Chapter 4

Methodology

This chapter details the methods used in the analysis of Nêhiyawêwin Order. The primary research question investigated in this dissertation is: how, and in what way, can Nêhiyawêwin order be understood as an alternation that can be predicted through morphosyntactic, surface-syntactic, and lexical-semantic features. This chapter describes the corpus used, the univariate analysis, and the multivariate analysis. The methodologies used in this analysis are based off those bivariate and multivariate statistics described in Bresnan et al. (2007); Divjak and Gries (2006); Gries (2003) and Arppe (2008), in particular the combination of univariate and multivariate techniques. This chapter does not detail the methods used for creating the underlying corpus (information detailed at length in Arppe et al. (2020)) or the process by which verbs and nouns were semantically clustered for inclusion as predictors (described in Chapter 3).

4.1 The Corpus

The underlying corpus from which the data set used in this dissertation is the Ahenakew-Wolfart corpus (Arppe et al., 2020). The Ahenakew-Wolfart corpus is likely the largest morphosyntactically tagged corpus of all Canadian Indigenous languages, let alone Nêhiyawêwin. Although there has been attemtps in the last few decades to increase the amount of texts in Nêhiyawêwin, there is still a paucal amount of texts written in Nêhiyawêwin, and many of those texts that are publicly available, are written in a nonstandard Roman orthography. The Ahenakew-Wolfart corpus is unique in that it is meticulously standardized. The texts that make up the corpus were collected

by Freda Ahenakew and H. C. Wolfart between the 1970s and 1990s. These texts have previously been published in Ahenakew (2000); Bear et al. (1992); Kâ-Nîpitêhtêw (1998); Masuskapoe (2010); Minde (1997); Vandall and Douquette (1987); Whitecalf (1993). These texts are mainly dialectic or narrative discussions between two or more native Nêhiyawêwin speakers. Together, these texts contain 142,192 tokens (20,503 types), though some of these tokens are English, French, or Michif words; fragments; or other items. Focusing only on Nêhiyawêwin items, there are 80,221 tokens (16,532 types). Each of these tokens has been morphosyntactically tagged by automatic and hand-parsed means (Arppe et al., 2020). Tokens were tagged for their lemma as well as both verbal and nominal features. For verbs: preverbs, tense, word class, Order, commitative morphemes, and conjugation class; For nouns: person/number marking, possession, declension, and diminutive morphemes; Both nouns and verbs were marked for the feature of semantic class. An example token with its relevant tags is found in (13).

(13) ê-ohci-pimâtisit
pimâtisiw PV/e PV/ohci V AI Cnj 3Sg @PRED-AI
'S/he lived thus / make a living thus'

Beyond this, the corpus has been further syntactically tagged by an automatic constraint grammar (Schmirler et al. 2018, Schmirler Forthcoming). Among other features, this constraint grammar marks tokens for their predicate, actor, and goal status.

To create the data set used in this dissertation, I extracted only verbs from the above corpora and further restricted the data set by selecting only verbs that contained a classification as described in Chapter 3. This results in a data set of 13,628 tokens (2032 types). In addition to the morphosyntactic tags seen above, verbs were marked for arguments (and those arguments' morphosyntactic features) when arguments were syntactically present (as opposed to represented only by verbal agreement). This results in an entry such as (14).

(14) pimâtisiw PV/e PV/ohci V AI Cnj 3Sg @PRED-AI AI-state
kikâwînaw N A D Px1Sg Sg @ACTOR> NDA-Relations

From here, each token and its accompanying analyses were transformed into a

Table 4.1: Extract from Data Frame

Lemma	PRED-AI	PV/ahci	...	PV/e	PV/ohci	PV/pe	V	AI	Cnj	3.actor	3.goal	AI-state	Sg.actor
pimâtisiw	TRUE	FALSE	...	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE

data frame of 'dummy' variables: every verb lemma token makes up a row, while every morphosyntactic tag constitutes a logical column. For every lemma token, if a morphosyntactic feature is observed, a value of `TRUE` is set for the corresponding column, otherwise a value of `FALSE` is set. Dummy variables allow for easily interpreted results, especially when dealing with covariance (Baayen, 2012). Given the example of (14) the data frame extract in Table 4.1 is produced.

For the sake of fitting Table 4.1 to the page, the majority of the columns are not shown, but every feature present in (14) would have a column value of `TRUE` for the token `pimâtisiw`, and all features not present are given a value of `FALSE`. The exception to this is the actor and goal marking morphemes. Although the corpus marks person and number of morpheme as one unit (e.g. `1Sg`), the data frame used for analysis in this dissertation split the features up (i.e. there were seperate columns for `3` and `Sg` for both actors and goals). Finally, a number of tokens were removed because their part of speech was not reliably identified in the corpus. There were 310 of these tokens, the majority of which (301 tokens) were the verb *ayâw*. In addition to basic locative use, *ayâw* may also be used to describe the state of 'having' something. In the corpus, *ayâw* was marked as both `VAI` and `VTI`. Because the VTI form of the verb inflects the same as the VAI form, and because syntactic arguments are usually not present in a sentence beyond verbal agreement (and even then, only in the VTA), determining which of these two classes the lemma was acting in was difficult for the non-native speakers annotating the corpus. Three further lemmas, *manitowi-kîsikâw* (4 tokens), *misi-paskwâw* (3 tokens), and *nanamipayiw* (1 token), were removed as there was disagreement between the corpus and dictionary sources. In the first two cases, these forms were given in the corpus as VIIs, while dictionary sources cited them as also NIs. This disagreement is understandable, as VIIs that deal with time or space often describe substantives. The final case, *nanamapayiw* is given as an VII in the corpus, while Wolvengrey (2001) analyses it as an VAI and LeClaire et al. (1998) offers an

Table 4.2: Preverb Class Tokens and Types

	Types	Tokens
Discourse	4	277
Position	15	285
Qual	30	316
Quant	7	10
Time	18	4720
Move	4	731
Start/Finish	5	229
Want/Can	4	195

analysis of both VAI and VII. Although context (through either native speaker annotation or translations by native speakers) would quickly resolve these ambiguities, the corpus being used had not yet been disambiguated in this sense, and given the small number of tokens, I opted to remove these 309 tokens from the data set.

Because Nêhiyawêwin contains a large number of possible preverbs (the model underlying the corpus could identify 267 unique preverbs), I undertook a manual classification of these morphemes. I identified 8 unique classes: Discourse, Position, Qual, Quant, Time, Move, Start/Finish and Want/Can. Of the 267 identified preverbs, only 86 preverb types were observed in the corpus. Table 4.2 lists the number of tokens and types in each of the preverb classes.

In all, the resulting data frame of non-imperative forms contains 13,292 lemma rows by 4777 columns. Due to errors in coding, 100 items were excluded from this, creating a data frame of 13,192 items. The use of such a logical data frame for predicting an alternation is present in Arppe (2008) and allows for the assessment of individual values of categorical variables through straightforward application of chi-squared analyses and logistic regression to predict a multinomial alternation, in this case Order.

4.2 Modelling the Alternation

In this dissertation, I will evaluate a univariate analysis given the morphosemantic features mentioned above to model a verb lemma's likelihood of occurring in various Order types. Although Chapter 2 identified five unique Conjunct Orders (along with

Table 4.3: Preverb Class Tokens and Types

		Types	Tokens
Independent		876	4390
ê-Conjunct		1480	6378
kâ-Conjunct		600	1696
Other-Conjunct		393	828
	Subjunctive	75	100
	Initial Change	18	21
	ka-Conjunct	344	707
Total		3349	13,294

the Independent), the majority of these classes have few tokens. Small counts can be problematic for statistical analyses, particularly for regression analyses. To address this, the ka-/ta-Initial, Initial Change, and Subjunctive Conjuncts were conglomerated into a single 'other' class. This results in the Order alternations as seen in Table 4.3.

In order to gain a wholistic understanding of Nêhiyawêwin Order, this dissertation will investigate 3 main alternations of these Orders:

- Independent vs Conjunct
- Independent vs ê-Conjunct
- Conjunct Type: ê-Conjunct vs kâ-Conjunct vs Other-Conjunct

The first of these alternations, Independent vs Conjunct, will inform about the difference between the two Orders broadly. The second, Independent vs ê-Conjunct, will investigate the difference between the two most similar Order forms which are often conceived as synonyms and used roughly interchangeably. The third alternation will be used to model the extent to which we can predict the modes through morphosemantic features from a corpus. Three main data frames were used:

- `AWnImp`: used in analyzing the Imperative vs. Conjunct alternation, representing all non-imperative forms minus the 100 errors previously mentioned
- `AWIvE`: used in analyzing the Independent vs. ê-Conjunct alternation, representing only forms of forms with `TRUE` `Ind` or `PV.e` forms

Table 4.4: `AWnImp` *statistics*

	Types	Tokens
Independent	876	4390
Conjunct	1722	8802
Total	2598	13,192

Table 4.5: `AWIvE` *statistics*

	Types	Tokens
Independent	876	4390
ê-Conjunct	1480	6378
Total	2356	10,768

- `AWCnj`: used in analyzing the Conjunct Type alternation, representing only forms with `TRUE` `Cnj` forms.

In Table (4.4) through (4.6) are relevant counts for each of the three dataframes.

4.3 Univariate Analyses

The term *univariate analysis* refers to an analysis that takes into account only one variable at a time. The most common form of univariate analysis for discrete variables is the chi-square test, originally introduced in Pearson (1900) and refined over the last century to produce the modern day chi-squared test (Agresti, 2013). The chi-square test makes use of contingency tables to measure the association/correlation of a (set of)

Table 4.6: `AWCnj` *statistics*

	Types	Tokens
ê-Conjunct	1480	6378
kâ-Conjunct	600	1696
Other-Conjunct	393	828
Total	2473	8902

variable(s) an outcome. This is calculated by comparing the expected frequency of an outcome/variable pair with with the observed frequencies of the same pairings. Chi-square tests provide a simple statistic, the eponymous χ^2 statistic, whose value reflects an estimated association. This statistic is given for the whole *set* of values of the explanatory and outcome variables tested. If one were to run a chi-square test to determine if the set of variables { 1Sg.actor , 2sg.actor , 3sg.actor , past tense , future tense , present tense } was associated with an increased likelihood of a lemma being in the Independent or Conjunct Order, the resulting χ^2 statistic would indicate the level of association for that set as a whole. To investigate the effect an individual variable has, one must make use of the Standardized Pearson Residual, calculated through the formula in (4.1), where *P* is the Standardized Pearson Residual, *O* is the observed frequency of a variable/outcome pair, *E* is the Expected frequency of a variable/outcome pair, t_i is the sum of a variable across all outcomes, and t_j is the sum of all variables for a given outcome (adapted from (Agresti, 2013, 81)). Note that in (4.1) the denominator represents its standard error.

$$P = \frac{O - E}{\sqrt{E(1 - t_i)(1 - t_j)}} \tag{4.1}$$

This produces a Standardized Residual which can be interpreted based on its magnitude and direction. A positive residual of at least 2.00 represents a significant positive association (i.e. one observes more instances of a variable/outcome pairing than would be expected) while a negative value of -2.00 or lower represents a negative association. Values greater than -2.00 but less than 2.00 represent an association not deemed to be significant (Agresti 2013, 81; exemplified in Arppe 2008, 79).

The chi-square test is best used with higher frequency data sets. According to Cochran (1954), the results of a chi-square test are not reliable when the contingency tables for a given variable has more than 20% of its expected values <5. In these cases, it is suggested that researchers make use of an alternative test, such as the Fisher's Exact Test that forms the basis of Gries' Collostructional Analysis (2004). Some authors, however, believe that Fisher's Exact Test is too conservative (D'Agostino et al., 1988), increasing the risk for Type II errors in hypothesis testing. For this dissertation, I will simply consider phenomena with sufficient frequencies for a chi-square statistic.

Univariate Models

In building models for univariate analysis, all variables with a minimum occurrence of 10 were selected for a given conjugation class for each alternation. This restriction was chosen to exclude incredibly infrequent items which make statistical modelling difficult or unreliable, while including as many variables as possible. Because univariate analysis considers variables on their own basis, manual scrutiny of variable selection was not performed at this point.

4.4 Bivariate Analyses

Following Arppe (2008), after univariate analyses were conducted and a set of variables were selected, I conducted bivariate analyses. Bivariate analysis is simply measuring the assosciation between two variables. Bivariate analysis as done by Arppe (2008) can be a useful tool for creating models for mixed effects modelling. Bivariate analysis for this dissertation makes use of the `associations` function from the `polytmous` package (Arppe, 2013). This function calculates Theil's Uncertainty Coefficient (Henri, 1970a) for every combination of variables passed to it. This coefficient is a mutual information measure and describes the extent to which knowing about one variable can inform our understanding of another variable via a reduction of entropy (Arppe, 2008, 90).

4.4.1 Bivariate models

For this dissertation, I make use of Theil's Uncertainty Coefficient to identify to identify potential covariance which could impede the fitting of mixed effects models. Bivariance was tested for each of the four alternations mentioned above. Variables for each alternation were chosen only from those items with a significant χ^2 statistic ($p < 0.05$). Automatic and manual classes were tested separately, as there was a great deal bivariance between automatic and manual class variables.

4.5 Multivariate Analysis

Using the methodology of Arppe (2008), following bivariate analysis, the resulting variable sets were used to form a set of variables to perform multivariate analysis. The fundamental technique used in this analysis was logistic mixed effects regression. Logistic regression is a generalized form of linear regression as applied to categorical outcomes. Logistic regression models a binary outcome such to what extent individual predictor affects an outcome (Agresti, 2013, 163). Like all generalized linear models, logistic regression attempts to predict outcomes by representing the distribution of the data. Specifically, the technique allows researchers to specify a set of predictors and models the data so that researchers can determine the extent to which an individual predictor influences a particular outcome given a set of parameters (variables/effects).

Logistic regression (with a single independent variable, for example) can be modelled with the equation in (4.2) (Agresti, 2013, 163), where x represents the independent variable, α represents a model intercept, and β is the slope of x.

$$\pi(x) = \frac{e^{\alpha+\beta x}}{1 + e^{\alpha+\beta x}} \tag{4.2}$$

Equation (4.2) represents the odds ratio for the effect of an independent variable on a particular outcome (e.g. the effect of age on the use of one of two synonyms). These ratios are bounded between 0 and ∞. More commonly, logistic regression models are fit with the logit function, derived from (4.2) and seen in (4.3) (Agresti, 2013, 163).

$$\mathtt{logit}[\pi(x)] = \log \frac{\pi(x)}{1 - \pi(x)} = \alpha + \beta x \tag{4.3}$$

The resulting estimates given by the logit are given in log odds, rather than odds. These values are *not* bounded between 0 and ∞, but instead $-\infty$ and $+\infty$. Positive values represent an increase in likelihood of a an outcome for a particular variable; negative values represent a decrease in likelihood; a value of zero represents no effect on the outcome.

This dissertation makes use of mixed effects models in its logistic regression. In terms of regression for language data, mixed effects models have now become the norm (Barth and Kapatsinski, 2018, 100). In comparison to models that make use of only fixed effects,

those variables for which all possible values are represented in the data, mixed effects models allow for the researcher to control for variables in which random variation can be expected (Baayen, 2012). For the data used in this dissertation, morphosyntactic features like `Actor.1`, which are dummy variables that represent the presence or absence of a feature (in this case, whether or not a verb is marked for first person), are **fixed effects** because all possible values (`TRUE, FALSE`) are represented in the data. Conversely, the `Lemma` variable (a multi-level variable containing all lemmas of the corpus) are *samples* of the total lemma set in Nêhiyawêwein and thus can be expected to contain some amount of random variability/outcomes not present in the corpus; thus, `Lemma` is best modelled as a **random effect**. In a mixed effects model, the random variability of a random effect is 'controlled' for, allowing for estimations of fixed effects without the confounds of the random effect.

Fixed effects are analyzed relatively straightforwardly: for each of the dummy variables, one of the two possible levels are chosen to act as a baseline reference (Baayen, 2012). By default, R uses the `0/FALSE` level as a base line, though one could use the alternate level as a reference if needed. For the dummy variables in this dissertation, this means the reference level represents the absence of a particular variable. In modelling an outcome, the logistic regression analyses each observation in its training data and, if an outcome is *not* observed, assigns the variable a value of 0 for the outcome; otherwise, if the variable *is* observed, a value of 1 is given to the variable for the outcome Baayen (2012). Importantly, a model's intercept represents the variables' reference levels (Baayen, 2012). Random effects are not given a reference level; instead, each level can be thought of as adjustments to each fixed effect (Baayen, 2012). As an example, given the fixed effect `actor.1`, the logistic model would make adjustments to `actor.1`'s slope based on observations of each level of `Lemma`. In this sense, there is no reference level the others are compared to.

This analysis makes use of the `lme4` package in R (Bates et al., 2015).

4.5.1 Binarization of the Alternation

Making use of logistic regression, this dissertation will investigate the behaviour of four three alternations based on the order presented in Chapter 2:

1. The Independent vs. the Conjunct (as a whole)
2. The Independent vs. the ê-Conjunct
3. The ê-Conjucnt vs. the kâ-Conjunct vs. the Other-Conjunct (Comparing modes within the Conjuncts)

The comparisons made here allow for investigation of a a wide range of Order behaviour, while still specifying the alternations not generally explainable by semantics: The ê-Conjunct and the Independent along with the general Conjunct and the Independent.

Logistic regression assumes a dichotomous decision by default. This is the case, for example, when comparing the Independent and the Conjunct. For the final two alternations above, however, there are more than two outcomes being compared. In polytomous cases, there are multiple methods by which the data can be binarized. One such technique is the *one-vs-rest* (OVR) heuristic. In one-vs-rest comparisons, a model compares one class against all other possible classes (Frank and Kramer, 2004). In this way it can be thought of as comparing χ against $\neg\chi$. One can also make use of the *pairwise comparison*, which creates pairs of the outcome given a predictor, tests each pair, and selects which outcome is most likely to occur with each predictor (Frank and Kramer, 2004).In the present study, each Conjunct class would be paired with one other class at a time, eventually being paired with each other Conjunct type. For each pairing, which-ever type is most probable given each set of predictors, would be picked as the 'winner' for that pairing, and would be proposed to the model. For each morphological feature, that class which is most often proposed would be selected and given as the likeliest Order type. In addition to one-vs-rest and pairwise comparisons, we can make use of *nested dichotomies*. The concept of a nested dichotomy is straightforwardly described: a group of outcomes being compared is split into mutually exclusive dichotomies repeatedly until unary classifications are created (Frank and Kramer, 2004). Each dichotomy is given a statistical probability which can be multiplied together to determine the overall probability of the dichotomous tree. Similarly, other binarization techniques such as Baseline-Category (Fox, 1997, 468) classification offer alternative binarization techniques. Arppe (2008) presents a brief

overview of all the above techniques. For this dissertation, I will use the OVR heuristic in the polytomous Conjunct Type alternation due to its conceptual and computational ease and simplicity, particularly in modelling maximally three outcomes in a single alternation.

The resulting logistic models provide estimated adjustments for every variable, even if these effects are not considered statistically significant. The results provided by the `lme4` package calculates the p-value for each effect using the asymptotic Wald tests for generalized linear models (Bates et al., 2015). Recognizing that the use of p-values are not without controversy (Gelman, 2016), this dissertation will still use p-values to determine which effects are most pertinent in modelling the Order alternation.

4.6 Model Assessment

In addition to the results described above, one can assess the overall performance of a logistic model. This assessment gives us invaluable information and allows us to see how well, and in what ways, a model represents Order type selection in terms of morphosyntactic and semantic features. In particular, one can see how well a model is able to evaluate a given form as the correct Order type without raising false positives (precision), as well as how many instances of a given Order type it classifies correctly, regardless of false positives (recall). Recall, precision, and overall accuracy are measured per Conjunct type for each conjugation class (not for the model in its entirety). As an additional way of assessing model fit, one may also use the τ value to determine how much better the model performs than selecting based solely on overall proportions/through random assignment (Goodman and Kruskal, 1959, 745-747) (ranging from 0.00 to 1.00, with 1.00 being a perfect model). According to Arppe (2008, 140), τ values of roughly 0.5 and above suggest a 'good' model fit. Because the models used in this dissertation are logistic, a true R^2 *Likelihood* score is inapplicable. Instead, a so called Pseudo-R^2 value must be used. As Hosmer and Lemeshow (2000, 167) point out that Pseudo-R^2 Likelihood scores for logistic regression are generally much smaller than in other statistics, such as R^2 values given in standard linear models. Another important difference between the R^2 measure and Pseudo-R^2 Likelihood is

that the former can be used as a measure of how much variance is explained by the model under consideration; Pseudo-R^2 Likelihood can never report explained variance (Hosmer and Lemeshow, 2000, 164). Instead, Pseudo-R^2 Likelihood can be seen as a measure of reduction in the *badness* of fit. The specific form of Pseudo-R^2 that I will use is McFadden's Pseudo-R^2 (ρ^2) (Domencich and McFadden, 1975) as reported by the `ModelStatistics` function (Arppe, 2013). Macfadden's Pseudo-R^2 appears to have a stable, but non-linear, relationship with a general R^2, wherein a ρ^2 value of 0.2, 0.3, and 0.4 are roughly equivalent to an R^2 of 0.3, 0.5, and 0.73 respectively (Domencich and McFadden, 1975, 124). As with other Pseudo-R^2 measures, a ρ^2 of over 0.25 is indicative of a fairly well fit model. Furether, Han et al. (2013) suggest that, in their experience, Pseudo-R^2 Likelihood scores of nearly 0.30 are indicative of very good models without risk of over-fitting.

Because models are fit separately for each Conjunct type in each conjugation class in the Conjunct Type alternation, estimated probabilities add up to something close to, but not exactly, 1.00. In order to achieve this range of 0.00-1.00, the `ModelStatistics` function (Arppe, 2013) aggregates all models and performs a normalization of estimated probabilities, such that they add up to exactly one.

In addition to the models described above, one can use logistic regression with only random effects. These models present only those lemma-specific effect, and can be used as a base line against which one can assess how much of an effect morphological features have on the ability to predict Conjunct type (cf. the discussion of Harrigan and Arppe (2015) regarding lemma-specific preferences on occurrence in the Independent or Conjunct orders). For each of the mixed-effect models, Pseudo-R^2 Likelihood, Accuracy, and τ measures for models with only random effects will be also be given. By comparing fixed effects models against the mixed effects models, we can determine the extent to which random effects affect the fit of our modelling of Order.

Based on the above methodology, the following predictions are proposed:

1. Overall, modelling will be successful though constrained (likely due to the small size of the corpus).

2. Due to a lack of syntactic data, mixed effects modelling based on the Nêhiyawêwin

corpus will be able to provide some insights, but model fits will rarely be significantly informative (as measured by the exceeding of an ρ^2 of 0.2).

3. Semantic variables will do more to explain variance than morphological variables (as in (Arppe, 2008)).

4. The Conjunct Type alternation will be significantly less cohesive in its results (due to the straightforward syntactic/semantic choices driving it, which are not reflected in the variables of the data set).

5. The alternation between the Independent and the ê-Conjunct will be the most robust and well-fit model (because the two forms are nearly synonymous in many cases).