Root and Morpheme Frequency and Dispersion in Ichishkíin Legends

Keegan Livermore

LING 593

Fall 2020

University of Oregon

In order to design authentic, intentional language curricula, vocabulary, linguistic functions, and possible contexts need to be imagined for what type of engagement the learners will have during the course. A language corpus can provide all of those and more through statistical analyses and easy access to instances and examples. For low-resource languages, however, there may be limited materials to fill that corpus, be they digitized or even physical. Many endangered language communities often need to work with the materials on hand to glean enough to help create or supplement their class materials. For Yakima Ichishkíin, an indigenous language spoken in current central Washington state, one of the primary sets of materials that has permeated most language classes thus far are tribal legends. These represent a limited number of registers and genres, creating potential issues about the representative reach of the corpus.

Another major challenge with these materials, however, lies in the highly polysynthetic nature of the language, as there are many ways that they could be parsed, tagged, and processed in order to yield helpful products for future language projects. Many of the current corpus tools have built their tokenization procedures off of more isolating, higher-resource languages, so there may need to be different considerations to allow for an opportunity or revitalization corpus to still process data in a meaningful manner. This project will explore strategies for using materials from polysynthetic language contexts and consider ways to use lexeme and morpheme occurrence data to support classroom curriculum development.

Literature Review

Corpus Building in a Low-Resource/Revitalization Language Context

It is difficult to escape the perspective and priorities of prototypical corpora when approaching a revitalization/documentation language context. Cox (2011) points out that the goals of corpus linguistics and language documentation can align on many fronts because of their shared goals of rigorous storage

and study of language materials. There are concerns of representativeness of the language, a balance of genres, and a focus on data from authentic communicative settings (Vinogradov, 2016). This is a major area of concern for groups working to build corpora for low-resource and low-density languages, because there is not often such a plethora of options available for study. Because of this, there are often concerns about what types of research are possible because of deficiencies in terms of some of these key factors that support solid linguistic analysis and description. The limitations of the genres present based on available digital files as well as authentic, natural sources of language input can pose problems when trying to determine what direction to start moving in. Rice and Thunder (2017) give examples of using planned traditional stories and personal narratives, edited written stories, and some unplanned conversations to try to get enough variety and access to different forms. Vinogradov (2016) points out that many smaller corpora often do not have the luxury of rejecting material that may be too similar to others already in the corpus, as the size will help guarantee a sufficient base for any analytic conclusions based on the data.

Community capacity building is often intertwined with these documentation and corpus goals in order to effectively gather enough materials for an opportunity corpus while honoring the different stakeholder relationships one needs for intensive language work. Rice and Thunder (2017) point out that any community member is able to help document materials through all of the technology at their fingertips in the modern age. This both takes advantage of accessing speakers wherever they are at as well as establishing more community stakeholders who can provide input into what types of language they see around them as important for documentation. Elicitations can be helpful in getting at specific language forms or functions that may be hard to isolate, but even a smattering of files from multiple speakers can provide compelling initial results. While archives are not the end goal of this type of work, they do help provide a jumping off point towards annotation and further research after some processing (Cox, 2011). Although the statistic verifications based on size and instance quantity may not be as

encouraging compared to prototypical, high-density languages, worthwhile results can be gathered from initial analyses that help the future community and research projects. Klavans (2018) points out, however, that there should not be a disconnect between the analytic side and the community-facing side of these projects, as they often can and should inform each other about how they can go about doing their work and making sure it follows the tangible path laid about by the community planners and activists.

Corpus Design Issues for Polysynthetic Languages

For languages with rich morphology systems to add and modify root meanings, automatic parsing may be difficult. Arkhangelskiy and Lander (2016) describe the difficulties of identifying base forms of roots and affixes from derived surface forms that may be adapting to their local environments. While some programs have tried to work from reference grammars to reproduce patterns seen in previous analyses, there will inevitably be errors that lead to incorrect glosses and tags. Often times, additional elicitations or manual data parsing will be needed in order to inform the model how to more accurately and wholly parse particular morphemes away from roots (Schreiner et al., 2020). However, this kind of interlinear glossing practice can be used to inform the model to then later inform on-the-spot glossing and provide some checks to the system with language informants present. There may be issues with the level of documentation on some morphemes and phenomena or variation in orthography or phonemes that may not have been accounted for yet in that refinement process.

Another question for polysynthetic languages that many have considered is the level of detail in the root-, stem-, morpheme-, and word-level tags. The normal annotation schemes for prototypical corpora with part of speech tags are not helpful to languages that stack meaning via morphemes onto roots (Rice & Thunder, 2017). A major issue with this stacking is when the “part of speech” of the derived stem is different than the original root’s, creating a conversation about how many layers of

meaning is needed to store within the annotations. There may also be difficulties in assigning a lemma to a stem that has significant derivational affixes, as the stem could be considered a lemma on its own versus assigning the lemma to the root's meaning (Arkhangelskiy & Lander, 2016). There may be a purpose in designing and assigning more general categories of tags, depending on how efficient and conservative the corpus wants to be (Galves et al., 2017). There are models that have packed tags into their data, but it depends on what types of information the analysis methods intend to pull out for efficiency's sake. Many low-resource polysynthetic languages start out trying to annotate whole word meaning, as that could be considered the most immediate meaning accessible within the syntax structure.

Pedagogical and Teaching Purposes for Low-Resource/Revitalization Corpora

One of the biggest reasons for building digital language resources is for the larger community of learners to have access to authentic supplementary materials. Many of these languages are not spoken socially or in public spaces as often as English or other dominant languages, so language input needs to come from other places. Lots of this accessibility manifests as open portals, often web-based, to allow for dictionary and grammar searches by learners who are still acclimating to new words or patterns (Cox, 2011; Rice & Thunder, 2017). While it may yet be seen or understand exactly how these portals are utilized by the larger community, teachers and higher-proficiency speakers may be able to navigate them with more fluidity and confidence

Some corpora are being developed specifically to support the development of curricula and lesson plans for classroom instruction. Gatbonton et al. (2015) describe their process of focusing on teaching target expressions from elders that can be used in conversation and trying to build elicitation and corpus processing procedures that would yield helpful results for the classroom purposes. In their analysis, they were looking for specific polysynthetic "n-gram" expressions that could essentially help

populate their unit lists to help build useful utterance chunks for learners. Metadata tags had to be reconsidered in order to provide context about the types of questions that elicited specific responses. Although they were not aiming at creating a full corpus in the image of a prototypical corpus, they were successful in still extracting collocates and frequencies of particular lexemes and morphemes, both of which they did not have to work with before (Rice & Thunder, 2017; Vinogradov, 2016). For teaching purposes, new insights from these types of data can help produce new pathways to language engagement.

Research Questions

This project examined the contents of five Yakima Ichishkíin legends with the intent to more analytically describe the lexical contents to better support curriculum development around these existing materials. The research questions were:

1. What are the most common roots and morphological affixes across the five Yakima Ichishkíin legends?
2. How can these roots and affixes be grouped linguistically and categorically by semantic domain?

These two inquiries will provide better estimates of possible lexeme lists that learners should be either expected to know or be encouraged to learn prior to engagement with the legend for better comprehension of the text.

Methodology

Corpus Design

Due to Ichishkíiníin's status as a language in revitalization, there are few centralized digitized materials that are already formatted linguistically to use as the corpus of study. A major part of this

project was conceptualizing the different possibilities for both a working corpus format for the research questions in focus as well as a more multi-purpose corpus format for future analysis on a variety of research topics. Materials that were available from L1 speakers included a handful of documented legends, a memoir text with both personal stories and descriptions of cultural activities, and some procedural texts. Three factors were considered in the selection process: the registers (and consequent language varieties) represented, the current utility of those texts in current Ichishkíin language classrooms, and the familiarity of the analyst with the texts. The legends were selected for this project as an opportunity corpus because those are the most integrated into current curricula and were therefore also the most known to the analysts. They are:

- Ḵw'ashḵw'ashyáy (The Legend of Crane)
- Sɬkni (Yellowbell)
- Spilyáy ku Ḵ'áx̱nu (Coyote and Prairie Chicken)
- Wax̱púuya ku Asumyáy (Rattlesnake and Eel)
- Łkw'i ku Sts'at (Day and Night)

These legends include both narration and conversational speech, providing multiple registers albeit without a sense of their relative proportions. They will also be in a slightly more formal register and focused on third person participants, so they would not be able to serve as a true corpus for general Ichishkíin language usage. However, they will be a great start and be able to at least provide insight into narration and storytelling language usage.

All of the legend texts were already digitized, but were presented in their original, unparsed form with stylized full-sentence English translations. In order to make them machine readable using Python, there had to be linguistic analysis integrated into the text in order to process the texts into meaningful parts for analysis. There were conversations with colleagues about how much information

would be most helpful to the project and desired types of analyses with these texts. Ichishkíin's high degree of polysynthetic structure means that significant energy and time would need to go into preparing the texts for the most ideal analysis. It was agreed that full parsing and glosses would be most useful in the long-term in order to best approach any analysis, but with the time available this quarter, full morphological parsing would fulfill the identification needs alongside personal linguistic and declarative knowledge for further analysis. The texts were edited to include specific number codes at all morpheme boundaries depending on what kind of division it was: a prefix, a verb root, a suffix, or an enclitic. These insertions were done by hand using personal knowledge of many of the derivational and inflection morphemes and what has been considered a standalone lexeme in most cases. Once these five texts were coded as such, they were put through a parsing script that split up the words at those boundaries and inserted appropriate symbols to briefly annotate what type of morpheme it was: a "+" for prefixes and suffixes, a "-" for verb roots, and a "=" for enclitics. These were then reprinted individually by file in-line for future analysis and prototype archival and distribution purposes.

Analysis Procedures

These parsed and coded text files were then read back into Python in order to generate frequency counts for each unique morpheme and lexeme. This was done at two levels: combining the counts for all five text files into a master list as well as comparing the frequency counts between the legends to look at their overall distribution. This was done to better understand the utility of the lexemes semantically across multiple contexts and to eliminate the possibility of a lexeme having a high frequency and being marked important despite only being present in one or two of the five legends under analysis. It was important to consider how useful these lexemes and morphemes would be to the learners from a long-term perspective if they are to be able to work with all five texts over the course of their learning. A list of total unique tokens was gathered between the five legends, and then was used to gather the individual frequency counts for each token across the legend set. A table was generated that

listed the lexeme/morpheme, its part of morphology (e.g. root or affix), and then the respective frequency counts as well as a count of how many legends had instances of that particular token. The list was finalized by identifying all orthographic lexemes that occurred in at least two legends. This may have left out some instances if they had an allomorphic surface form or different stress marking due to a stress-changing affix in the original data, but without additional processing time or a lemma dictionary file, correcting this would have been too time-intensive for this test drive analysis. Lexemes and morphemes remaining were split into a list of root lexemes and a list of affix morphemes.

At this stage, the linguistic analysis was able to happen by hand. Glosses and categories relied on learner-speaker and linguistic intuition as well as senses of how vocabulary items could be grouped together based on behavior or meaning. Glosses were assigned to all items on the lists, and then a version of the conceptual category was assigned based on that gloss and the other glosses on the list. The first round focused more on part of speech for the list of roots (nouns, descriptors, verbs, meta-language, etc.) while the list of affixes focused more on general function type (person, derivational, case, etc.). After this first pass, the list of roots was then subdivided into more semantic categories (people, environment, etc.). These sets of categories were then examined across the five legends to look at how many unique tokens in these categories occurred with respect to how many legends it was present in as well as how many total instances of each category were present in the set of legends.

Results

With the five texts parsed morphologically, there were approximately 3900 tokens to compile and analyze. The only annotations present were the symbols used to indicate what type of morpheme the instance was (root, prefix, affix, or enclitic). Part of the parsing process was also generating parsed files in both in-line and vertical formats for archival and easier access for future analysis. These files were put through the dispersion calculator script, yielding counts of the type of unique affix instances as

well as of the overall token instances across all five legends with respect to how many legends each unique token is in.

Unique Count	# of Legends					
Morpheme Type	1	2	3	4	5	Total
affix	70	29	17	13	10	139
root	591	89	31	4	9	724
Total	661	118	48	17	19	863

Table 1: Counts of Unique Token Roots and Affixes across 5 Ichishkíin Legends

Sum Totals	# of Legends					
Morpheme Type	1	2	3	4	5	Total
affix	110	137	172	418	851	1688
root	1001	371	271	65	500	2208
Total	1111	508	443	483	1351	3896

Table 2: Sum Counts of Token Roots and Affixes across 5 Ichishkíin Legends

Looking at the Counts of Unique Tokens in Table 1, approximately 84% of the tokens are roots, either noun, verb, or descriptor of some kind. There is a more fixed set of affixes and enclitics that are used as well as the corpus source is comprised of 5 legends that all focus on different characters and activities, so there is a wide range of lexemes. Additionally, 81.6% of those roots occur in only a single legend, meaning that significant vocabulary studying would be needed in order to fully prepare for the first reading and comprehension experience. Despite all of this, roots make up only 56.6% of all tokens in the current version of the corpus, which reflects the highly polysynthetic and agglutinative natures of the language. Comparing these numbers puts into context where the most fruitful analytic work will be regarding the content: on the roots and affixes that are in more than one legend (in their current forms). This accounts for 202 unique tokens, approximately a quarter of the overall list, as well as 2,785 of the overall token count, approximately 71% of all tokens present. While this may not provide all of the lexical details that a reader would need when reading the legends, they would be able to get the grammatical and narrative gist of the piece.

Root Morphemes

The first pass at describing the different unique root types focused on the grammatical category, which could almost be considered a part of speech. The "description" label includes both adjective, adverb, and descriptive nominal lexemes. The "meta" label includes demonstratives, non-"action" verbs (e.g. "to become" and the copula *wa*), and grammatical words (e.g. "maybe", "must", and question words). Pronoun lexemes were pulled out separately because they are optional inclusions within the grammar that can help clarify the scenario of the event or disambiguate the meaning of a particular clause. The "vocalization" found is an "aah" sound that was transcribed into the legends and thus included in the counts.

Count of Roots	# of Legends				
Root Category	2	3	4	5	Total
description	11	3	1	0	15
meta	16	14	5	8	43
noun	14	5	1	0	20
pronoun	2	2	0	0	4
verb	23	3	3	0	29
vocalizations	0	1	0	0	1
Total	66	28	10	8	112

Table 3: Counts of Unique Root Instances in their Grammatical Categories in at least 2 Ichishkíin Legends

The majority of the root types are in the "meta" category and are in two or three of the five legends. This list includes many of the demonstratives that are individually marked for prepositional case. This list of "meta" roots does still have some commonalities across all five legends, so it is worthwhile. There are also a few nouns, verbs, and descriptors that are shared between more than two legends, showing how different these five texts are from each other content-wise. Thus, there could be a shorter list of "essential" lexemes and morphemes that could be analyzed, but including the full 112 allows for some more content areas to be present.

Sum of Roots	# of Legends				
Root Category	2	3	4	5	Total

description	42	11	27	0	80
meta	51	108	105	540	804
noun	84	34	5	0	123
pronoun	4	33	0	0	37
verb	69	32	39	0	140
vocalizations	0	12	0	0	12
Total	250	230	176	540	1196

Table 4: Sum Counts of Root Instances in their Grammatical Categories in at least 2 Ichishkíin Legends

While the "meta" category still leads the number of overall instances, it can be seen that approximately 20% of the instances would be categorized as nouns and verbs. This might suggest that the lexical content is concentrated and/other so particularized to each legend that there would need to be specialized vocabulary lists for each legend in use already. Comparing this with the statistic that two-thirds of all roots in this set fall into the "meta" category suggests that the framing and narrative conventions may be more helpful to study explicitly. The low pronoun usage reflects that they are optional grammatically and therefore may not be completely necessary to study at the beginning to understand the legends' contents.

These categories are broken down further into more content-related categories in Tables 5 and 6. These category distinctions were made personally by the analyst based on their own language intuitions and experiences with language curriculum development. Full lists of these tokens can be found in the Appendices with their corresponding category descriptions.

Count of Roots	# of Legends				
Content Category	2	3	4	5	Total
activity	6	0	1	0	7
adjective	6	1	1	0	8
being	2	0	1	0	3
body	1	0	0	0	1
character	5	1	0	0	6
environment	3	0	0	0	3
helpful	8	12	4	8	32
living	9	1	1	0	11
location	1	0	0	0	1

movement	8	0	1	0	9
number	1	1	0	0	2
pronoun	2	2	0	0	4
second-helpful	8	5	1	0	14
state	3	1	0	0	4
thing	3	3	0	0	6
verbalization	0	1	0	0	1
Total	66	28	10	8	112

Table 5: Counts of Unique Root Token Type by Content Category

The category with the most unique token types is the "helpful" category, which includes the copula forms, nominative and locative demonstrative items, and question words (e.g. "when", "where", "how"), among other items. The "second-helpful" category includes what could be considered more "intermediate" items, including more prepositional case-marked demonstratives and question meanings. Most of the other categories of root items occur in only two legends out of the five, so these would be the beginnings of unit vocabulary lists that additional items could be added to in order to flesh them out sufficiently.

Sum of Roots	# of Legends				
Content Category	2	3	4	5	Total
activity	19	0	12	0	31
adjective	26	3	27	0	56
being	15	0	5	0	20
body	7	0	0	0	7
character	36	8	0	0	44
environment	15	0	0	0	15
helpful	29	110	90	540	769
living	25	17	6	0	48
location	5	0	0	0	5
movement	25	0	21	0	46
number	2	3	0	0	5
pronoun	4	33	0	0	37
second-helpful	22	24	15	0	61
state	9	5	0	0	14
thing	11	15	0	0	26
verbalization	0	12	0	0	12
Total	250	230	176	540	1196

Table 6: Sum Count of Instances of Root Tokens by Content Category

Seeing that the category with the most overall instances of that type of token is the “helpful” category affirms many of those decisions based on anecdotal usage patterns. The next categories that stepped up are the “adjective”, “character”, “living”, and “movement”, many of which feel like they could be commonly used in curricula under a variety of contexts and settings. These categories overall could provide some guidance about the direction to take assigning root vocabulary to new language learners.

Affix Morphemes

Dividing the affix morphemes into categories is much easier as linguistic and morphological categories can be used as labels. The “meta” category here includes only an enclitic that means “I wonder,” which is currently understudied morphologically and syntactically. These categories overall do count surface forms of these morphemes as separate unique tokens, so they do not compile allomorphs into one gloss that is counted.

Count of Affixes	# of Legends				
Affix Category	2	3	4	5	Total
case	1	2	3	1	7
derivational	6	4	1	1	12
directional	3	2	0	1	6
meta	1	0	0	0	1
number	2	0	2	0	4
person	3	3	4	2	12
suffix	3	1	1	0	5
TAM	3	1	1	4	9
transitivity	1	3	3	1	8
Total	23	16	15	10	64

Table 7: Counts of Unique Affix Tokens in their Grammatical Categories

These are more spread throughout the legends, as they are foundational to the meaning created through the polysynthesis. The required morphology, specifically in the person, transitivity, tense/aspect, and case markers are present in a majority of the legends, cementing them as key language function skills that need to be acquired for communication and storytelling.

Sum of Affixes	# of Legends				
Affix Category	2	3	4	5	Total
case	2	20	53	20	95
derivational	26	27	24	16	93
directional	20	13	0	18	51
meta	6	0	0	0	6
number	8	0	55	0	63
person	23	27	145	325	520
suffix	9	12	22	0	43
TAM	8	16	109	393	526
transitivity	2	44	56	79	181
Total	104	159	464	851	1578

Table 8: Sum Counts of Affix Instances in their Grammatical Categories

Looking at the sum counts of all affix instances in Table 8, the person and tense/aspect markers are required for all verbs, so they are understandably each a third of the total affix instances. Transitivity marking stands out as a key function skill, which can require intensive focus and study to learn the paradigms of marking speech act and non-local participants around verbs. The other two relative standouts in terms of overall number of instances are the different case markers and derivational morphemes, both of which modify lexemes to fit into the intended meaning of the utterance. These are not used as proportionally as the regular grammatical relation and verbal morphology, but should still be noted as being relatively frequent for their categories.

Discussion

Word Lists & Impact on Curricula

The produced lists of words shared by multiple legends appear to cover several important bases and concepts of language that beginning and intermediate learners would need to know as a foundation for everyday use. The categories of roots cover personal description, environmental and living conditions, physical and emotional description, and metalanguage markers that help connect utterances and ideas together. While these categories were grouped together by a curriculum developer, this is but

one way that they could be arranged, depending on the learning environment. The dominant presence of those “meta” markers (e.g. “again”, “must”, “maybe”, negation) suggests that teachers may need to think about how to introduce and practice those markers to help learners structure their discourse more effectively. The affixes were grouped more linguistically with less subjective decisions made about how a particular affix could be considered. These factors of frequency and dispersion across texts seem to be good indicators of potential importance when overlaid with each other. A next step would be to compare them to existing curriculum and word lists to both assess where there might be differences in linguistic or content priorities from personal needs assessment or anecdotal evidence.

In terms of the method of deriving the word lists, this process seems to be more of a deep-dive than many other polysynthetic language communities have attempted. Many analysis teams left words unparsed as versions of complex stems, n-grams, or collocates to still work within their analyses to collect a different type of data. Some were working on automatic parsers based on provided grammatical data, which seems promising depending on how much variation is present within the language’s grammar and phonology. This study was focused more on lexeme frequency than trying to extract common utterance constructions or examples, perhaps from more of an assessment or evaluation perspective. These lists are absolutely more data than the Ichishkíin curriculum-designing community had before, so this should be considered a success on that front.

Limitations

One of the biggest limitations in this study, however, was the time available for the preparation and analysis of available materials. Parsing was the course of action chosen for this project, as instance counts were integral to the assessment of types of functions and content expressed. However, without the use of tools like FLEX or other automatic parsers, this work had to be done by hand. This put several limitations on what types of text could be used for these purposes, as it would have been more ideal

that the linguists were already familiar with the texts enough to be able to recognize the morphemes and roots on sight rather than needing to translate everything for the first time. Alongside this manual parsing, the glossing had to occur by hand. Because of the surprising large quantity of tokens, it was decided that the glossing would happen in preparation for the report writing rather than in the raw data stage. This does mean that one needs to understand enough Ichishkíin on their own before they can truly analyze the data.

Next Steps

The small number of materials available could be considered a limitation, but the Ichishkíin research community has learned how to dig further into the things they already have to find new insights and meanings. With these initial corpus building tools in hand, more computer applications and programs are needed to help parse, gloss, and prepare the corpus for future analysis. Tools like FLEX will be able to help expedite these tasks in order to start extracting and analyzing phenomena that we haven't gotten to yet linguistically. The corpus will allow us to sample from our multiple sources more easily and create more continuity amongst the resources for our language community.

References

Arkhangelskiy, T. A., & Lander, Yu. A. (2016). Developing a polysynthetic language corpus: Problems and solutions. *Компьютерная Лингвистика и Интеллектуальные Технологии*, *15*(22), 40–49. https://publications.hse.ru/en/articles/184655926

Cox, C. (2011). Corpus linguistics and language documentation: Challenges for collaboration. In *Corpus-based Studies in Language Use, Language Learning, and Language Documentation* (pp. 239–264). Brill Rodopi. https://doi.org/10.1163/9789401206884_013

Galves, C., Sandalo, F., Sena, T. A. de, & Veronesi, L. (2017). Annotating a polysynthetic language: From Portuguese to Kadiwéu. *Cadernos de Estudos Lingüísticos*, *59*(3), 631–648. https://doi.org/10.20396/cel.v59i3.8651003

Gatbonton, E., Pelczer, I., Cook, C., Venkatesh, V., Nochasak, C., & Andersen, H. (2015). A Pedagogical Corpus to Support a Language Teaching Curriculum to Revitalize an Endangered Language: The Case of Labrador Inuttitut. *International Journal of Computer-Assisted Language Learning and Teaching*, *5*(4), 16–36. https://doi.org/10.4018/IJCALLT.2015100102

Klavans, J. L. (2018). Computational Challenges for Polysynthetic Languages. *Proceedings of the Workshop on Computational Modeling of Polysynthetic Languages*, 1–11. https://www.aclweb.org/anthology/W18-4801

Rice, S., & Thunder, D. (2017). *Community-based corpus-building: Three case studies*. http://scholarspace.manoa.hawaii.edu/handle/10125/42052

Schreiner, S. L. R., Schwartz, L., Hunt, B., & Chen, E. (2020). Multidirectional leveraging for computational morphology and language documentation and revitalization. *Language Documentation & Conservation*, *14*, 69–86. http://scholarspace.manoa.hawaii.edu/handle/10125/24917

Vinogradov, I. (2016). Linguistic corpora of understudied languages: Do they make sense? *Káñina*, *40*(1), 127–130. https://doi.org/10.15517/rk.v40i1.24143

Appendix A: Root Frequency & Dispersion List

morpheme	Gloss	Category	Semantic	ḵw'ashḵw'ashyáy	spilyáy	sɬkni	waxpush	ɬkw'i	Total	counter
wáḵ'ish	alive	description	adj	4	1	0	0	0	5	2
ayáyat	beautiful	description	adj	7	0	5	0	0	12	2
nch'i	big	description	adj	1	0	0	1	0	2	2
káakɬm	full	description	adj	1	1	0	0	0	2	2
shíx	good	description	adj	20	4	2	1	0	27	4
iksíks	little	description	adj	1	1	0	1	0	3	3
ḵ'ínupa	looking	description	adj	0	1	2	0	0	3	2
k'aywá	short	description	adj	1	0	0	1	0	2	2
wíyat	far	description	location	3	2	0	0	0	5	2
naxsh	one	description	number	0	1	1	0	1	3	3
nápu	two.people	description	number	1	0	0	0	1	2	2
k'asáwi-	be.cold	description	state	1	0	0	1	0	2	2
anáwi-	be.hungry	description	state	0	2	0	1	0	3	2
payúwi-	be.sick	description	state	3	0	0	0	1	4	2
shaláwi-	be.tired	description	state	2	1	0	0	2	5	3
ánach'axi	again	meta	helpful	0	3	2	0	0	5	2
tɬ'áax̱w	all	meta	helpful	5	14	5	0	2	26	4
míimi	already	meta	helpful	1	0	1	0	1	3	3
táaminwa	always	meta	helpful	6	1	3	0	0	10	3
kú	and	meta	helpful	63	32	23	17	7	142	5
wá-	COP	meta	helpful	45	18	8	2	5	78	5
wachá-	COP.PAST	meta	helpful	13	5	3	2	2	25	5
tx̱ána-	happen	meta	helpful	6	2	1	0	1	10	4
paysh	maybe	meta	helpful	7	1	0	3	0	11	3
cháw	NEG	meta	helpful	30	14	6	2	1	53	5
áw	now	meta	helpful	25	13	12	9	2	61	5

mish	Q	meta	helpful	6	5	0	1	0	12	3
íx̱wi	still/later	meta	helpful	2	3	6	0	0	11	3
íkw'ak	that	meta	helpful	7	0	2	2	0	11	3
kwnák	that.LOC	meta	helpful	17	9	1	2	0	29	4
íkwnak	that.LOC	meta	helpful	2	1	0	0	0	3	2
awkú	then	meta	helpful	76	30	27	2	9	144	5
íchi	this	meta	helpful	9	7	8	1	0	25	4
íchna	THIS.loc	meta	helpful	3	2	0	2	0	7	3
íkush	thus	meta	helpful	11	6	3	2	2	24	5
páyu	very	meta	helpful	5	0	0	1	0	6	2
tun	what	meta	helpful	3	3	2	0	0	8	3
anakú	when	meta	helpful	6	1	1	3	2	13	5
mun	when	meta	helpful	0	3	1	0	0	4	2
mɨnán	where	meta	helpful	3	7	1	0	0	11	3
miin	where.ALL	meta	helpful	0	1	0	1	0	2	2
shin	who	meta	helpful	1	0	0	0	1	2	2
ii	yes	meta	helpful	3	0	0	1	0	4	2
iiii	yes	meta	helpful	2	0	1	0	0	3	2
kw'áx̱i	again	meta	second	0	2	1	0	0	3	2
kuuk	at.then	meta	second	2	1	0	0	0	3	2
ának	later	meta	second	2	0	1	1	0	4	3
laak	might	meta	second	1	1	0	0	2	4	3
huuy	must	meta	second	4	1	1	0	0	6	3
míshkin	Q.INST	meta	second	1	0	0	1	0	2	2
íkwaal	so.long	meta	second	1	0	1	0	0	2	2
ana	SUB	meta	second	12	1	1	0	1	15	4
íkuuni	that.ALL	meta	second	4	2	0	1	0	7	3
íkwɨn	that.DAT	meta	second	3	1	0	0	0	4	2
kunkínk	that.INST	meta	second	1	0	0	0	1	2	2
kwmak	that.PL	meta	second	2	1	0	0	0	3	2

túkin	what.INST	meta	second	1	0	2	0	0	3	2
shiin	who.OBJ	meta	second	1	1	0	1	0	3	3
aah	verbal	n/a	verbalization	3	7	2	0	0	12	3
kákya	bird	noun	being	3	0	0	0	1	4	2
myánash	child	noun	being	2	9	0	0	0	11	2
tíin	Indian.person	noun	being	2	0	1	1	1	5	4
tpɨsh	face	noun	body	6	0	1	0	0	7	2
spilyáy	Coyote	noun	character	1	6	0	0	1	8	3
pát	older.sister	noun	character	3	0	3	0	0	6	2
yanwáy	poor.thing	noun	character	1	0	0	1	0	2	2
tmáy	umarried.girl	noun	character	10	0	2	0	0	12	2
áyat	woman	noun	character	6	1	0	0	0	7	2
láymut	youngest.girl	noun	character	7	0	2	0	0	9	2
tiichám	land	noun	environment	4	0	2	0	0	6	2
ámchnik	outside	noun	environment	2	0	0	1	0	3	2
wána	river	noun	environment	0	3	0	3	0	6	2
ikúuk	today	noun	helpful	5	2	4	0	0	11	3
tkwátat	food	noun	thing	0	3	1	0	0	4	2
watít	legend	noun	thing	2	1	0	0	0	3	2
táatpas	shirt	noun	thing	3	0	1	0	0	4	2
áan	sun	noun	thing	1	0	3	1	0	5	3
chíish	water	noun	thing	2	0	1	1	0	4	3
anwíkt	year	noun	thing	0	1	0	1	4	6	3
niimí	1PL.GEN.PN	Pronoun	pronoun	1	0	1	0	0	2	2
ínk	1SG.NOM.PN	Pronoun	pronoun	8	4	0	1	0	13	3
imk	2SG.NOM.PN	Pronoun	pronoun	1	0	0	1	0	2	2
pɨnk	3SG.NOM.PN	Pronoun	pronoun	17	2	1	0	0	20	3
wiláalakw-	compete	verb	activity	0	0	0	2	1	3	2
tíwi-	fight	verb	activity	1	1	0	0	0	2	2
iɬamayk-	lose	verb	activity	0	1	2	0	0	3	2

itax̲shi-	make.wake.up	verb	activity	3	0	2	0	0	5	2
sáyp-	serve	verb	activity	0	2	0	1	0	3	2
pnú-	sleep	verb	activity	3	1	5	0	3	12	4
táx̲shi-	wake.up	verb	activity	0	1	2	0	0	3	2
kú-	do	verb	helpful	7	2	0	1	0	10	3
mí-	do	verb	helpful	1	3	0	1	0	5	3
háashhaash-	breathe	verb	living	0	0	1	0	1	2	2
náx̲ti-	cry	verb	living	2	1	0	0	0	3	2
wɨshúwa-	get.ready	verb	living	0	0	3	1	0	4	2
wɨ́np-	grab	verb	living	1	1	0	0	0	2	2
yík-	listen	verb	living	2	1	0	0	0	3	2
tk'í-	look.at	verb	living	1	1	3	1	0	6	4
waník-	name	verb	living	1	0	1	0	0	2	2
ƛ̲'ínu-	see	verb	living	0	1	2	0	0	3	2
ɨn-	tell	verb	living	10	3	0	4	0	17	3
px̲wí-	think	verb	living	2	0	1	0	0	3	2
tƛ̲'íx̲-	want	verb	living	2	0	0	0	1	3	2
wyáalakw-	abandon	verb	movement	1	2	0	0	0	3	2
wyánawi-	arrive	verb	movement	2	1	0	0	0	3	2
wína-	go	verb	movement	12	2	0	6	1	21	4
winá-	go	verb	movement	1	0	0	0	1	2	2
at-	go.out	verb	movement	0	2	2	0	0	4	2
wíihayk-	go.up	verb	movement	0	1	1	0	0	2	2
túx̲-	return	verb	movement	3	1	0	0	0	4	2
wináchik-	ride.in	verb	movement	1	0	3	0	0	4	2
kwíita-	walk.by	verb	movement	1	2	0	0	0	3	2

Appendix B: Affix Frequency & Dispersion List

morpheme	Gloss	Category	ɣw'ashɣw'ashyáy	spilyáy	sɬkni	waɣpush	ɬkw'i	Total	counter
'+knik	ABL	case	4	5	0	2	2	13	4
'+kan	ALL	case	1	2	0	2	0	5	3
'+ay	BEN	case	0	0	0	1	1	2	2
'+yaw	DAT	case	8	3	2	6	1	20	5
'+mí	GEN	case	7	6	2	0	0	15	3
'+ki	INST	case	1	2	1	0	1	5	4
'+pa	LOC	case	23	6	2	4	0	35	4
'+ɬá	AGT	derivational	3	1	0	0	0	4	2
'+ani	APPL	derivational	2	1	1	0	0	4	3
'+nani	APPL	derivational	2	2	0	0	0	4	2
shapá+	CAUS	derivational	3	4	1	0	0	8	3
'+wát'a	DES	derivational	1	0	0	0	2	3	2
wíi+	go	derivational	1	0	0	1	0	2	2
'+t	NLZR	derivational	7	3	2	2	2	16	5
pápa+	RECIP	derivational	2	0	0	1	3	6	3
piná+	REFL	derivational	11	2	10	1	0	24	4
'+ni	STAT	derivational	7	0	1	0	1	9	3
'+yi	STAT	derivational	1	3	0	0	0	4	2
'+tɣaw	superlative	derivational	6	3	0	0	0	9	2
'+m	CSL	directional	8	6	2	1	1	18	5
'+ɨm	CSL	directional	8	0	4	0	0	12	2
'+úu	DIR	directional	4	3	2	0	0	9	3
'+shamsh	IMPV.CSL	directional	1	1	0	0	0	2	2
'+ɨnkik	TSL	directional	1	1	2	0	0	4	3
'+kik	TSL	directional	5	0	0	1	0	6	2
'=ɣash	I.wonder	meta	2	4	0	0	0	6	2

'+in	ASSOC.DUAL	number	6	5	2	1	0	14	4
'+yin	ASSOC.DUAL	number	1	0	0	4	0	5	2
'+ma	PL	number	21	14	5	0	1	41	4
'+íma	PL	number	2	0	1	0	0	3	2
'=natash	1PL.EXCL	person	5	0	1	0	0	6	2
'=na	1PL.INCL	person	10	4	2	0	1	17	4
'=nash	1SG	person	17	10	1	5	0	33	4
'=ish	1SG	person	2	1	1	0	0	4	3
'=sh	1SG	person	6	1	0	3	0	10	3
'=ash	1SG	person	6	3	0	0	0	9	2
'=pam	2PL	person	5	2	6	0	0	13	3
'=nam	2SG	person	32	14	1	7	0	54	4
'=am	2SG	person	5	3	0	0	0	8	2
pa+	3PL.S	person	38	18	13	3	9	81	5
i+	3SG.S	person	121	47	36	25	15	244	5
pá+	INV	person	20	10	3	8	0	41	4
'+tya	actually	suffix	3	1	0	0	0	4	2
'+ch'a	also	suffix	13	1	7	1	0	22	4
'+x̲i	also	suffix	1	1	0	0	0	2	2
'+k'a	intensifier	suffix	9	0	0	2	1	12	3
'+sim	only	suffix	1	2	0	0	0	3	2
'+tax̲nay	COND	TAM	2	0	0	0	1	3	2
'+ta	FUT	TAM	61	18	10	13	9	111	5
'+x̲a	HAB	TAM	10	4	2	0	0	16	3
'+k	IMPER	TAM	2	0	1	0	0	3	2
'+na	PAST	TAM	48	26	22	21	10	127	5
'+ya	PAST	TAM	17	10	2	3	3	35	5
'+a	PAST	TAM	50	32	16	11	0	109	4
'+sh	PERF	TAM	1	0	0	0	1	2	2
'+sha	PROG	TAM	55	28	23	12	2	120	5

'=matash	1>2	transitivity	2	3	1	0	0	6	3
'=mash	1SG>2SG	transitivity	3	6	1	2	0	12	4
á+	3O	transitivity	48	24	4	2	1	79	5
áw+	3O	transitivity	10	3	3	0	0	16	3
'=pat	3PL>3	transitivity	13	7	2	0	0	22	3
'+nɨm	ERG	transitivity	1	0	1	0	0	2	2
'+maman	OBJ.PL	transitivity	5	6	2	0	1	14	4
'+nan	OBJ.SG	transitivity	15	3	1	11	0	30	4