	SW	dikmek	E	whitefish (n)
mitig	SW	mtik	E	tree (n)
nagamo	E	n'gamo	E	sing (vai)
giiwese	SW	giiwse	E	hunt (vai)
baashkigizige	SW	baashkzige	E	shoot (vai)
agindaaso	SW	n'gidaaso	SW	read (vai)
babaamose	SW	bbaamse	E	walk about (vai)
bikwaakwad	SW	bkwaakwat	E	ball (n)
gimiwan	SW	gmiwan	SW	rain (n)
waabooz	E	waaboos	E	rabbit (n)
bakwezhigan	SW	bkwezhgan	E	bread (n)
waasechigan	SW	waasechgan	SW	window (n)
wewebizo	SW	wewebza	E	swing (vai)
akwaandawe	SW	kwaandw	E	climb (vai)
bagizo	SW	bgiza	SW	swim (vai)