Syntactic Annotation of a Hupa Text Corpus[1]

Justin Spence[*], Zoey Liu[*], Kayla Palakurthy[†], and Tyler Lee-Wynant[*]
[*]University of California, Davis and [†]University of California, Santa Barbara

1. Introduction

As usage-based approaches to understanding synchronic and diachronic phenomena have become increasingly prevalent in the field of linguistics (Bybee and Beckner 2010), searchable annotated text corpora now play a central role as the empirical foundation upon which language description and linguistic theory are built. However, the vast majority of corpus-based research focuses on a handful of languages with large quantities of pre-existing machine-readable text data. This raises the unfortunate possibility that insights derived from these methods will be limited in typological scope and cross-linguistic validity, making it imperative to develop similar corpora for less-studied languages as well. Searchable electronic corpora are already being developed for a variety of less-studied languages, including several Native American languages (e.g., Nordhoff, Tuttle, and Lovick 2016, Taff 2013, Garrett 2011, Garrett et al. 2013), and text corpora have figured prominently in published studies of Dene languages in particular (Holton and Lovick 2008, Berez and Gries 2010, Lovick and Tuttle 2012). The goals of these corpus development projects vary greatly depending on the specific research interests of the people involved in their creation, and so, accordingly, do database formats and annotation schemes. As similar projects are undertaken for increasing numbers of Dene and other less-studied languages, it is worthwhile to consider the details of how particular corpora are constructed in order to highlight the strengths and limitations of their design and implementation.

The present paper reports on efforts to enhance an existing text corpus of Hupa, a Dene language of northwestern California, with annotations that will facilitate syntactic analysis. In recent years, detailed analysis of Hupa texts has yielded insights into various syntactic and discourse phenomena (Spence 2008, Newbold 2010, Newbold and Escamilla 2012, Spence 2013:157-161), but these studies have been based largely on opportunistic samples of the available documentation of the language. Each researcher invested significant energy into locating tokens of relevant phenomena before analysis could proceed, and the resulting ad hoc collation and coding of data has not, for the most part, been transferable to subsequent research efforts. While an electronic corpus of Hupa with a robust lexical concordance now exists to facilitate some aspects of this kind of work (described in §2 below), queries must still proceed on an item-by-item basis – an improvement over traditional paper-and-pencil methods for linguistic analysis of text material, but still somewhat cumbersome. The current project aims to enhance this corpus with explicit syntactic annotations, yielding a treebank that will facilitate more efficient and exhaustive data exploration and analysis moving forward.

Since this work is ongoing and the annotations we have implemented thus far are still largely provisional, the primary aim of this paper is to highlight methodological aspects of the work that might be relevant to others who have undertaken (or are considering) similar endeavors: the

[1] We wish to thank audiences at the 2017 UC Davis Symposium on Language Research and the 2017 Dene Languages Conference for helpful comments on this research, which has been supported by an NSF Documenting Endangered Languages grant (BCS#1500851).

syntactic framework and annotation scheme we have adopted, our annotation procedures and workflows, difficulties we've encountered in implementation, and solutions to some of the most pressing problems. In §2, we provide an overview of the Hupa language, available sources of text documentation, and the existing electronic corpus that is the basis for syntactic annotation. §3 focuses on the details of the Universal Dependencies annotation framework and aspects of our procedures and workflow that highlight some of the strengths and limitations of our approach and illustrate how similar efforts could be developed for other Dene languages. It is important to emphasize that syntactic annotation per se is not the goal. Rather, it is a means to develop a better understanding of syntactic phenomena in Hupa, hopefully in a way that is revealing both for traditional academically-oriented research and for language revitalization efforts in the contemporary Hupa community. Accordingly, despite the provisional nature of some of the specific analytic decisions, in §4 we also present some preliminary results that address clause-level constituent order, building on previous research by Newbold (2010), by way of illustrating ways that the annotated corpus can be applied to specific research questions. §5 concludes, summarizing the main points and suggesting directions for future research.

2. Hupa Texts

2.1. Text Documentation

Hupa is a Dene language of northwestern California, traditionally spoken in Hoopa Valley on the lower Trinity River in present-day Humboldt County, with closely related dialects known as Chilula and Whilkut on nearby Redwood Creek and surrounding areas. Hupa is by far the most extensively documented California Dene language, including a large quantity of text material summarized in Table 1. The earliest attested Hupa text is found in Jeremiah Curtin's unpublished field notes from the late 1880s, now archived at the National Anthropological Archives. Pliny Earle Goddard worked on the language in the first decade of the 20th century, publishing texts both for Hupa (1904, 1911) and Chilula (1914); additional unpublished text material can be found in field notes archived at the American Philosophical Society. Edward Sapir transcribed a large number of texts in the summer of 1927 which were eventually published several decades later thanks to the efforts of Victor Golla and Sean O'Neill (Sapir and Golla 2001). Due to their reliability and accessibility, the Sapir texts have played a key role in most recent studies of Hupa grammar, including the syntactic annotation project reported here. Later in the 20th century, Mary Woodward, Victor Golla, and Sean O'Neill all collected Hupa text material: audio recordings and/or transcriptions are archived at the Survey of California and Other Indian Languages at UC Berkeley. Some of this material is accessible through the California Language Archive web portal; Golla's transcriptions were published (1984). Since the early 2000s, Hupa elder Verdena Parker has produced a large number of recordings of Hupa texts, some of them on her own, others in collaboration with members of the Hoopa Valley Tribe or with linguists working under the auspices of the Hupa Language Documentation Project (HLDP) originally based at UC Berkeley. Some of these recordings are archived at SCOIL and accessible through CLA, with additional deposits from the HLDP due in the near future under the current grant-funded effort.

Researcher	Date	Published	Archive	Catalog Numbers
Curtin	1888-1889	no	NAA	NAA MS 2063
Goddard	1900s	yes	APS	Na20a.1, Na20a.2 (Hupa) Na20g.1, Na20g.2 (Chilula)
			Bancroft	CU 23-1 12(2)
Sapir	1927	yes	APS	Na20a.4
Woodward	1953	no	SCOIL	Woodward.002
Golla	1963	yes	SCOIL	LA 119
O'Neill	1990s	no	SCOIL	(in progress)
Parker	2003-	no	SCOIL	LA 256

Table 1: Summary of Hupa Text Documentation[2]

2.2. *Corpus*

As is commonly the case for the documentation of Native American languages collected over several decades by researchers with different levels of training and ability, the text documentation summarized in Table 1 is extremely eclectic. Each researcher used his or her own idiosyncratic transcription system, and the original materials vary greatly in quality. Audio recordings include both analog and digital media of various sorts. One of the primary goals of the HLDP has been to assemble these diverse strands of Hupa text documentation and compile them into a single resource. Under development since 2008, the corpus has now grown to over 36,000 glossed units (single words or multi-word expressions) from over a century of documentation: three published collections (Goddard 1904, Sapir and Golla 2001, Golla 1984) and transcriptions of field recordings created with Mrs. Parker, all normalized to a unified practical orthography. The corpus is fully concordanced with an associated lexical database based originally on a learner-oriented print dictionary (Golla 1996), but expanded significantly with new entries harvested from the text corpus and detailed verb paradigms elicited in consultation with Mrs. Parker. Both the lexicon and text corpus can be searched through a single interface, the Hupa Online Dictionary and Texts website.[3] The website consists of XML backend databases and a PHP search interface. The HLDP has recently converted the text corpus to a new

[2] NAA = National Anthropological Archives; APS = American Philosophical Society; SCOIL = Survey of California and Other Indian Languages

[3] http://nalc.ucdavis.edu/hupa/hupa-lexicon.html

XML schema largely conforming to Text Encoding Initiative (TEI) standards (Text Encoding Initiative Consortium 2015; cf. Czaykowska-Higgins, Holmes, and Kell 2014). A robust, flexible, and widely-adopted standard for electronic text corpora, TEI offers ready-made solutions to a number of common encoding problems and provides long-term stability by ensuring that structural information encoded in the database will be interpretable well into the future.

Syntactic relations are straightforwardly represented in the TEI schema as annotations at the word level in the XML hierarchy. Further details about the implementation of these annotations are provided below; the important point here is that we are annotating a text database that already has a well-defined format and structure, so there is less flexibility in terms of how to implement the annotation framework than we would have had starting from scratch. This is appropriate considering that the text corpus is intended to encode more than just syntactic information, but it does create practical problems since the database must be converted to a different format in order to take advantage of syntactic annotation tools that have been developed for corpora conforming to different standards.

3. Syntactic Annotation

The goal of the annotation project outlined here is to develop a treebank of Hupa, a set of sentences where syntactic relationships are represented explicitly in order to facilitate subsequent data exploration and analysis. In this section we outline some of the details of how we have approached this effort, both at the conceptual/theoretical level and in the details of implementation. In making these decisions explicit and transparent, we hope to facilitate dialogue with others engaged in similar activities with an eye towards collaborative efforts to share data in collaborative projects moving forward.

3.1. Framework

In constructing the Hupa treebank, we have adopted a dependency grammar approach to representing syntactic relationships, which has historical origins in the early work of Tesnière (1959). This was desirable for several reasons. First, a dependency grammar treebank was already under development for Karuk, another Native American language of northwestern California (Garrett et al. 2013, Mikkelsen 2015). The Karuk text database uses an XML format whose organization is similar to the Hupa database (not accidentally, since both were developed by researchers currently or formerly affiliated with the Survey of California and Other Indian Languages at UC Berkeley). In the short term, this made it easy to adapt analogous components of their annotation procedures to ours. In the long term, it has the potential to facilitate direct comparison of syntactic relationships in the two languages - although Karuk is not a Dene language, language contact effects in grammatical and functional domains have been a topic of interest for languages of the region (Conathan 2004, O'Neill 2008). Moreover, the Karuk treebank project faces similar kinds of analytic problems as the Hupa project does: as Jurafsky and Martin (2017, ch. 14) point out, dependency grammar is well suited to analyzing languages like Hupa and Karuk with relatively free word order (discussed in §4). The surface orientation of dependency grammar is also advantageous (Garrett et al. 2013), since it does not require positing phonologically null elements or underlying word orders whose empirical justification is not yet established.

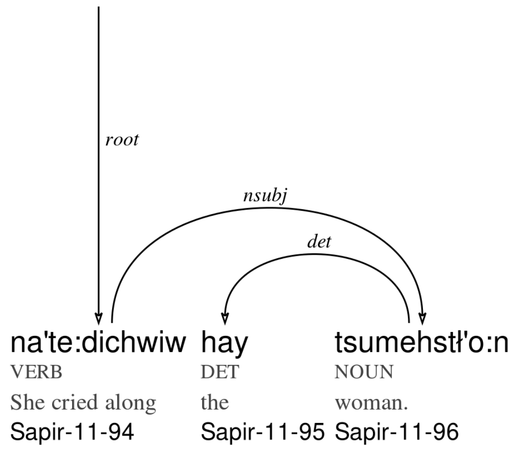
The specific annotation framework we have adopted is known as Universal Dependencies (UD henceforth) (Nivre et al. 2016), which provides a constrained set of core dependency relations that are intended to be applicable for any human language (hence “universal”), but with some flexibility to allow for language-specific variability. UD is explicitly designed to be applied to a typologically diverse set of language: the latest version (v. 2.1, Nivre et al. 2017) provides multilingual corpora for 60 languages with 102 dependency treebanks in total. Starting with a pre-defined and constrained set of dependency relations was considered advantageous insofar as one of the potential dangers in designing an annotation scheme from scratch for a language whose syntax is not well understood is to posit such a broad range of annotation options that generalizations are obscured and the scheme becomes difficult to apply consistently. UD’s cross-linguistic focus also offers the possibility of including the Hupa treebank in broader comparative and typological studies. Another advantage is that UD is strongly lexicalist in orientation, so the corpus does not need to be exhaustively parsed and analyzed morphologically - a daunting challenge in any Dene language - before syntactic annotation can proceed. Finally, there are a number of tools to assist with annotation and analysis for corpora constructed following UD conventions.

In a dependency grammar framework, the syntactic structure of a sentence relies on the individual words that comprise it. The main verb is treated as the head of the full sentence and assigned the dependency label [root]. Each word is in a head-dependent pair, and within the pair, the head and the dependent are associated directly with a binary dependency relation. Each dependent can have only one head, whereas the same head can have more than one dependent. To illustrate this, consider the Hupa sentence given in (1a) with associated dependency graph in (1b).[4] In this example, the verb *na’te:dichwiw* ‘she cries along’ is labeled as the root of the sentence. Each head-dependent pair is connected with a directed arc, which stems from the head and points to the dependent, with the dependency relation indicated on the arc. For instance, the two words *na’te:dichwiw* and *tsumehstł’o:n* ‘woman’ form a head-dependent pair, where *tsumehstł’o:n* is the dependent and *na’te:dichwiw* is the head. The dependency relation between these two words is labeled as [nsubj] on the dependency arc, indicating that *tsumehstł’o:n* is the subject of *na’te:dichwiw*. Similarly, the determiner element *hay* (often glossed with the English definite article ‘the’) is a dependent of *tsumehstł’o:n*, with the dependency relationship labeled [det].

(1a) na’te:dichwiw hay tsumehstł’o:n
She cried along the woman.
‘The woman cried as she went back along.’ (S&G 11.11)

[4] All examples are taken from the text collection published as Sapir and Golla (2001), referenced as “S&G” followed by the text number and line number separated by a period (so example (1) is taken from text 11, line 11). Like other texts available on the Hupa Online Dictionary and Texts website, Hupa words are rendered using the practical orthography of Golla (1996). Some interlinear glosses have been modified slightly from the original source for ease of exposition. Dependency graphs were created with the online annotation tool Arborator (Gerdes 2013).

(1b)



3.2. Implementation and Workflow

To date, the project has completed a preliminary annotation of 23 of the 74 texts in the Sapir and Golla (2001), over 4,700 glossed units in 814 numbered lines - approximately 26% of the Sapir collection and 13% of the corpus overall. The starting point for annotation is the (mostly) TEI-conformant XML text database underlying the Hupa Online Dictionary and Texts website. Syntactic relations in each sentence ultimately are realized as word-level annotations in the XML hierarchy. Consider, for example, the short sentence in (1a) above. The subject of the sentence, *tsumehstł'o:n* 'woman', would have the following XML representation in the database:

(1c)
```
<w xml:id="Sapir-11-96">
  <ref type="dependentOf" target="#Sapir-11-94">nsubj</ref>
  <reg>
    <m>tsumehstł'o:n</m>
  </reg>
 </w>
```

The syntactic dependency between the verb and its subject is represented with the <ref> tag in the second line of this example, which includes a pointer to the verb's unique identifier in the database (#Sapir-11-94) as well as the specific syntactic relation obtaining between them (nsubj).

Inserting these dependency labels directly in the XML structure is extremely cumbersome and prone to error, however. Therefore, the project uses homegrown Perl scripts to convert the XML representation to the tabular CoNLL-U format used by UD databases. This allows the project to take advantage of existing tools such as Arborator, an online dependency grammar annotator (Gerdes 2013, https://arborator.ilpga.fr/). A modified tab-separated CoNLL-U representation of sentence (1a) above is given as follows:

(1d)
```
1   na'te:dichwiw   She cried along   VERB   682    _   0   root    _   Sapir-11-94
2   hay             the               DET    3097   _   3   det     _   Sapir-11-95
3   tsumehstł'o:n   woman.            NOUN   3457   _   1   nsubj   _   Sapir-11-96
```

The syntactic dependency annotations appear in the seventh and eighth columns, so “1 nsubj” in the third line indicates that *tsumehstł’o:n* is an [nsubj] dependent of the word in the first line. These columns are empty when the sentence is exported from the XML database and filled in using the graphical user interface provided by Arborator. Note we are using some of the CoNLL-U fields in non-standard ways. The third column, for example, is normally used for a representation of the lemma to which a word belongs. We are instead using this field for the English gloss associated with each word, which facilitates annotation since this information is displayed in the online Arborator tool (cf. the third line below the dependency graph in 1b above). We use the fifth column, ordinarily reserved for language-specific for the dictionary id numbers associated with lexical items, which helps annotators look up items in the Hupa Online Dictionary and Texts interface. The last CoNLL-U field is reserved for miscellaneous project-specific annotations, which we use to store the unique identifier of each word in the corpus. A Perl script uses this identifier to insert the UD annotations added to the CoNLL-U file back into the XML database according to the conventions outlined in (1c).

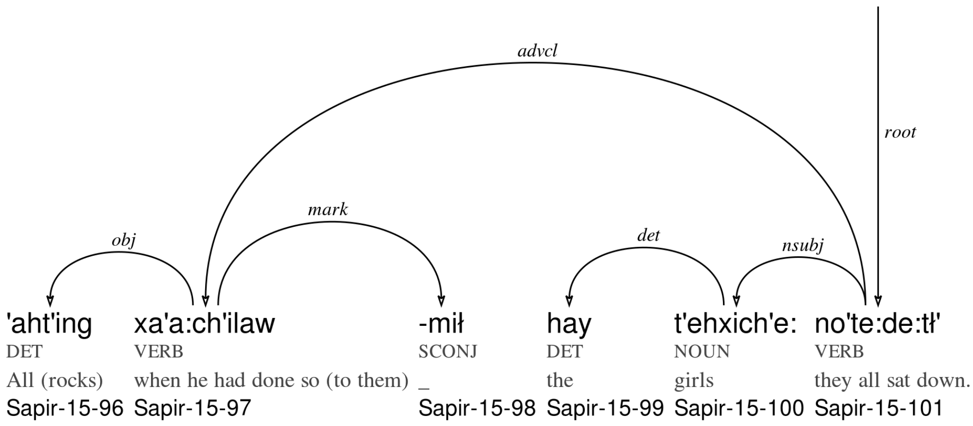
Our emphasis thus far has been on developing standard ways of annotating common constructions, with each text annotated by two members of the project team in order to ensure consistency. Once two annotators have completed a text, a third compares the results, reconciling discrepancies in the two annotation files by drawing on comparisons with similar previously annotated sentences, reference to the UD guidelines, and in some cases consulting specific UD corpora to determine how similar kinds of constructions are handled in the annotation of other languages. Decisions regarding conflicting annotations are documented for future reference and shared with the rest of the project team. Annotation decisions are guided by the extensive online documentation of UD principles and dependency relations (http://universaldependencies.org/), as well as ongoing project-specific documentation continually refined as new structures are encountered in the data. This is done iteratively, and texts that were annotated early on are re-checked and corrected as annotation conventions evolve in light of new data.

As mentioned above, the annotation method proceeds by first identifying the head of the sentence, typically the main verb, and then the dependency relations of each additional word in the sentence. This method works for relatively simple sentences like (1a) above, and is equally appropriate for more complex sentences, such as the one in example (2). Here, the verb of the main clause labeled [root] is *no’te:de:tł’* ‘they all sat down’, which has both a subject *t’ehxich’e:* ‘girls’ and an adverbial clause labeled [advcl]. The verbal head of the adverbial clause, *xa’a:ch’ilaw* ‘he had done so’, has both a subordinating enclitic *-mił* labeled with the UD [mark] dependency relation, and a direct object of its own, *’aht’ing* ‘all’.

(2a)	’aht’ing	xa’a:ch’ilaw-mił	hay	t’ehxich’e:	no’te:de:tł’
	All (rocks)	when he has done so (to them)	the	girls	they all sat down.

‘When he had done this to all the rocks, every one of the girls sat down.’ (S&G 15.11)

(2b)



Note that unlike syntactic theories where functional elements like determiners and subordinating conjunctions are the heads of their respective phrases, one of the overarching principles of the UD framework is the primacy of lexical categories, which are treated as heads with functional elements as dependents wherever possible. A second principle of UD annotation is promotion: when a content word is elided, an item that would ordinarily be analyzed as its dependent can assume the syntactic relation that the elided word would otherwise have had. In example (2), the object *tse:* ‘rocks’ in the adverbial clause is not expressed, and the quantifier *’aht’ing* ‘all’ is promoted to be the object of *xa’a:ch’ilaw*.

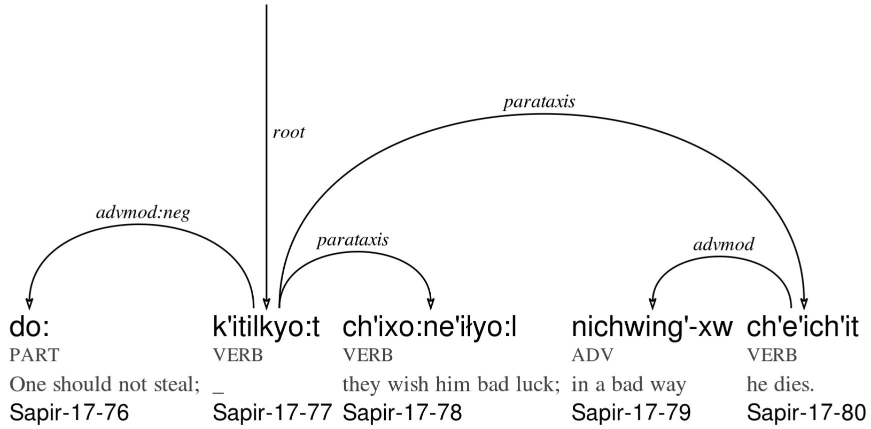
While identifying dependency relations in the aforementioned cases is relatively straightforward, there are many cases where it is much less so. This is due to one of the fundamental problems encountered in this kind of syntactic annotation for less-studied languages (Garrett et al. 2013): annotation is undertaken as a way to develop analyses of poorly-understood syntactic structures, but adding annotations to a corpus presupposes that the correct analysis is already known. The annotation procedure itself thus becomes a method for discovering rigorous syntactic analyses, with annotations refined iteratively as better understandings of recurring phenomena are developed. While the range of grammatical phenomena for which this iterative discovery procedure is needed in less-studied languages like Hupa may be greater than it is for languages with a longer history of formal linguistic analysis, it should be noted that this is fundamentally the same justification offered by Chomsky (1957) for developing computationally explicit models of grammar even for relatively well-studied languages like English: “By pushing a precise but inadequate formulation to an unacceptable conclusion, we can often expose the precise source of this inadequacy and, consequently, gain a better understanding of the linguistic data.”

Three examples of problematic areas of analysis and our treatment of them can be given by way of illustration. First, each text in the Sapir and Golla (2001) collection is divided into a sequence of numbered lines, each of which corresponds to a sentence of the text’s English free translation. Most numbered lines contain a main verb that can be assigned the label [root], but many include additional verbs that do not have any explicit marker of coordination or subordination, as in example (3) - note the use of semicolons in the English gloss line:

(3a)	do: k’itilkyo:t	ch’ixo:ne’iłyo:l	nichwing’-xw	ch’e’ich’it
	One should not steal;	they wish him bad luck;	in a bad way	he dies.

‘He doesn’t steal; people swear at him, and he dies in a bad way.’ (S&G 17.14)

(3b)



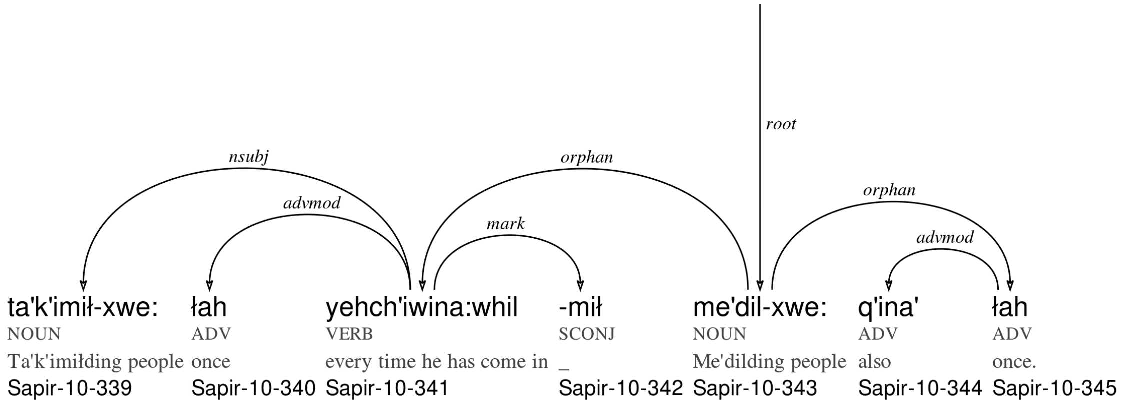
In such cases, the annotation team assigns the [root] label to the first verb of the line (here, *k’itilkyo:t* ‘someone steals’). Subsequent verbs are analyzed as dependents of the root with the label [parataxis] (http://universaldependencies.org/u/dep/parataxis.html). This analysis implies that the paratactic clauses are more closely connected to the root than they are to, say, verbs in preceding or subsequent lines, or to each other. However, since the line divisions themselves may be artifacts of the units that were deemed appropriate for rendering English free translations, it is not clear that this analysis is wholly appropriate: it might tend to reify aspects of the text that reflect post hoc considerations related to the process of translation. Nonetheless, by invoking the [parataxis] relation in such cases, we are able to annotate them consistently, making it easy to locate them and modify the analysis later if necessary. This highlights what has been one of our guiding annotation principles (following a suggestion by the Karuk treebank annotation team): to aim for consistency, even where correctness might be unobtainable for the time being.

A second example of analytic difficulty involves ellipsis. Example (4) is a case where there is no main verb to bear the [root] label for the sentence as a whole:

(4a)	ta’k’imił-xwe:	łah	yehch’iwina:whil-mił	me’dil-xwe:	q’ina’	łah
	Ta’k’imiłding people	once	every time he has come in	Me’dilding people	also	once.

‘Each time the *ta’k’imiłxwe:* go in to dance once, the *me’dilxwe:* also do so once.’ (S&G 10.38)

(4b)



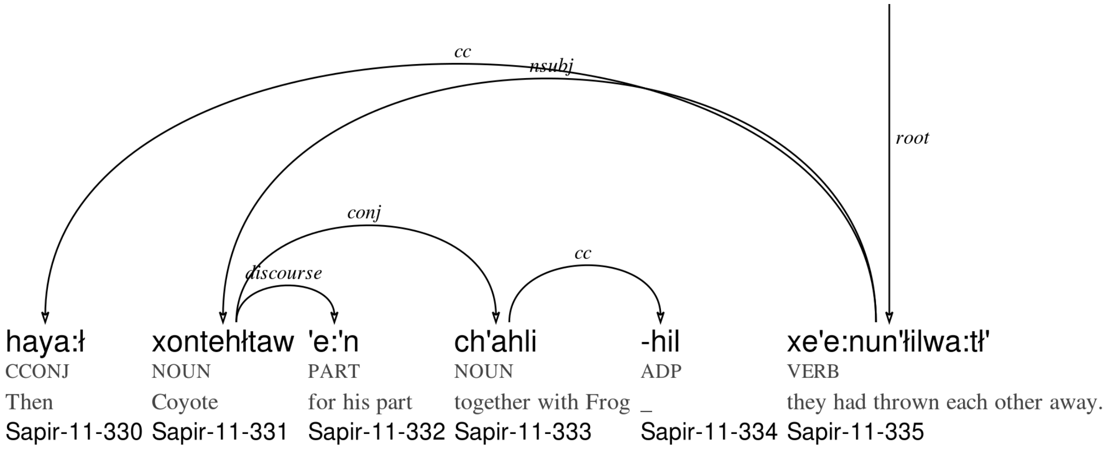
Here, the main verbal predicate is elided, and a remnant nominal subject *me'dil-xwe:* 'Me'dilding people' is assigned the [root] label according to the principle of promotion discussed above. Other clause-level elements in the sentence are assigned the label [orphan], which is used in cases where promotion under ellipsis "would result in unnatural and misleading dependency relation" between the promoted head and the remnants of ellipsis (http://universaldependencies.org/u/dep/orphan.html). In this example, the annotators gave preference to the nominal subject of the elided verb, perhaps on the principle that nouns are more canonically content words than adverbs. But other options for promotion to [root] might be preferable instead - perhaps the repeated adverbial modifier *łah* 'once' that, together with *q'ina'* 'also', seems to license the ellipsis, or perhaps the head of the subordinate clause *yehch'iwina:whil* (despite it being explicitly marked with the subordinating enclitic *-mił*). Although the correct resolution to this promotional ambiguity is not known at present, once again invoking the [orphan] relation makes it easy to locate this and similar examples later on.

Finally, example (5) illustrates the use of the non-core UD dependency relation [discourse]. The particle *'e:'n* 'for his part' is analyzed as dependent on the preceding nominal. The UD dependency [discourse] is intended for interjections and similar words "which are not clearly linked to the structure of the sentence, except in an expressive way" (http://universaldependencies.org/u/dep/discourse.html). While the Hupa particle *'e:'n* does typically express some degree of affective meaning (it is sometimes glossed with an English interjection such as 'Indeed!'), often it also performs a more well-defined discourse function, in example (5) probably indicating a shift in topic from the preceding line of the text. However, since the correct analysis of this particle is not known at present, treating it as a [discourse] dependent on the preceding noun is another example of an ad hoc annotation decision we've made in the interests of consistency. This will require revision in the future since it is unclear whether [discourse] dependents are more adequately treated as dependents of the clausal head, or whether some other dependency relation is more appropriate for this and similar elements.

(5a)	haya:ł	xontehłtaw	'e:'n	ch'ahli-hil	xe'e:nun'łilwa:tł'
	Then	Coyote	for his part	together with Frog	they had thrown each other away.

'Coyote and Frog had walked out on each other.' (S&G 11.44)

(5b)



4. Application to Word Order Phenomena

As noted above, our annotation of the Hupa text corpus is subject to revision as the annotation process itself leads to a better understanding of Hupa syntax. However, in this section we present data pertaining to surface level constituent order: while these results should still be considered preliminary, they are unlikely to be substantially affected by subsequent changes to our implementation of the UD annotation scheme. This is because identification of clause-level grammatical relations such as subject and object are generally more straightforward than some other dependency types. The main point, however, is not to assert the correctness of a particular analysis, but rather to illustrate how a well-annotated corpus can be brought to bear in addressing research questions that might be time-consuming or otherwise difficult to explore systematically with other methods.

As is the case for many Dene languages, the study of grammar above the level of the word in Hupa is very much in its infancy. One area that has received some attention is clause-level constituent order, specifically the ordering of noun phrases with respect to the verb. Dene languages are typically described as having SOV word order, sometimes quite rigidly so (e.g., Jung 2000 for Jicarilla and Lipan Apache). Golla (1970:295-296) observed that Hupa allows the placement of subject and object NPs in postverbal position such that (S)VO and (O)VS orders are not uncommon. This is illustrated in example (1), where the subject *tsumehstł'o:n* 'woman' occurs after the verb *na'te:dichwiw* 'she cried along' - cf. example (2), where the subject *t'ehxich'e:* 'girls' is in preverbal position. Newbold (2010) quantifies the relative distribution of preverbal vs. postverbal NPs encountered in a subset of the texts in Sapir and Golla (2001), showing that subject, direct object, and oblique NPs occur in postverbal position approximately 28% of the time.[5] Building on an observation of Conathan (2004:76-79), Newbold argues convincingly that the appearance of NPs in preverbal vs. postverbal position is sensitive to their

[5] Newbold found NPs to occur in postverbal position at a fairly consistent rate regardless of their type: subject (28.8%), object (28.2%), or locative/oblique (27.3%).

discourse status as new versus old information: NPs appearing in postverbal position tend to be discourse old, i.e., previously mentioned in the text.[6]

Analysis of the annotated Hupa text corpus largely confirms Newbold's findings with regard to the frequency of subject, direct object, and oblique NPs occurring in postverbal position. Results were obtained with a Perl script that takes the XML text database as input and searches UD syntactic dependency labels specified by the user, returning examples containing those labels and the labels of their dependents in tabular format. We were thus able to generate an exhaustive list of clauses with overt NPs by searching for clause-level labels (such as [root] and [parataxis]) with dependents [nsubj] (subject), [obj] (direct object), and [obl] (oblique). The distribution of main clause NPs with respect to the verbal head are shown in table 2:

	preverbal	postverbal
subject	174 (69.6%)	76 (30.4%)
object	196 (70.3%)	83 (29.7%)
oblique	431 (77.7%)	124 (22.3%)
total	801 (73.9%)	283 (26.7%)

Table 2: Distribution of preverbal and postverbal NPs in matrix clauses

These results are highly similar to Newbold's, an encouraging finding since it suggests that the results are not due to idiosyncrasies of the coding schemes used the two studies, or to the particular subset of the Sapir and Golla (2001) text collection that she analyzed.[7]

At this early stage of development, the corpus annotations discussed here are purely syntactic: they do not encode discourse and semantic properties such as definiteness and the old

[6] Newbold discusses some exceptions to this tendency, most of which fall into a few fairly well-defined classes, such as copular clauses and split NPs (where a portion of the NP, such as a quantifier, appears preverbally and the remainder postverbally: Spence 2008).

[7] For the most part, the texts analyzed by Newbold and the ones annotated in the present effort do not overlap: Newbold worked primarily with texts from the end of the Sapir and Golla (2001) collection, which features genres identified as "Myths and Tales" and "Legends and Traditional History," whereas we have annotated texts mainly from the beginning of the collection, which features descriptions of ceremonies and other aspects of traditional ways of living. A naïve chi-squared test using the prop.test() function in R (R Core Team 2013) suggests that the differences between Newbold's findings and ours are not statistically significant, both overall for all NP dependency types combined (p = .4343) or for the disaggregated dependencies individually, e.g., p = .3191 for obliques, the dependency type with the largest percentage difference between our findings (22.3% postverbal) vs. Newbold's (27.3% postverbal).

vs. new status of NPs.[8] We are therefore unable to directly evaluate the central explanatory claims of Newbold's study, that the status of NPs as new vs. old in the discourse is one of the main factors determining their relative order with respect to the verb. We can, however, offer indirect support for Newbold's analysis by extending the descriptive coverage to include adverbial clauses, as shown in Table 3:

	preverbal	postverbal
subject	45 (93.8%)	3 (6.3%)
object	29 (90.6%)	3 (9.4%)
oblique	63 (92.6%)	5 (7.4%)
total	137 (92.6%)	11 (7.4%)

Table 3: Distribution of preverbal and postverbal NPs in adverbial clauses

Unlike in main clauses, postverbal nominals are relatively rare in adverbial clauses: out of 148 examples, only 11 NPs (7.4%) in adverbial clauses occur postverbally. This difference between main vs. adverbial clauses has a natural interpretation that provides circumstantial support for the discourse-based account of Newbold (2010). According to the diachronic analysis of Conathan (2004:71-79), postverbal nominals are innovative in Hupa relative to the rest of the Dene language family and due to discourse-level functional convergence effects with neighboring languages such as Yurok and Karuk, which use similar word order alternations to encode information structure. Bybee (2001) notes that main clauses are cross-linguistically much more susceptible than subordinate clauses to diachronic changes involving word order. Bybee attributes this to the fact that main clauses are "pragmatically richer" than subordinate clauses - hence information foregrounding and backgrounding achieved with variable constituent orders occurs commonly in main clauses but not in subordinate clauses. The relative rarity of postverbal NPs in Hupa adverbial clauses, then, provides additional support for Newbold's discourse-based analysis of constituent order. More importantly for present purposes, the investment of effort into annotating the corpus made it possible to generate these results in a matter of minutes, and they can be updated as more of the corpus is annotated.

Another extension of Newbold's study involves the relative ordering of constituents in main clauses that have multiple NP dependents of the verb. Modifying the dataset used to generate Table 2 reveals another distributional asymmetry illustrated in Table 4, which shows the position

[8] This is in contrast to some other annotation projects for Dene languages, especially the one described by Nordhoff, Lovick, and Tuttle (2016), which does explicitly encode discourse-functional features such as newness and contrastiveness.

of NPs with respect to the verb for each attested pair of the subject, object, and oblique dependency types:[9]

	NP$_1$ NP$_2$ V	NP$_1$ V NP$_2$	V NP$_1$ NP$_2$
{subject, object}	8 (36.3%)	13 (59.1%)	1 (4.5%)
{subject, oblique}	29 (34.5%)	50 (59.5%)	5 (6.0%)
{object, oblique}	47 (47.5%)	50 (50.5%)	2 (2.0%)
{oblique, oblique}	40 (55.6%)	30 (41.7%)	2 (2.8%)
total	124 (44.8%)	143 (51.6%)	10 (3.6%)

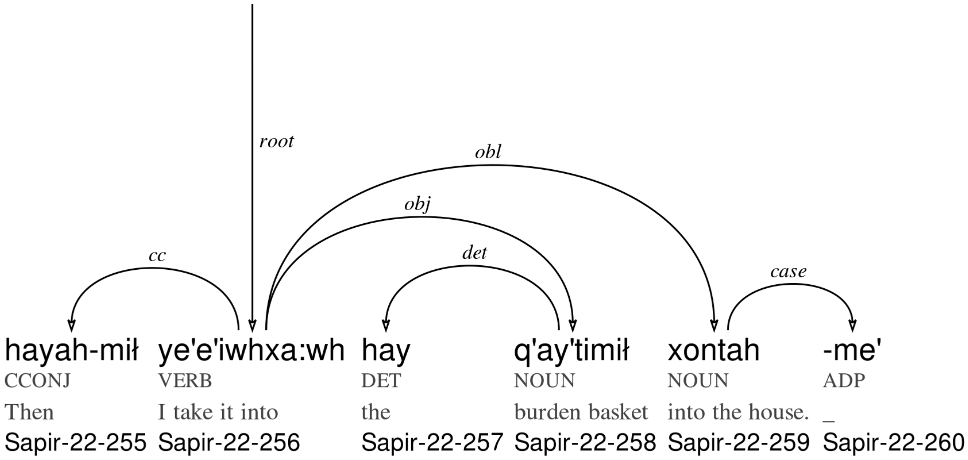
Table 4: Distribution of multiple NPs in main clauses

Table 4 shows that when there are exactly two NPs in a main clause, they are roughly equally likely to appear either both preceding the verb (NP$_1$ NP$_2$ V), or with one preceding the verb and the other following it (NP$_1$ V NP$_2$). Although not unattested, as illustrated in example (6), cases where both NPs follow the verb (V NP$_1$ NP$_2$) are exceedingly rare.

(6a)	hayah-mił	ye’e’iwhxa:wh	hay	q’ay’timił	xontah-me’
	Then	I take it into	the	burden basket	into the house

‘I take the pack basket into the house.’ (S&G 22.30)

(6b)



[9] In the first column of Table 4, curly braces are intended to be evocative of an unordered list - so “{subject, object}” includes cases where a subject NP precedes a direct object NP, and vice-versa.

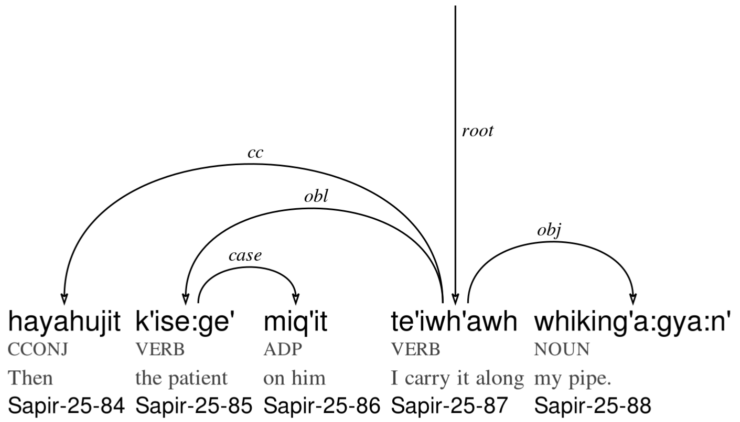
Opportunistic inspection of these data suggests that the cases where more than one NP occurs in postverbal position do tend to conform to the expectations of Newbold's analysis: in (6), for example, both *q'ay'timił* 'burden basket' and *xontah* 'house' are discourse old, mentioned in the two immediately preceding lines of the text.

However, under an analysis in which discourse-old NPs are expected to occur postverbally, it is somewhat surprising that cases such as (6) are not more common. In fact, it is not difficult to locate examples like (7), where the oblique *k'ise:ge'* 'patient' and the direct object *whiking'a:gya:n'* 'my pipe' are both discourse old, but only one occurs in postverbal position:

(7a)	hayahujit	k'ise:ge'	miq'it	te'iwh'awh	whiking'a:gya:n'
	Then	patient	on him	I carry it along	my pipe.

'Then I move my pipe all over the patient.' (S&G 25.14)

(7b)



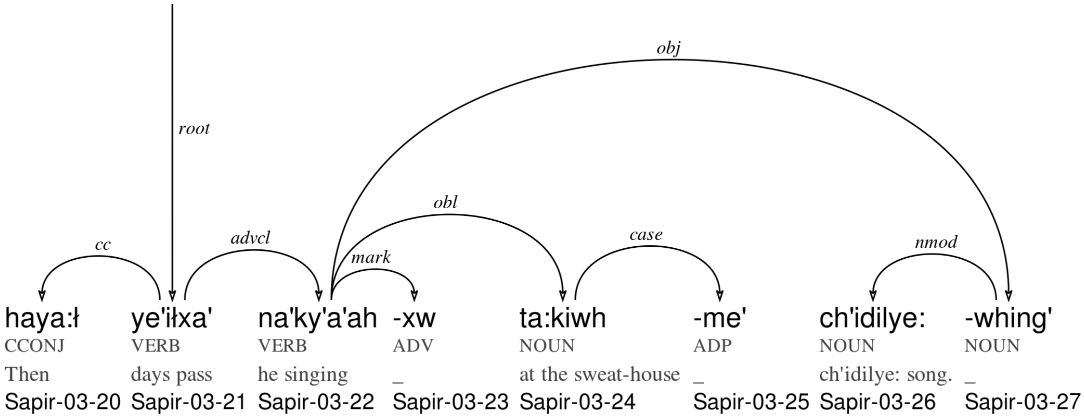
Although both *k'ise:ge'* and *whiking'a:gya:n'* are previously mentioned, perhaps relevant here is the fact that the former occurs only once, in the first line of the text, whereas the latter is mentioned explicitly or is an understood argument several times immediately prior example (7) - suggesting an analysis whereby a discourse-old NP that is more highly "activated" (cf. Lambrecht 1994:93-101) than another preferentially occupies the postverbal position. Resolving such issues definitively through more systematic consideration of the full range of examples must be deferred until the discourse-informational status of NPs is represented in the Hupa treebank (again pointing out the limitations of the strictly syntactic annotations that we have pursued thus far). Nonetheless, the key point for present purposes is that the syntactic annotations in the corpus, even at this relatively early stage of development, have heuristic value for data exploration, making it easy to identify the distributional asymmetry shown in Table 4, which in turn leads to an interesting hypothesis concerning constituent ordering in main clauses with multiple NP dependents of the root.

A final point worth noting is that the distributional data obtained from analysis of the annotated corpus reflect tendencies rather than absolute rules. There are discourse new nominals in postverbal position, discourse old nominals in preverbal position, and postverbal nominals in adverbial clauses. There is even the following example of an adverbial clause with both a direct object (*ch'idilye:-whing'* 'ch'idilye: song') and a locative oblique (*ta:kiwh-me'* 'at the sweathouse’) occurring after of the verb *na’ky’a’ah*, whose subordinate status is marked with the enclitic *-xw* and indicated with the label [advcl] in (8b):

(8a)	haya:ł	ye’iłxa’	na’ky’a’ah-xw	ta:kiwh-me’	ch’idilye:-whing’
	Then	days pass	he singing	at the sweat-house	ch’idilye: song.

‘He would sing World Renewal songs in the sweathouse for days on end.’ (S&G 3.4)

(8b)



This is a clear example of both dispreferred orders, with two postverbal nominals occurring in an adverbial clause. That such exceptions to otherwise robust tendencies can be identified when the full range of data is examined demonstrates how corpus methods are well-suited to discovering usage patterns involving phenomena where elicited grammaticality judgments might be unlikely to yield crisp judgments.

5. Conclusion

The aim of this paper has been to describe an ongoing project to enhance a text corpus of Hupa through syntactic annotation. The discussion above explicitly considers our choice of a dependency grammar framework and details of its implementation, which we hope can be a resource for similar projects in other Dene languages by highlighting the strengths and limitations of our approach. Although still in the early stages of development, already the annotated corpus is able to reveal usage-based patterns in the texts and contribute to our understanding of Hupa syntax. Moving forward, the project will continue expanding the set of texts in the corpus that have syntactic annotations and refining annotation procedures and analyses with the inclusion of more data. As the discussion in §4 suggests, an important future step will be to expand the strictly treebank-oriented syntactic annotations to include discourse-functional information along the lines described by Nordhoff, Tuttle, and Lovick (2016) for their corpus of Alaskan Dene languages, which will be relevant to developing more robust explanations for phenomena such as non-canonical constituent orders. Overall, while Hupa and other Dene languages may never have text corpora as large as ones that have been developed for English and other languages with international cachet, there is rich potential for applying corpus-based methods to develop a better understanding of their structure.

As discussed in §3, one of the reasons we chose to adopt the Universal Dependencies framework is its explicit orientation towards cross-linguistic comparison. Accordingly, we also hope that this report on the particulars of the Hupa corpus annotation project can contribute to a broader conversation among Dene language scholars regarding the feasibility of developing standards that will facilitate the sharing of corpus data for historical-comparative research. This has obvious appeal for traditional academic research questions, such as the possibility of a data-driven reconstruction of Proto-Athabaskan syntax. It is also potentially relevant for language revitalization, especially where sparsely documented languages must look to those with more documentation in order to fill in empirical gaps as revitalization efforts get underway: this has been the case, for example, for Wailaki, another Dene language of California, which has drawn on comparative data from Hupa in the development of a language revitalization program (Begay, Spence, and Tuttle to appear). Adopting common standards and data formats will facilitate the creation of software and methods that can be applied to corpora for different languages. Attending to these issues sooner rather than later, while many corpus projects are still in their early stages of development, will help avoid time-consuming problems related to resolving inconsistencies that will inevitably result from uncoordinated efforts. The time therefore seems ripe for engaging in discussions about the prospects for sharing corpus data among Dene language scholars, teachers, and language activists in order to facilitate collaborative work moving forward.

References

Begay, Kayla, Justin Spence, and Cheryl Tuttle. To appear. "Teaching Wailaki: Archives, Interpretation, and Collaboration." In *Translating Across Time and Space*, edited by Adrianna Link, Patrick Spero, and Abigail Shelton. Lincoln: University of Nebraska Press.

Berez, Andrea and Stefan Th. Gries. 2010. "Correlates to Middle Marking in Dena'ina Iterative Verbs." *IJAL* 76(1): 145-165.

Bybee, Joan. 2001. "Main Clauses are Innovative, Subordinate Clauses are Conservative." In *Complex Sentences in Grammar and Discourse: Essays in Honor of Sandra A. Thompson*, edited by J. Bybee and M. Noonan, 1-17. Amsterdam: Benjamins.

Bybee, Joan and Clay Beckner. 2010. "Usage-Based Linguistics." In *The Oxford Handbook of Linguistic Analysis*, edited by B. Heine and H. Narrog, 827-855. Oxford University Press.

Chomsky, Noam. 1957. *Syntactic Structures*. The Hague: Mouton.

Conathan, Lisa. 2004. *The Linguistic Ecology of Northwestern California: Contact, Functional Convergence, and Dialectology*. PhD dissertation, UC Berkeley.

Czaykowska-Higgins, Ewa, Martin D. Holmes, and Sarah M. Kell. 2014. "Using TEI for and Endangered Language Lexicon Resource: The Nxaʔamxcín Database-Dictionary Project." Language Documentation and Conservation 8: 1-37.

Escamilla, Ramón and Lindsey Newbold. 2012. "Notes on the Structure of the Hupa Personal Narrative." In *Working Papers in Athabaskan Languages 2011*, edited by James A. Crippen, Christopher Cox, Richard Dauenhauer, and Nora Marks Dauenhauer. Alaska Native Language Center Working Papers 10. Fairbanks: Alaska Native Language Center.

Garrett, Andrew. 2011. "An Online Dictionary with Texts and Pedagogical Tools: The Yurok Language Project at Berkeley." *International Journal of Lexicography* 24(4): 405-419.

Garrett, Andrew, Clare Sandy, Erik Maier, Line Mikkelsen, and Patrick Davidson. 2013. "Developing the Karuk Treebank." Presentation at the Fieldwork Forum, UC Berkeley, November 13, 2013.

Gerdes, Kim. 2013. "Collaborative Dependency Annotation." In *Proceedings of the Second Conference on Dependency Linguistics,* 88-97. Prague: Charles University, MatfyzPress.

Goddard, Pliny Earle. 1904. "Hupa Texts." *University of California Publications in American Archaeology and Ethnology* 1(2): 89-368.

Goddard, Pliny Earle. 1911. "Athapascan (Hupa)." In *Handbook of American Indian Languages, Part 1*, edited by Franz Boas, 85-159. Bureau of American Ethnography B-40(1). Washington, DC: Government Printing Office.

Goddard, Pliny Earle. 1914. "Chilula Texts." *University of California Publications in American Archaeology and Ethnology* 10(7): 289-379.

Golla, Victor. 1970. *Hupa Grammar*. PhD dissertation, UC Berkeley.

Golla, Victor. 1984. *Hupa Stories, Anecdotes, and Conversations.* Hoopa, CA: Hoopa Valley Tribe.

Golla, Victor, compiler. 1996. *Hupa Language Dictionary*, 2nd ed. Hoopa, CA: Hoopa Valley Tribal Council.

Holton, Gary and Olga Lovick. 2008. "Evidentiality in Dena'ina Athabaskan." *Anthropological Linguistics* 50(3/4): 292-323.

Jung, Dagmar. 2000. "Word Order in Apache Narratives." In *The Athabaskan Languages: Perspectives on a Native American Language Family*, edited by Theodore Fernald and Paul Platero, 92-100. New York: Oxford University Press.

Jurafsky, Dan, and James Martin. 2017. *Speech and Language Processing.* 3rd ed. draft. https://web.stanford.edu/~jurafsky/slp3/

Lambrecht, Knud. 1994. *Information Structure and Sentence Form: Topic, Focus, and the Mental Representations of Discourse Nominals*. New York: Cambridge University Press.

Lovick, Olga and Siri Tuttle. 2012. "The Prosody of Dena'ina Narrative Discourse." *IJAL* 78(3): 293-334.

Mikkelsen, Line. 2015. "Exploring Karuk Morphology in a Parsed Text Corpus." Paper presented at the 2015 SSILA Winter Meeting, Portland, OR.

Newbold, Lindsey. 2010. "Word Order in Hupa." In *Working Papers in Athabaskan Languages 2009*, edited by Siri Tuttle and Justin Spence, 89-106. Alaska Native Language Center Working Papers 8. Fairbanks: Alaska Native Language Center.

Nivre, Joakim, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, et al. 2016. "Universal Dependencies v1: A Multilingual Treebank Collection." In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, edited by Nicoletta Calzolari et al., 1659-1666. Paris: European Language Resources Association.

Nivre, Joakim, Željko Agić, Lars Ahrenberg, Lene Antonsen, Maria Jesus Aranzabe, Masayuki Ashahra, Luma Aseyah, et al., 2017. *Universal Dependencies 2.1.* LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, http://hdl.handle.net/11234/1-2515.

Nordhoff, Sebastian, Siri Tuttle, and Olga Lovick. 2016. "The Alaskan Athabascan Grammar Database." In *Proceedings of the Tenth International Conference on Language Resources* *and Evaluation*, edited by Nicoletta Calzolari et al., 3286-3290. Paris: European Language Resources Association.

O'Neill, Sean P. 2008. *Cultural Contact and Linguistic Relativity among the Indians of Northwestern California*. Norman: University of Oklahoma Press.

R Core Team. 2013. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.r-project.org

Sapir, Edward and Victor Golla. 2001. "Hupa Texts, with Notes and Lexicon." In *Northwest California Linguistics*, Vol. 14 of *The Collected Works of Edward Sapir*, edited by Victor Golla and Sean O'Neill, 19-1011. New York: Mouton de Gruyter.

Spence, Justin. 2008. "Some Discourse Effects on Split Nominals in Hupa." Presentation at the Syntax of the World's Languages III conference, Berlin, September 28, 2008.

Spence, Justin. 2013. *Language Change, Contact, and Koineization in Pacific Coast Athabaskan*. PhD dissertation, UC Berkeley.

Taff, Alice. 2013. *Woosh Een áyá Yoo X̲'atudli.átk: We're Talking Conversation.* 30 hours of Tlingit Conversation in Video with Bilingual Time-Aligned Text. University of Alaska Southeast. http://www.uas.alaska.edu/arts_sciences/humanities/alaska/languages/cuped/video-conv/

Tesnière, Lucien. 1959. *Éléments de Syntaxe Structurale*. Paris: Klincksieck.

Text Encoding Initiative Consortium. 2015. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. http://www.tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf.

Tuttle, Siri and Olga Lovick. 2014. "Alaskan Athabascan Commands: Grammatical Documentation from a Database Project." Paper presented at the 2014 SSILA Winter Meeting, Minneapolis, MN.