1

A Computational Analysis of Navajo Verb Stems

DAVID EDDINGTON AND JORDAN LACHLER

1 Introduction

One of the principal goals of linguistics is to find, classify, and describe relationships between words. Many formal mechanisms such as rules and constraints have been devised in order to show systematic relationships. Inflectional paradigms are a crucial component of a linguistic analysis that has applications for pedagogical grammars. For example, over the past 20 years there have been numerous Navajo textbooks produced that are aimed at beginning learners of the language. These include works such as *Diné Bizaad Bóhoo'aah* (Navajo Language Institute 1986), *Diné Bizaad: Speak, Read, Write Navajo* (Goossen 1995), and *The Navajo Verb: A Grammar for Students and Scholars* (Faltz 1998). However, there is one important feature of Navajo grammar which none of these works deals with directly, namely the inflection of the verb stem. For instance, Faltz (1998), while providing a remarkably lucid treatment of the verbal prefix systems, has comparatively little to say about patterns of verb stem inflection. This is true also of the two large Young & Morgan dictionaries (1987, 1992) that serve as the primary reference works on Navajo.

Verb stem inflectional patterns in Navajo are arguably one of the most intractable problems in modern Athabaskan linguist studies. While all Athabaskan languages display complex systems of verb stem inflection (see

Empirical and Experimental Methods in Cognitive/Functional Research.
Sally Rice and John Newman (eds.).

Axelrod 1993 for Koyukon; Rice 1989 for Slave), the system of Navajo is decidedly more complex. As Leer (1979: 19) notes,

> Navajo presents a much more diversified and idiosyncratic system of verb stem variation than those of Alaskan languages, almost certainly due to analogical innovation, which is thus quite difficult to analyze synchronically.

Our goal in this paper is to determine whether there are tendencies and patterns in the verb stem morphology of Navajo that may not be apparent on the surface, paying particular attention to the inflectional patterns in the momentaneous aspect. To this end, we submitted the stems to a number of computational analyses. As far as actual processing of the stems is concerned, we suggest that Navajo speakers store all known verbs stems in the lexicon in a network of stems that are interconnected based on their semantic and phonological similarity. Verbs with similar patterns reinforce each other and are available to exert analogical influence on new stems or stems that are unavailable due to imperfect memory. Our findings should be of interest to theoretical linguists as well as to applied linguists since the results have implications for pedagogical grammars of Navajo.

2 Previous Work on Athabaskan Verb Stems

Comparative work on the verbal morphology of Athabaskan languages has shown that verb stems in these languages were originally bimorphemic, composed of a lexical verb root followed by an inflectional suffix carrying modal and/or aspectual information (Hardy 1979; Leer 1979). Over time, the boundary between the root and the suffix weakened and the two elements began to fuse. In some of the Northern Athabaskan languages it is still possible to see evidence of these historical suffixes in the surface forms of the modern languages. However, in Navajo, the process of fusion between the root and the suffix has been carried further so that now there are essentially no readily segmentable modal/aspectual suffixes.

Nonetheless, most previous analyses of the Navajo verb stem system (Pike & Becker 1964; Stanley 1969; Pinnow 1974; Hardy 1979; Rice 1995) have attempted to relate the occurring surface forms to hypothetical underlying forms by positing abstract suffixes and complex rules of concatenation that essentially recapitulate the diachronic processes that are assumed to have given rise to the current system. For instance, Rice (1995) posits that all perfective verb forms in Navajo are formed by the addition of a suffix composed solely of the feature [Sonorant Voice], which is then realized differently in particular phonological environments.

Krauss (1990) follows a similar set of assumptions. In his treatment, he discusses the differences between the highly complex patterns of stem-variation found in Athabaskan and the notably more transparent system that is present in Eyak,[1] where the boundary between the verb root and the inflectional suffix is clear. With respect to the Athabaskan verb stem, he states that

> [t]hrough analysis involving a great deal of abstraction, these underlying or historical elements, no longer segmental, must still account in some sense for the speaker's ability to generate the involved stem variants without memorizing thousands of forms. (Krauss 1990: 154)

In the present paper, our analysis will focus not on deriving the stem forms from abstract underlying stems, but rather on relating the actually occurring stem forms to one another in a direct fashion, employing a usage-based, schematic approach of the type originated in Bybee (1985) and further outlined in work by Langacker (1991), Bybee (1995), and others. Before presenting the details of the analysis of the verb stem it will be useful to look at the structure of the Navajo verb as a whole.

3 Structure of the Navajo Verb

As is well known, the morphological composition of the verb in Navajo is extremely complex (Kari 1976; Young & Morgan 1987; McDonough 1990; Faltz 1998). Structurally, it is composed of a prefix string, which may contain either a single prefix or a conglomeration of many prefixes, and a verb theme. The verb theme in Navajo is itself internally complex. It is composed of a classifier prefix plus a verb stem. There are four classifiers in Navajo, *-ø-*, *-ɬ-*, *-d-*, and *-l-*, typically used to mark notions of transitivity.[2] The verb stem appears on the right of the classifier. Verb stems usually have rather abstract meanings which is why the glosses we provide must be considered only suggestive. Stems are typically of the shape CVC, where V stands for a vowel which is either short or long, low tone or high tone, oral or nasal. Changes in the shape of the verb stem, in combination with the addition of various prefixes mark the numerous categories of mode and aspect on the verb. As an example of what is involved, consider a partial paradigm for the momentaneous aspect and imperfective mode of the verb meaning 'card wool' (examples taken from Faltz 1998: 283). The classifier

[1] Eyak is a sister language to all the Athabaskan languages, being grouped with Athabaskan in the Na-Dene family.

[2] See Kibrik (1993) and (1996) for excellent discussions on the interaction of the classifier prefixes with degrees of transitivity in Athabaskan.

prefix here is *-ɬ-* and the verb stem is *–chaad*. (In Navajo orthography *ch* = /ʧ/, *sh* = /ʃ/, *zh* = /ʒ/, *ɬ* = / ɬ /, ‘ = /ʔ/. High tone is indicated with accentuated vowels, and nasality is marked with a nasal hook beneath a vowel.)

	Singular	**Duoplural**	**Distributed Pl.**
1st person	*hanishchaad*	*haniilchaad*	*hadaniilchaad*
2nd person	*hanółchaad*	*hanołchaad*	*hadanołchaad*
3rd person	*hainółchaad*	*hainółchaad*	*hadeinółhaad*
4th person	*hazhnółchaad*	*hazhnółchaad*	*hadazhnółchaad*

Table 1: Momentaneous Imperfective Forms of ‘Card Wool’

Duoplurals refer to a subject composed of only two members while distributed plurals consist of three or more members. Fourth person refers to indefinite subjects and is also used as a polite form in certain contexts. In Table 2, we present the same verb in the same aspect, but in the perfective mode. Note that in this case, the verb stem is again *-chaad*, with the difference in mode being expressed solely by changes in the prefixes. In fact, some of the forms are identical to their imperfective counterparts.

	Singular	**Duoplural**	**Distributed Pl.**
1st person	*haníółchaad*	*haniilchaad*	*hadaniilchaad*
2nd person	*hanínółchaad*	*hanoołchaad*	*hadanoołchaad*
3rd person	*hainíółchaad*	*hainíółchaad*	*hadaneeshchaad*
4th person	*hazhníółchaad*	*hazhníółchaad*	*hadazhneeshchaad*

Table 2: Momentaneous Perfective Forms of ‘Card Wool’

Lastly, we consider the same verb, again in the same aspect, but this time in the future mode. In this case, not only are there changes in the prefixes, there is also a different verb stem form, *-chał*.

	Singular	**Duoplural**	**Distributed Pl.**
1st person	*hadínéeshchał*	*hadínúilchał*	*hadadínúilchał*
2nd person	*hadíníółchał*	*hadínóołchał*	*hadadínóołchał*
3rd person	*haidínóołchał*	*haidínóołchał*	*hadeidínóołchał*
4th person	*hazhdínóołchał*	*hazhdínóołchał*	*hadazhdínóołchał*

Table 3: Momentaneous Future Forms of ‘Card Wool’

While the details of inflection vary greatly from one class of verbs to the next, the complex changes illustrated by ‘card wool’ are fairly representative of Navajo verbs in general. In the remainder of this paper we restrict

our attention to those changes that affect the verb stem. For a detailed treatment of the verb prefixes see Faltz (1998).

4 The Navajo Verb System

In Young and Morgan (1992), the inflection of each verb stem in the language is indicated by simply listing all the variant shapes that the stem can take, identified with which combination of modal and aspectual features each form expresses. This listing of stem forms is called the stem set. As an example, the stem set for the verb 'have a bad dream' is shown in the Table 4. The aspects appear on the left side of the table and the modes across the top.[3] The modes are imperfective, iterative, perfective, future, and optative. Note also that the forms of usitative mode are the same as the iterative mode; the stem forms of verbs that have progressives are identical to the future which is why these modes are not listed separately.

	IMPF	**ITER**	**PERF**	**FUT**	**OPT**
Momentaneous	*gą́ą́sh*	*gąsh*	*gąązh*	*gąsh*	*gą́ą́sh*
Continuative	*gąąsh*	*gąsh*	*gąązh*	*gąsh*	*gąąsh*
Diversative	*gązh*	*gąsh*	*gązh*	*gąsh*	*gązh*
Repetitive	*gąsh*	---	---	---	---
Semelfactive	*gąsh*	*gąsh*	*gąsh*	*gąsh*	*gąsh*
Neutral	*gąsh*	---	*gąązh*	---	---

Table 4: Stem Set for 'Have a Bad Dream'

Examining this stem set, we note several striking features: (1) Not all combinations of mode and aspect are attested. For instance, the repetitive aspect occurs only with the imperfective mode, and not with any of the other four. (2) There are fewer distinct stem forms than there are combinations of aspect and mode categories. That is, while this verb occurs in a total of 23 aspect and mode combinations, it only has five distinct stem forms (*gąsh, gązh, gąąsh, gąązh, gą́ą́sh*) that are distributed in a seemingly haphazard fashion. The brunt of distinguishing between all the varied combinations of aspects and modes is borne by the aspectual and modal prefixes, not by the verb stem itself, as we saw above in the 'card wool' example. (3) All five stem forms follow the same CVC pattern (where V can be long or short, low-tone or high-tone, nasal or oral). None of the five have any readily segmentable suffix added to them and it is not clear whether any one of the

[3] In this table, and those that follow, we follow the practice of Young & Morgan (1992) in not writing hyphens before the verb stem, even though it is indisputably a bound morpheme.

five forms is in some way more basic, serving as the form from which the other four may be derived.

This simple listing of forms in a stem set begs the question of predictability. That is, do verbs stems in Navajo follow any definable patterns in their mode and aspect inflection, or are the forms essentially random which would require memorization of all forms? Young and Morgan address this issue briefly.

> Although the existence of regular patterns and rules governing the derivation of stems from underlying roots is quite apparent, and although some of them have been formulated, research in stem derivation is far from complete or conclusive as far as the Navajo language is concerned. (1992: 807)

Our goal is to shed light on how predictable the patterns alluded to by Young and Morgan are.

5 Analyses

Given the difficulty of finding broadly applying paradigms within the verb stems we thought it was of interest to probe them computationally. In order to do so, we converted all of the verb stems listed in Young & Morgan's (1992) lengthy Navajo dictionary into computer readable format.[4] Each stem was represented as a series of nine variables. For example, the imperfective momentaneous form of the stem *chozh* to shatter' is *chóósh,* which yielded the following variables:

#	Variable	Example
1	Initial consonant	/tʃ/
2	Initial consonant type	affricate
3	Vowel	/o/
4	Vowel length	long
5	Vowel nasality	oral
6	Vowel tone	high
7	Final consonant:	/ʃ/
8	Final consonant type	fricative
9	Additional vowel after CVC stem	none

Table 5: Example of Variables Used in the Simulations

[4] The database is available at http://linguistics.byu.edu/faculty/eddingtond/navajo

A total of 5,895[5] stem forms from 437 different verbal stem sets were thus encoded, although one must keep in mind that there are not 5,895 unique stem forms; identical stem forms are often used in several different aspects and modes. For instance, the form *chóósh* is not only the imperfective momentaneous form of *chozh*, but the optative momentaneous as well.

Each verb stem contains both lexical and morphological information. The lexical information indicates which stem set the particular stem form belongs to, which in turn allows its meaning to be identified. The morphological information relates to the aspect and mode of the stem. Lexical identification of the stems is carried out mainly by the initial consonant and vowel quality of the stem in spite of a small degree of ablaut and some variation in the initial consonant quality. Morphological information in the stem, on the other hand, is conveyed principally by variations in the final consonant and in the nasality, tone, and length of the stem's vowel. The variations in the stem work in tandem with prefix combinations, but the morphological role of the prefixes is beyond the scope of our present work.

5.1 Data Mining

In order to gain insight into the patterns among the verb stems we examined these data using a machine learning program called C4.5 (Quinlan 1993). The program is designed to 'mine' a data set and find all possible generalizations or relationships. Data mining programs such as C4.5 have many uses. For example, large medical databases contain information about patients' symptoms, blood pathology, etc., along with the diagnosis of each patient. These databases are mined in order to discover what combination of symptoms is most likely to indicate which malady. In political polling, the political leanings of individuals may be predicted based on combinations of spending habits, educational level, television programs viewed, brand of wine recently purchased, etc. Data mining techniques provide the information on which such predictions are based.

The first task we had C4.5 perform was to find any generalizations about the phonological make-up of stems with different aspects. Table 6 summarizes our findings. Keep in mind that there are 5,890 stem forms and 437 different stem sets. The program assigns momentaneous as the default that applies to all of the cases not specified by these eight generalizations. This default is not surprising since 35% of the stems in the database have

[5] The aspect database contains only 5,890 forms, while there are 5,895 in the mode database. This is because we chose not to include the 5 forms of the verb stem *t'óód*[2] which is the only verb that appears in the dictionary with the persistive aspect. The oddity of this particular aspect is highlighted by the authors of the dictionary themselves (Young & Morgan 1992) who mark it with a question mark.

the momentaneous aspect. In all, these rules and the default yield a 35.9% overall success rate. This low rate should not be surprising because if significant patterns existed they would have been described in previous analyses.

The outcome of the program may be read as an if/then statement. For example, the first line states that if a stem has a high tone and ends in /n/, then it is predicted to have neutral aspect. This pattern holds true in 100% of the cases, but is not a very general statement since it only applies to four stems. The second generalization is much more broad. There are 1414 stems with long high vowels. The generalization states that these stems have momentaneous aspect, although it is only correct in about half the cases. It misapplies in 709 of the cases for a misapplication rate of 50.1%.

Vowel Length	Vowel Orality	Vowel Tone	Final Cons.	Pred. Aspect	# Appl.	# Misappl.	% Correct
--	--	high	n	Neutral	4	0	100
long	--	high	--	Moment.	1414	709	49.9
short	oral	high	l z ʒ ʔ	Neutral	3	0	100
short	nasal	low	h	Semelf.	41	25	39.0
long	--	low	ʃ	Contin.	66	37	43.9
short	--	low	ø	Contin.	81	53	34.6
short	--	low	l	Contin.	100	66	34.0
long	nasal	--	--	Moment.	415	175	57.8

Table 6: Generalizations Found by C4.5 for Navajo Aspect.

We applied the same program to predict the mode of the stems (Table 7). These predictions together achieve an overall accuracy of 43.9%, but only when the iterative mode is assumed to apply to any instances not covered by these generalizations. The first row states that if a stem has a long vowel and ends in either /s/ or /ʃ/, then its mode is optative, although the prediction is borne out only in 41.3% of the instances and misapplied 58.7 percent of the time. The second generalization is noteworthy because it does not entail a combination of variables but a single consonant. About 48% of the stems ending in the lateral fricative /ɬ/ are future. If one interprets this as a future suffix, it would be the only clearly segmentable morpheme found in the verb stems keeping in mind that stem forms of verbs that have progressives are identical to the future.

Vowel Length	Vowel Orality	Vowel Tone	Final Cons.	Pred. Mode	# Appl.	# Mis-appl.	% Correct
long	--	--	s ʃ	Optative	344	202	41.3
--	--	--	ɬ	Future	1103	574	48.0
short	oral	high	d ʔ ø l ʒ z n	Imperf.	65	24	63.1
long	oral	--	h s ʃ	Imperf.	629	391	37.8
--	oral	low	d n	Imperf.	405	265	34.6
long	--	--	d ʔ ø l ʒ z n	Perfect.	707	218	69.2
--	--	--	d ø l ʒ z	Perfect.	426	194	54.5
long	nasal	--	h	Iterative	337	176	47.8
--	--	low	h	Iterative	752	496	34.0

Table 7. Generalizations Found by C4.5 for Navajo Mode.

While it is of interest to find broad correspondences between verbs stem features and the modes and aspects in general, identifying patterns within a particular mode and aspect combination is more germane to the search for paradigms within the verb stems. Analyzing each and every combination is beyond the scope of our study. What we want to demonstrate is that such generalizations exist and may be arrived at computationally. However, we did examine the most frequent aspect, the momentaneous, which yielded the generalizations in Table 8. The analysis correctly accounts for 52.8% of the cases when the imperfective is considered the default. Once again /ɬ/ emerges as a possible future morpheme.

Vowel Length	Vowel Orality	Vowel Tone	Final Cons.	Pred. Mode	# Appl.	# Mis-appl.	% Correct
--	**--**	--	l z ø ʒ	Perfect.	193	28	85.5
long	**--**	--	ʔ	Perfect.	100	28	72.0
--	**--**	low	d ʔ ø l ʒ z n	Perfect.	277	140	49.5
long	oral	**--**	h s ʃ	Imperf.	421	229	45.6
long	**--**	**--**	s ʃ	Imperf.	38	22	42.1
short	**--**	**--**	h s ʃ	Iterative	397	223	43.8
long	nasal	**--**	h	Iterative	196	109	44.4
--	**--**	**--**	ɬ	Future	384	156	59.4

Table 8. Generalizations Found by C4.5 for Navajo Momentaneous Forms.

Although these patterns only account for a small portion of the verb stems, we trust that these generalizations will prove useful to the commu-

nity of Athabaskan scholars and teachers, however we do not assume that Navajo speakers need to glean this kind of information from the language data in order to process their language. Furthermore, we note that generalizations for only four of the 13 possible aspects, and for only five of the six modes were found. The success rates of the simulations also seem artificially inflated by the algorithm's broad application of the default. In any event, Navajo speakers surely have some sort of system for representing the entire modal and aspectual system of their language which is what we turn our attention to now.

5.2.1 Predicting Stem Forms by Analogy

A full account of all the inflectional intricacies of the Navajo verbal lexicon remains to be fleshed out. However, we would like to examine the stems from a framework that emphasizes a dynamic view of the lexicon that focuses on the relationship between surface forms that are stored in the lexicon rather than on rules that generate surface forms from underlying roots. Krauss (1990) argues that speakers must possess some sort of stem inflection rules in order to produce the myriad of Navajo verb forms. However, the striking lack of generality such rules would have makes it seem much more likely that speakers have memorized the actual stem forms themselves, much as Young and Morgan portray them in their stem sets. We would like to explore the idea that speakers do store every stem they have knowledge of. However, we assume that the stems are not stored in isolation from each other but are crucially linked through a massive network of connections. Following Bybee (1985, 1988, 2001), stems would be linked on the basis of semantic and phonological similarities.

As an example, consider the following four stems:

(1)

nééh	the imperfective momentaneous form of the verb stem *ná* 'migrate'
né	the imperfective diversative form of the verb stem *ná*
nééɬ	the optative momentaneous form of the verb stem *ná*
nééh	the optative conclusive form of the verb stem *nah* 'forget'

In terms of semantics, the first three forms are linked because they are all forms of the same verb stem *ná* 'migrate'. The first two also have the imperfective mode in common, while the first and third share the momentaneous aspect. All of the forms would be associated because they share the same initial consonant and have the vowel /e/ with a high tone. The three forms with the long vowels would be linked in regards to that feature as well. The two forms of *nééh* are linked in regards to their phonological

overlap, but have no semantic associations. This sort of network among stored items allows relationships between items to be identified. For example, the relationship between *nééɬ* and *né* is such that *né* has no final consonant in contrast to the /ɬ/ of its partner; it also has a short /e/ instead of a long /e/. We speak of relationships rather than derivations since we assume that one stem form is not derived from another nor from a separate underlying form.

Now, imagine that a speaker hears a hypothetical new word containing the optative momentaneous root *tłééɬ* and wants to use the stem in the imperfective diversative,[6] but s/he is not familiar with that particular form of the stem. According to analogy, the speaker would utilize the commonalities that *tłééɬ* has with other verb stems of the language in order to find the closest verb stem to *tłééɬ*. If the most similar stem to *tłééɬ* were *nééɬ*, the speaker would have a model relationship upon which to determine the imperfective diversative stem related to *tłééɬ*. That relationship is the one that holds between the optative momentanous *nééɬ* and its imperfective diversative counterpart *né*. Therefore, the speaker could predict that the diversative form is *tłá* by performing a simple proportional analogy:

(2) *nééɬ* is to *né* as
tłééɬ is to ? (= *tłé)*

This analogical process need not apply only to new or unknown forms but also in cases in which imperfect memory or noise in the system impedes access to a known form. Our goal is to determine the extent to which Navajo verbs stems are predictable from other surface forms by analogy. Our analysis emphasizes analogy based on phonemic and morphological properties, but we assume that other semantic properties that we have not identified come into play as well.

5.2.2 Analogical Algorithm

At this point it is necessary to present the model we will apply to the task at hand. Analogical models, which are also known as exemplar or memory-based models, have been applied to investigate a wide variety of linguistic phenomena such as word recognition (Goldinger 1996), Arabic and German plural formation (Nakisa, Plunkett, & Hahn 2001), Spanish stress and gender assignment (Eddington 2000, 2002a), linking elements in Dutch noun

[6] The imperfective describes and uncompleted action such as the English present tense. "The diversative describes movement 'here and there' among something, roaming or wandering, or a process" (Young & Morgan 1992: 871).

compounds (Krott et al. 2002), phonological alternations in Turkish stems (Rytting 2000), Dutch stress assignment (Gillis et al. 1993), Italian verb conjugations (Eddington 2002b), and phonotactic knowledge in Arabic and English (Frisch et al. 2001). A number of analogical models have been developed: Nosofsky's Generalized Contextual Model (Nosofsky 1990), Pierrehumbert's exemplar model (Pierrehumbert 2001), and Analogical Modeling of Language (Skousen 1989, 1992). We have chosen to utilize the Tilburg Memory-based Learner (TiMBL; Daelemans et al. 2001) since it has a good track record for dealing with morphological variation of the type we are covering, is readily available, and has ample documentation and instructions for its use.

TiMBL works by taking an input and calculating which items in the database of exemplars are the most similar to the input. These are known as the nearest neighbors of the input. To illustrate how this works, consider the task of predicting the past tense form of an English verb from its present tense form. The database would contain present tense forms along with a specification of the relationship between each past and present form. The relationship between *sing* and *sang,* for example, is that the vowel is /i/ in the present tense and /æ/ in the past. During the training session, series of variables that represent instances of forms are stored in memory along with their behavior, so *sing* could be represented by the phonemic variables /s, I, ŋ/. In the case that the same verb is encountered more than once in the database, a count is kept of how often each stem form is associated with a given relationship to another stem. In addition to storing instances and counting duplicates, the extent to which each variable helps predict the correct outcomes is used in the calculation of similarity. This is known as information gain.

During the testing phase, when an input such as *sing* is given to the program, the algorithm searches for *sing* in the database. If *sing* is found it is given the behavior that it has been assigned in the majority of cases (i.e. the vowel is /i/ in the present tense and /æ/ in the past.) If the item is not found in the database, a similarity algorithm is used to find the most similar stem(s)–its nearest neighbor(s). *Ring* could presumably be the nearest neighbor of *sing*. The behavior of the nearest neighbor(s) is then applied to the stem in question, so *ring* is to *rang* as *sing* is to ___.[7] If two or more items are equidistant from the stem in question the most frequent behavior of the tied items is applied to the stem in question. In the algorithm we use in our simulations the similarity between the values of a variable is precalculated and used to adjust the search for nearest neighbors accordingly. This

[7] We assume proportional analogy here, but non-proportional analogies are also possible (see Eddington 2004: 78; Skousen 2002: 42-43).

precalculation allows certain values of the variables to be regarded as more similar to each other than other values.

5.2.3 Simulations Using One Stem to Predict Another

We chose the momentaneous aspect for the Navajo simulations since the 2,073 momentaneous forms comprise about 35% of all verb stems, and most stem sets have momentaneous verb stems. The momentaneous verbs stems appeared in the database as a series of nine variables as described in section 5. In addition, one variable was included that specified the relationship between the two stem forms in question.

Consider the following two entries from the database used to predict the imperfective stem from the future stem.

(3) k o e l o h h f 0 same
(4) k o o s o l s f 0 long-e

In (3), the future momentaneous 'become confused' is represented by variables indicating that it begins with k which is an obstruent, that is followed by the vowel /e/, that is a long, oral, high tone vowel. The stem ends in the consonant /h/ that is a fricative, and no (0) other vowel appears after the final consonant. The relationship between these future and imperfective momentaneous stem forms is that they are the same. In (4), the future stem *kos* of the set stem *kéézˇ* 'to cough' is represented as a series of variables in like manner. However, the imperfective form *kees* differs from *kos* which is specified by the final variable which in essence states that the two stems are identical except that *kees* has a long /e/ in contrast to the short /o/ of *kos*.

The verb stems of each of the five modes were used to predict the momentaneous form of the remaining four modes. This entailed running 20 separate simulations. In each simulation, the verb stems were removed from the database one at a time, which allowed each to serve as the test item. Analogies were drawn from the remaining database items. An example of how this was done is in order. Consider the first simulation which entailed predicting the imperfective form based on the future form. First, a future stem form was selected as the test item and the algorithm calculated the single most similar future stem to it in the database. This stem is known as the test form's nearest neighbor (k=1). We also calculated using the three nearest neighbors (k=3). Once the nearest neighbor was found the relationship that that particular future stem bears to its imperfective forms was applied to the test stem in order to predict the imperfective form of the test item. For example, the nearest neighbor of the future stem *kos* is the future stem *koh* whose imperfective is *kóóh*. Therefore, by analogy the imperfec-

tive form corresponding to *kos* is incorrectly predicted to be **kóós* rather than the correct *kees* (*koh* is to *kóóh* as *kos* is to ? = *kóós*). The success rates for these 20 simulations are summarized in Table 9.

Using A to predict B	**k=1**	**k=3**	**Using A to predict B**	**k=1**	**k=3**
A Future B Imperfective	51.9	55.4	A Iterative B Future	**93.3**	**93.3**
A Future B Optative	43.6	44.1	A Iterative B Imperfective	52.4	55.7
A Future B Perfective	32.2	31.8	A Iterative B Optative	44.7	43.7
A Future B Iterative	83.0	**85.3**	A Iterative B Perfective	38.1	40.3
Average	52.7	54.2	**Average**	57.1	58.3

Using A to predict B	**k=1**	**k=3**	**Using A to predict B**	**k=1**	**k=3**
A Imperfective B Future	73.9	73.7	A Perfective B Future	**70.0**	66.4
A Imperfective B Optative	**77.7**	76.6	A Perfective B Imperfective	53.0	55.8
A Imperfective B Perfective	34.5	39.8	A Perfective B Optative	46.0	49.5
A Imperfective B Iterative	69.0	71.0	A Perfective B Iterative	67.4	67.4
Average	63.8	65.3	**Average**	59.1	59.8
A Optative B Future	71.5	74.3			
A Optative B Perfective	38.4	44.5			
A Optative B Imperfective	84.0	**86.5**			
A Optative B Iterative	69.3	72.0			
Average	65.8	69.3			

Table 9: Success Rates of the First Simulation

The individual success rates range from a low of 31.8% to a high of 93.3%. The average ability of each mode to predict the other four modes ranges from 52.7% to 69.3%. The best overall predictor of the form of the other four forms is the optative; the worst is the future. On the other hand, the future is most easily predicted by the other four modes. That is, the average ability of the other four modes to correctly predict the future is 77.2%

at k=1. This is followed by the iterative (72.0%), the imperfective (60.3%), and the optative (53.0%). The most difficult to predict is the perfective (35.8%).

Keep in mind that some verb stems have two (or more) possible stem forms. For example, the momentaneous perfective of the verb *tłah* is either *tłah* or *tła*. The perfective mode contained more doublets of this sort than the other modes. If each of these doublets appeared in the database, when one of the doublets is correctly predicted that necessarily entails that the other would not be. In order to eliminate this possibility we only included the first member of the doublet that appeared in Young & Morgan (1992). In this way the predictions for a mode with many doublets could be compared with modes with few doublets. However, we did include doublets when there were alternative forms for the entire paradigm.

One potential difficulty we see with using one member of a paradigm to predict the form of another is that paradigmatic relationships involve more than one member; therefore, the first simulations may have missed significant analogies. In order to remedy the situation we carried out another set of simulations.

5.2.4 Simulations Using a Paradigm to Predict a Single Verb Stem

In the previous simulations, the nearest neighbor(s) of a single verb stem, such as a future stem whose imperfective form we want to predict, was found in the database. The relationship between the future stem of the nearest neighbor and its imperfective stem, for instance, was used to predict the form of the unknown imperfective. Consider the momentaneous paradigm of the verb stem *ji'* whose imperfective, iterative, perfective, future, and optative forms are:

(5)	IMPF	ITER	PERF	FUT	OPT
	jiih	*jih*	*ji'*	*jih*	*jiih*

In the second set of simulations the goal is to predict the form of one stem by drawing analogies based on the remaining four stems. In order to do so, databases were created that contained the variables of four of the forms, but not of the stem form to be predicted. For example, to predict the future form the database would contain information about the other four stems forms:

(6)	IMPF	ITER	PERF	FUT	OPT
	jiih	*jih*	*ji'*	*?*	*jiih*

The idea here is that speakers may be familiar with all of the members of a paradigm except one. Therefore, they will base their prediction of the stem shape of the missing member by analogy to the paradigm that is found to be most similar to the paradigm they know.

One difficulty that these databases pose is the issue of deciding which relationship to use. When the relationship of one stem is based on the other, as in the first simulations, there is only one relationship on which to analogize. Recall the earlier example in which the hypothetical verb stem *tłééł* is found to be most similar to the stem *nééł*. Since the relationship between *nééł* and *né* is known (i.e., *né* has no final consonant in contrast to the /ł/ of its partner; it also has a short vowel instead of a long one), the relationship can be extended by analogy to *tłééł* predicting *tłé*. When analogies are drawn on four of the members of a paradigm there are four possible relationships. Rather than agonize over which to use, we considered all four possible relationships. Once again 20 separate simulations were run, the results of which appear in Table 10.

Predict the **Future** basing the relationship on the:		
	k=1	**k=3**
Imperfective	81.6	81.4
Perfective	63.8	67.5
Iterative	**93.1**	92.6
Optative	76.9	78.7
Average	78.9	80.1

Predict the **Iterative** basing the relationship on the		
	k=1	**k=3**
Imperfective	78.7	79.7
Perfective	62.5	64.8
Future	**93.8**	93.3
Optative	75.7	75.4
Average	77.7	78.3

Predict the **Imperfective** basing the relationship on the:		
	k=1	**k=3**
Future	**86.8**	**86.8**
Perfective	61.3	68.2
Iterative	81.9	83.1
Optative	**86.8**	85.6
Average	79.2	80.9

Predict the **Perfective** basing the relationship on the:		
	k=1	**k=3**
Imperfective	58.6	62.3
Future	64.3	64.5
Iterative	63.0	**65.8**
Optative	57.6	63.8
Average	60.9	64.1

Predict the **Optative** basing the relationship on the:		
	k=1	**k=3**
Imperfective	**85.6**	85.1
Perfective	55.8	65.2
Iterative	78.1	78.7
Future	80.1	84.4
Average	74.9	78.4

Table 10: Success Rates of the Second Simulation.

The individual success rates on these simulations ranges from a low of 55.8% to a high of 93.8% which is a substantial increase from the 31.8% to 93.3% range obtained in the previous simulations that utilized only one form. The average success rate by mode in the present simulations ranges from 60.9% to 80.9% which is also an improvement from the 52.7% to 69.3% range obtained in the earlier simulations. This demonstrates that when more members of a paradigm are considered more accurate predictions about the phonological shape of the missing member may be made.

6 Conclusions

The goal of this paper has been to examine the Navajo verb stems by computational means in order to discover patterns that may not be outwardly apparent. In accordance with what other researchers have observed, we found the Navajo verb stems to contain very little in the way of overt paradigms, although we were successful in identifying a small number of patterns which are summarized in Tables 6-8. However, the lack of widely-applying surface-apparent paradigms leads us to believe that each stem is memorized and stored in the mental lexicon with connections to other stems based on phonological and semantic similarity. Our simulations of the momentanous aspect suggest that if Navajo speakers analogize on memorized stem sets, the phonological shape of a new or unavailable form may be correctly predicted in about 75% of the cases. It is important to reiterate that the shape of a verb stem works in tandem with the prefixes in setting out the intended mode and aspect. Nevertheless, our simulations suggest that there are important paradigmatic patterns among the verb stems that may be utilized. The analogical simulations we carried out do not specifically indicate what those patterns are, but further analysis along the lines of Lachler (2000) should ultimately identify them.

References

Axelrod, M. 1993. *The Semantics of Time: Aspectual Categorization in Koyukon Athabaskan*. Lincoln: University of Nebraska Press.

Bybee, J. 1985. *Morphology: A Study of the Relation between Meaning and Form*. Amsterdam. John Benjamins.

Bybee, J. 1988. Morphology as Lexical Organization. *Theoretical Approaches to Morphology,* eds. M. Hammond & M. Noonan, 119-41. San Diego: Academic Press.

Bybee, J. 1995. Regular Morphology and the Lexicon. *Language and Cognitive Processes* 10: 425-55.

Bybee, J. 2001. *Phonology and Language Use*. Cambridge: Cambridge University Press.

Bybee, J. & D. Slobin. 1982. Rules and Schemas in the Development and Use of the English Past Tense. *Language* 58: 265-289

Daelemans, W., J. Zavrel, K. van der Sloot, & A. van den Bosch. 2001. TiMBL: *Tilburg Memory Based Learner, Version 4.1, Reference Guide. Induction of Linguistic Knowledge Technical Report.* Tilburg, Netherlands: ILK Research Group, Tilburg University. (http://ilk.kub.nl/).

Eddington D. 2000. Spanish Stress Assignment within the Analogical Modeling of Language. *Language* 76: 92-109.

Eddington, D. 2002a. Dissociation in Italian Conjugations: A Single-Route Account. *Brain and Language* 81: 291-302.

Eddington, D. 2002b. Spanish Gender Assignment in an Analogical Framework. *Journal of Quantitative Linguistics* 9: 49-75.

Eddington, D. 2004. *Spanish Phonology and Morphology: Experimental and Quantitative Perspectives.* Amsterdam: John Benjamins.

Faltz, L. M. 1998. *The Navajo Verb: A Grammar for Students and Scholars*. Albuquerque: University of New Mexico Press.

Frisch, S.A., N.R. Large, B. Zawaydeh & D.B. Pisoni. 2001. Emergent Phonotactic Generalizations in English and Arabic. *Frequency and the Emergence of Linguistic Structure*, eds. J. Bybee & P. Hooper, 159-179. Amsterdam: John Benjamins.

Gillis, S., W. Daelemans, G. Durieux & A. van den Bosch, 1993. Learnability and Markedness: Dutch Stress Assignment. *Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society*, 452-457. Hillsdale, N.J.: Erlbaum

Goldinger, S.D. 1996. Words and Voices: Episodic Traces in Spoken Word Identification and Recognition in Memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 22: 1166-1183.

Goosen, I.W. 1995. *Diné Bizaad: Speak, Read, Write Navajo*. Flagstaff, AZ: Salina Bookshelf.

Hardy, F. 1979. *Navajo Aspectual Verb Stem Variation*. PhD dissertation, University of New Mexico.

Kari, J. 1976. *Navajo Verb Prefix Phonology*. New York: Garland Publishing.

Kibrik, A.A. 1993. Transitivity Increase in Athabaskan Languages. *Causatives and Transitivity*, eds. B. Comrie & M. Polinsky, 47-67. Amsterdam: John Benjamins.

Kibrik, A.A. 1996. Transitivity Decrease in Navajo and Athabaskan. *Athabaskan Language Studies: Essays in Honor of Robert W. Young*, eds. E. Jelinkek, S. Midgete, K. Rice, & L. Saxon, 259-304. Albuquerque: University of New Mexico Press.

Krott, A., R. Schreuder. & R.H. Baayen. 2002. Analogical Hierarchy: Exemplar-Based Modeling of Linkers in Dutch Noun-Noun Compounds. *Analogical Modeling: An Exemplar-based Approach to Language*, eds. R. Skousen, D. Lonsdale, & D. B. Parkinson, 181-206. Amsterdam: John Benjamins.

Krauss, M. 1990. Typology and Change in Alaskan Languages. *Language Typology 1987: Systematic Balance in Language*, ed. W. Lehmann, 147-156. John Benjamins.

Lachler, J. 2000. Verb Stem Ablaut in Navajo: A Regular Irregularity. *Proceedings of the 1999 Mid-America Linguistics Conference*, ed. M. T. Henderson, 241-251. Lawrence, KS: University of Kansas Linguistics Department.

Langacker, R.W. 1991. *Concept, Image, and Symbol: The Cognitive Basis of Grammar*. Berlin: Mouton de Gruyter.

Leer, J. 1979. *Proto-Athabaskan Verb Stem Variation*. Part One: *Phonology*. Fairbanks: Alaska Native Language Center.

McDonough, J. 1990. *Topics in the Phonology and Morphology of Navajo Verbs*. PhD Diss., University of Massachusetts.

Nakisa, R.C., K. Plunkett & U. Hahn. 2000. Single- and Dual-Route Models of Inflectional Morphology. *Models of Language Acquisition: Inductive and Deductive Approaches*, eds. P. Broeder & J. Murre, 201-222. Oxford: Oxford University Press.

Nosofsky, R.M. 1990. Relations between Exemplar Similarity and Likelihood Models of Classification. *Journal of Mathematical Psychology* 34: 393-418.

Pierrehumbert, J. 2001. Exemplar Dynamics: Word Frequency, Lenition and Contrast. *Frequency and the Emergence of Linguistic Structure*, eds. J. Bybee and P. Hooper, 137-158. Amsterdam: John Benjamins.

Pike, K. L. & A. L. Becker. 1964. Progressive Neutralization in Dimensions of Navajo Stem Matrices. *International Journal of American Linguistics* 30: 144-154.

Pinnow, H.J. 1974. *Studie zur Verbstammvariation im Navaho*. Berlin: Indiana.

Quinlan, J.R. 1993. *C4.5: Programs for Machine Learning*. San Mateo CA: Morgan Kaufmann.

Rice, K. 1989. *A Grammar of Slave*. Berlin: Mouton de Gruyter.

Rice, K. 1995. The Representation of the Perfective Suffix in the Athapaskan Language Family. *International Journal of American Linguistics* 61: 1-37.

Rytting, C.A. 2000. An Empirical Test of Analogical Modeling: The /k/ ~ /ø/ Alternation. *Lacus Forum XVII: The Lexicon*, eds. A.K. Melby & A.R. Lommel, 73-84. Fullerton, CA: Linguistic Association of Canada and the United States.

Skousen, R. 1989. *Analogical Modeling of Language*. Dordrecht: Kluwer.

Skousen, R. 1992. *Analogy and Structure*. Dordrecht: Kluwer.

Skousen, Royal. 2002. Issues in Analogical Modeling. *Analogical Modeling: An Exemplar-based Approach to Language,* eds. R. Skousen, D. Lonsdale, & D.B. Parkinson, 27-48. Amsterdam: John Benjamins.

Stanley, R. 1969. *The Phonology of the Navajo Verb*. PhD dissertation, MIT.

Young, R. & W. Morgan. 1987. *The Navajo Language: A Grammar and Colloquial Dictionary*. 2nd ed. Albuquerque: University of New Mexico Press.

Young, R. & W. Morgan. 1992. *Analytical Lexicon of Navajo*. Albuquerque: University of New Mexico Press.