Word order in Cherokee: information structure, thematic structure, and variability

Brian Hsu (hsub@email.unc.edu), Benjamin Frey (benfrey@email.unc.edu)
University of North Carolina at Chapel Hill

Abstract: This paper examines the principles that determine clausal word order in Cherokee (Iroquoian; North Carolina and Oklahoma). Although the language is known to allow high flexibility in the ordering of major constituents in the clause, the principles that determine speakers' preferences among possible orders are not extensively described. We claim that a more comprehensive view of Cherokee grammar can be obtained by examining quantitative variation in word order in an annotated corpus of spoken narratives. Specifically, we investigate the properties that best predict the order of nominal and adverbial constituents relative to verbs, and the relative order among nominal and adverbial constituents. First, we confirm the effects of information-structural factors (referential accessibility and contrast) described in previous works, but find their influence to be probabilistic, rather than categorical. Second, we show that thematic roles (e.g. agents vs. themes) play an independent role in word order preferences. We use a logistic regression model to confirm that both types of factors interact cumulatively to influence word order probabilities. We further show that the principles we identify successfully explain the distribution of discontinuous nominal expressions in Cherokee. As broader theoretical contributions, our findings support claims that there is no strict division among languages based on whether thematic properties have any influence on word order, and that formal models should account for such probabilistic and cumulative patterns. Finally, our results suggest that annotated corpus methods can play a useful role in the documentation of typologically similar languages with variable word order in the clause.

Key words: Cherokee, syntax, word order, corpus linguistics, variability, cumulativity, information structure, thematic roles, discontinuous nominal expressions

1. Introduction

This paper is about the grammatical factors that influence word order preferences in languages with highly flexible word order, with a focus on Cherokee (Iroquoian, Oklahoma and North Carolina). As in many typologically similar languages, word order flexibility in Cherokee can be exemplified by the possible orders of transitive main verbs, agent nominal expressions, and theme nominal expressions. Following Dryer (1997), we illustrate this with the flexible orderings between transitive verbs and agent arguments, and between transitive verbs and theme arguments. Cherokee speakers accept each of these orders grammatical in an appropriate context, and each type of order can be found in spontaneous speech (Feeling and Pulte 1975; Scancarelli 1986; Beghelli 1996; Montgomery-Anderson 2008; 2015; Akkuş 2018), as shown in the examples below from recorded narratives. Like many languages with rich verbal agreement, Cherokee does not require all thematic arguments to be realized as nominal expressions. Examples (1a-b) lack an overt nominal theme, while examples (1c-d) lack an overt nominal agent.[1]

[1] We use the following glosses in the Cherokee examples (see Pulte & Feeling 1975; Montgomery-Anderson 2015 for additional discussion and definitions): 1 = first person, 3 = third person, AG = agentive, AMB = ambulative, AN = animate, CN = conjunction, CS = concessive, CIS = cislocative, DST = distributive, DT = delimiter, DUAL = dual number, EXCL = exclusive person, EXPP = experienced past, FUT = future, HAB = habitual aspect, INF = infinitive, IRR = irrealis,

(1) a. *Agent > verb*
ᎩᏟ ᎣᎩᏍᏓᏩᏛᏒ
[gitli] ogi-sdawadvs-v
dog 1.PL.EXCL-follow-EXPP
'The dog followed us.' (Feeling et al. 2017: 101)

b. *Verb > agent*
ᎠᎾᏓᏍᏕᎵᏍᎪ ᏴᏫ ᏧᎾᏍᏗ
a-n-adasdelis-g-o [yvwi j-u-n-asdi]
3-PL-help-PROG-HAB people DST-3-PL-little
'The little people help (others)' (Feeling et al. 2017: 43)

c. *Theme > verb*
Ꮓ ᎩᎳ ᎠᎹ ᎢᏙᏗᏍᎪ
No kil [am] ji-todis-g-o
Then until water 1-heat.water-PROG-HAB
'Then I heat some water.' (Feeling et al. 2017: 129)

d. *Verb > Theme*
ᎤᏍᏚᏅ ᎦᎶᎯᏍᏗᎢ
u-sdu-hnv [galohisdi?i]
3-close-EXPP door
'(he) closed the door.' (Feeling et al. 2017: 35)

The goal of this work is to understand the principles that determine Cherokee speakers' preferences among possible word orders, which have not been extensively investigated. It is relatively clear that Cherokee word order is strongly influenced by INFORMATION STRUCTURE factors such as the newness, salience, or relevance of individual constituents in the discourse context (Feeling and Pulte 1975; King 1975; Cook 1979; Scancarelli 1987; Montgomery-Anderson 2015). However, the extent to which ordering is determined by other properties is debated. For instance, Feeling and Pulte (1975) and King (1975) posit an independent tendency for noun placement to be determined by THEMATIC ROLE, such that agents preferentially precede themes. In contrast, Scancarelli (1986) posits that Cherokee follows the Newsworthiness Principle (Mithun 1992; 1995), which explicitly denies that grammatical properties like thematic role have any effect on word order.

This paper argues that this debate should be reframed, in light of growing evidence that word order in language can be determined by the interaction of multiple syntactic and discourse-related factors (whose preferences may conflict), and that the effect of individual factors can be probabilistic, rather than categorical (Bresnan et al. 2001; Manning 2003; Rosenbach 2005; Benor and Levy 2006; Szmrecsányi and Hinrichs 2008; Bresnan and Ford 2010; Schoenmakers et al. 2021; see Grafmiller et al. 2018 for a recent overview). Specifically, we propose that a more comprehensive understanding of Cherokee grammar can be attained by examining quantitative

---

ITR = iterative, LAT = lateral movement, LOC = locative, NEG = negation, NONF = nonfinite, PL = plural, OBJ = object, POS = possession, PRES = present tense, PROG = progressive aspect, REFL = reflexive, REL = relative, REPP = reported past, SG = singular, SPEC = specified action, SUBJ = subject, TR = translocative.

variation in word order in an annotated corpus of narratives. While the corpus methodology and statistical methods that we use are not novel in themselves, they have rarely been applied in the investigation of this type of pattern (the ordering of major constituents in a polysynthetic language), or in documentation work on similar languages (indigenous and/or highly endangered, and underresourced). Our study yields a number of descriptive, theoretical, and methodological contributions, which we preview below.

Our primary descriptive contribution is a newly comprehensive view of the factors that most strongly predict the order of major constituents in Cherokee clauses. First, we confirm the role of information structure as described in previous work, such as the preference for new information to precede old information, but find that the preference is PROBABILISTIC, rather than categorical. Second, we show that thematic information also plays a statistically significant role in word order preferences; thematic agents are likelier to occur early in the clause than theme arguments. In addition, we use a logistic regression model to show that information structure and thematic role are statistically significant, independent factors that interact CUMULATIVELY to influence word order probabilities. Finally, our proposal permits a new understanding of DISCONTINUOUS NOMINAL EXPRESSIONS in Cherokee, which have received little previous description aside from Williams (1996); we show that their distribution is best described using the information-structural and thematic word order preferences that we identify. While we cannot claim to present a fully exhaustive list of factors that predict word order in Cherokee, we are confident that our study has identified the most salient ones, given their high frequency of occurrence in naturalistic speech.

Although this research report does not aim to develop a formal account in this domain, our findings have several implications for theories of syntax. First, our findings offer new evidence that information-structural and thematic properties can interact in flexible word order patterns (Bader and Häussler 2010; Verhoeven 2014; Ellsiepen and Bader 2018). Furthermore, they suggest that there is not a strict division among languages based on whether thematic properties have any influence on word order (Payne 1987); Finally, the cumulative and probabilistic effects that we identify provide new support for formal models that can generate these types of patterns (Murphy 2017; Hsu 2021; Müller et al. 2022).

As a methodological contribution, our results suggest that annotated corpus methods can play a useful role in the documentation of typologically similar languages with variable word order, as it can uncover generalizations that are difficult to obtain from traditional elicitation methods (Tonhauser and Colijn 2010). This is particularly important, given that similar patterns of word order flexibility are common among indigenous languages of North America and Oceania (Dryer 2013). Furthermore, corpus analyses like the one that we present can facilitate the creation of pedagogically-oriented materials and grammars, one of many necessary tools needed to revitalize highly endangered languages like Cherokee. For example, it is easier to design instructional materials on flexible word order patterns if one can identify the most probable or frequent patterns, and the particular grammatical properties that condition them (Frey 2020).

The paper is organized as follows. Section 2 provides an overview of clausal word order in Cherokee, and prior analyses of its characteristics. Section 3 first describes the corpus and our annotation procedures for the tested predictors of word order. Section 4 presents quantitative evidence that information structure and thematic role independently contribute to word order preferences in Cherokee corpus. In Section 5, we discuss the effects of these factors on relative word order among nominal elements in longer sentences. Section 6 discusses the patterning of discontinuous nominal expressions, and how they can be understood in terms of the word order principles that we have identified. Section 7 concludes the paper.

2. Previous observations on clausal word order in Cherokee

Cherokee shows many morpho-syntactic properties that are common among polysynthetic languages (for a recent overview, see chapters in Fortescue et al. 2017). In addition to the aforementioned flexibility in word order and occurrence of discontinuous nominal expressions, all verbs contain a pronominal agreement prefix that inflects for properties of at least one thematic argument, including distinctions of person (first vs. second vs. third), number (singular vs. dual vs. plural), and clusivity (inclusive vs. exclusive); see Scancarelli (1987) and Montgomery-Anderson (2015) for comprehensive overviews of Cherokee morphology. It is relevant to note that contemporary Cherokee does not have productive noun incorporation (though it likely did at a historic stage; see Uchihara 2014); sentences in Cherokee likely contain more non-pronominal nominal expressions overall than equivalent sentences in languages with productive noun incorporation, including its Northern Iroquoian relatives.

All previous descriptions of the language concur that Cherokee word order is conditioned to a large extent by information-structural factors, which we briefly describe. First, there are effects of what we will call REFERENTIAL ACCESSIBILITY (Prince 1981). For example, nominal expressions that denote entities that are brand new to the discourse at hand tend to occur at the beginnings of clauses, while items that denote previously evoked entities tend to occur later. This is shown in a representative sequence of sentences in (2): *gitli* 'dogs' precedes the verb in the first sentence in the narrative in which they are mentioned; in a later sentence the now discourse-given *gitli* 'dogs' occurs after the verb.

(2) a.

ᎩᏟᏅ	ᏂᎪᎸ	ᏧᎵᏏᏅᏛ	ᎤᎾᏕᏯᏙᏗᎮ	ᎫᏍᏗ
gitli=hnv	nigolv	julsihnvd	u-n-adeytohdih-e	gusd.
dog=CN	always	nightly	3-PL-bother-REPP	something

'Every night something bothered their dogs.'

b.

ᏞᎦᏅ	ᎠᏎ	ᎠᎾᏂᏢᏍᎨ	ᎩᏟ
Hleg=hnv	ase	a-n-anhdlvs-g-e	**gitli**.
While=CN	maybe	3-PL-lie.down-PROG-REPP	dog

'The dogs would lie down for a while.' (Feeling et al. 2017: 81)

A second type of item with a tendency for early placement consists of phrases that bear CONTRAST. We define contrasted constituents as those whose referent(s) belong to a contextually relevant set of entities, evoked to the exclusion of those alternatives (Vallduví and Vilkuna 1998; Neeleman et al. 2007; Molnár 2002; Aissen to appear).[2] This is illustrated in the excerpt in (3): the first sentence (a) establishes that there is a relevant set of two men; each of the subsequent sentences comments on one of those men, to the exclusion of the other. Each of the corresponding nominal expressions bears contrast, and they each occur in a clause-initial position in (3b-c).

[2] Following these works, we assume that contrast is an independent notion from focus or topichood; not all types of focus involve an explicit alternative set, and information-structure topics can also bear contrast.

(3) a. **ᎾᏕTAi** **DhWᏢ** **DhᏫᏕᏫ**
wi-g-ajigo?-v a-ni-ta?li w-a-n-a?isv?-i
TR-1-see-EXPP 3-PL-two TR-3-PL-walk-AG
'I saw two men walking.'

b. **ᏌᏫZ** **ᎽWᏫ** **TBᏓ** **ᏂᏂᏢᏣ**
sagwu=hno kila=gwu iyvda jiy-olij-v
one=CN immediately=DT time 1.SG.SUBJ/3.AN.OBJ-recognize-EXPP
'One of them, I recognized immediately …'

c. **ᎤᎾᏢARTᏫᏍᎽh** **Ꮭ** **ᏍᏂᏂᏢᏙT**
u-n-aligos-v?i=sgini hla yi-jiy-olije?i
3-PL-be.partner-EXPP=CS NEG NONF-1.SG.SUBJ/3.AN.OBJ-recognize
'But his partner, I did not recognize.' (Feeling et al. 2017: 35)

Scancarelli (1986), discussing similar examples, proposes that Cherokee follows the Newsworthiness Principle (Mithun 1992; 1995), quoted below in (4), with our added roman numerals. We assume that "pragmatic function" refers to the information-structural factors that we have discussed, and that "syntactic function" includes thematic information. As such, the hypothesis asserts that thematic properties have no effects on word order in languages that follow this principle.[3] The elements defined in (i) correspond to discourse-new items, while elements in (ii) and (iii) correspond to items that bear contrast. In contemporary terms, the items in (ii) can be understood as contrastive topics or aboutness-shift topics and (iii) to contrastive foci (Vallduví and Vilkuna 1998; Neeleman et al. 2007).

(4) In a number of languages, the order of constituents does not reflect their syntactic functions at all, but rather their pragmatic functions: their relative newsworthiness within the discourse at hand. Constituents may be newsworthy because:
(i) they introduce pertinent, new information,
(ii) present new topics,
(iii) or indicate a contrast. (Mithun 1992: 58)

On the one hand, the Newsworthiness Principle correctly identifies phrase types that have a high propensity for early placement in the clause. On the other hand, it falls short of a complete descriptive or theoretical explanation of word order in Cherokee, for reasons that we now discuss.

First, there are aspects of the hypothesis that make it difficult to evaluate. The first arises from the claim that the word order of a sentence is determined by comparing the relative newsworthiness of each major constituent. While such comparisons may be possible among nominal expressions, there is no clear way to apply this to the ordering of nouns and verbs. Because nominal expressions

[3] Mithun claims that this principle holds in "pragmatically based languages." While there is no objective measure for how to identify such a language, Mithun (1995; 2017) suggests that all Iroquoian languages fall into this class, and suggests that this property develops as a consequence of having a rich set of agreement prefixes on verbs (which distinguish most combinations of person, number, and clusivity of subject and object arguments), and productive noun incorporation. While modern Cherokee differs from its Northern Iroquoian relatives in having lost productive noun incorporation, its ordering principles seem to strongly resemble those that Mithun describes for Cayuga (1992) and Tuscarora (1995), suggesting that Iroquoian languages are typologically highly similar in their word order principles.

typically refer to entities while verbs refer to predicates, verbs cannot be clearly characterized in terms of referential accessibility or contrast. Furthermore, because many verbal predicates must co-occur with nominal arguments, it is difficult to independently evaluate their pertinence in a sentence.

Second, the Newsworthiness Principle makes indeterminate predictions about the relative order of nominal constituents when they equally share or lack properties related to newness and contrast. For instance, sentence (5) occurs in a context in which both nominal expressions *na analsdelisgi* 'the helpers' and *gitli* 'dogs' have been the main participants of several preceding sentences; they appear to be equally discourse-given. Similarly, sentence (6) occurs at the very beginning of a narrative, suggesting that *daks* 'turtle' and *jiisd* 'rabbit' are equally discourse-new. Sentence (7) occurs in a context where two nominal expressions *sagwu=no* 'one (of the hunters)' and *junatana ahwi* 'big deer' are both contrasted with alternatives (there is a second hunter who kills small deer). Although the nominal expressions in these sentences are not clearly distinct in information-structure properties, they carry different thematic roles. In each of these examples, the agent argument precedes the theme, an ordering tendency separately posited by Pulte and Feeling (1975) and (King 1975).[4]

(5)	Ꮎ	ᎠᎾᎵᏍᏕᎵᏍᎩ	ᏫᏚᏂᏍᎦᏧ	ᏃᏊ	ᎩᏟ
	na	**a-n-alsdelis-g-i**	wi-d-u-ni-sgaj-e	nogwu	**gitli**.
	the	3-PL-help-PROG-AG	TR-DST-3-PL-call.off-REPP	now	dog

"The helpers called off the dogs/called the dogs back." (Feeling et al. 2017: 83)

(6)	ᏄᎵᏍᏔᏁᏥᏙᎸ	ᏓᎦᏏ	ᏚᎩᏴ	ᏥᏍᏚ
	N-uu-lstan-iidool-v	**daks**	d-uu-kiiy-v	**jiisd**
	NI-3B-happen-AMB-EXPP	turtle	DST-3B-beat.in.race-EXPP	rabbit

'How it happened that the/a turtle beat the/a rabbit.'

(7)	ᏌᏊᏃ	ᏧᎾᏔᎾ	ᎠᏫ	ᏓᎯᎮ
	[**Sagwu=no**]	[**j-u-n-atana**	**ahwi**]	d-a-hih-e
	one=CN	DST-3-PL-big	deer	DST-3-kill-REPP

'One (of the hunters) killed big deer' (Feeling et al. 2017: 53)

At this point, it is important to note that it may not be possible to get a clear understanding of the language's word order principles in a traditional approach that relies on the qualitative examination of individual sentences, even when their context within a conversation is considered. First, because every non-verbal major constituent in a clause is specified for properties related to both thematic role and information structure, it is difficult to isolate which property among multiple specifications is responsible for a particular observed word order. Second, word order principles in a language can be probabilistic; language grammars often follow formal principles that nonetheless tolerate occasional, but genuine exceptions (Bresnan et al. 2001; Manning 2003; Rosenbach 2005; Benor and Levy 2006; Szmrecsányi and Hinrichs 2008; Bresnan and Ford 2010;

[4] Pulte & Feeling (1975) describe this in terms of the ordering of *subjects* and *objects*. However, we can understand "subject" in this context as referring to agents of transitive verbs, and "objects" to themes of transitive verbs. In their terms (p. 353): "In simple declarative sentences in Cherokee, the subject of the sentence ordinarily precedes the verb with its modifiers and objects. In addition, objects of verbs ordinarily precede the verb, resulting in subject-object-verb word order."

Schoenmakers et al. 2021). We show this to be true of Cherokee; it is easy to find sentences produced by Cherokee speakers that are inconsistent with each of the tendencies that we identify, even as the overall trends follow clear grammatical principles.

In the remainder of the paper, we propose a new framework for investigating and describing clausal word order in Cherokee. First, we focus on describing the information-structural and thematic properties of nominal expressions, and the extent to which they predict (i) whether the nominal expression precedes or follows the verb of its clause, and (ii) the relative order of nominal expressions. We do not attempt to ascribe information-structural properties to verbs themselves. Second, we propose that we can achieve a more comprehensive understanding of Cherokee grammar by analyzing quantitative variation in a corpus of narratives (see Tonhauser and Colijn 2010 for a similar approach to Guaraní). For all sentences in the corpus, we annotated all major constituents for a range of properties related to referential accessibility, thematic role, contrast, animacy, and phonological length. We then use a mixed-effect logistic regression analysis to quantify the propensities of individual properties to condition word order while controlling for the effects of other predictors, yielding a more comprehensive model of how these factors interact.

Our results should not be taken as an exhaustive description of factors that condition word order preferences in Cherokee. Given the currently available resources on the language, we are not able to investigate a comprehensive inventory of information-structural properties beyond referential accessibility and contrast, such as factors related to prosodic structure and/or intonation (Shih et al. 2015), segmental phonological restrictions (Szmrecsányi and Hinrichs 2008; Shih and Zuraw 2017), or dialectal and other sociolinguistic factors (Bresnan and Ford 2010; Szmrecsanyi et al. 2017). Nonetheless, we believe that our corpus analysis identifies the most salient word order predictors in Cherokee, which occur most frequently in speech.

3. The Cherokee corpus and its annotated features

3.1 *Overview of the corpus*

Our annotated corpus consists of existing recorded narratives in Cherokee that have been published in Montgomery-Anderson (2008; 2015) and Feeling et al. (2017), the most extensive collections of glossed spoken Cherokee. All of these narratives are supplied by the original authors with a transcription (in both Cherokee syllabary and romanization), an interlinear morphemic gloss, and an English translation. This greatly facilitates the task of annotating the texts, given a general knowledge of Cherokee grammar and training in the classification of thematic and information-structure properties.

The corpus includes twelve narratives produced by nine speakers in total. Eight of these are speakers of Oklahoma Cherokee, with one narrative told by an Eastern Cherokee speaker.[5] The corpus consists primarily of personal narratives about events experienced by the narrator, family narratives that recount experiences of family members, and folk tales, with one procedural narrative 'How to make chestnut bread' and a historical narrative 'The search party.'[6] The list of narratives is given below in Table 1. Given the small current inventory of transcribed and

[5] We are not aware of dialectal differences in word order principles between Oklahoma Cherokee and Eastern Cherokee, but leave the question open for future study.

[6] To better control for genre and modality, we did not include narratives in Feeling et al. (2017) that are older, written texts ('The Good Samaritan,' 'Diary,' 'Legal Document'), back-and-forth conversations ('Hunting Dialogue,' 'Interview with Wilbur Sequoia'), or heavily code-switched ('Reminiscence').

morphologically segmented texts in the language, we acknowledge that we cannot rule out possible effects of genre or formality on our results, and that the effects of some word order factors may be better detectable in multi-participant dialogues. Nonetheless, we proceed with the assumption that this sample contains a reasonable representation of the core clausal word order principles of spoken Cherokee.

Text	**Description**	**Source**
'Ball of fire'	Personal narrative	Feeling et al. (2017)
'Cat meowing'	Personal narrative	Feeling et al. (2017)
'The invisible companion fox'	Personal narrative	Feeling et al. (2017)
'Little people'	Folk tale and family narrative	Feeling et al. (2017)
'Origin of evil magic'	Folk tale	Feeling et al. (2017)
'Spearfinger'	Folk tale	Feeling et al. (2017)
'Transformation'	Family narrative	Feeling et al. (2017)
'Two dogs in one'	Personal narrative	Feeling et al. (2017)
'Water beast'	Folk tale	Feeling et al. (2017)
'How to make chestnut bread'	Procedural narrative	Feeling et al. (2017)
'Rabbit and buzzard'	Folk tale	Feeling et al. (2017)
'Throw it home'	Personal narrative	Feeling et al. (2017)
'Wolf and crawdad'	Folk tale	Montgomery-Anderson (2008)
'The search party'	Historical narrative	Montgomery-Anderson (2008)
'The turtle and the rabbit'	Folk tale	Montgomery-Anderson (2008)

TABLE 1: Texts in the annotated Cherokee corpus

Within the corpus, we tag all *nominal expressions*, items which can in principle refer to an identifiable entity in the world, and all *thematic elements*, items in a thematic relation with a verbal predicate (these items need not be referential). Each item is tagged for all identifiable values for a range of grammatical properties, which are used as the independent variables of our statistical analyses in Section 4. Our corpus excludes filler particles like *nogwu* 'now,' non-thematic adverbs like *ase* 'maybe' or *do* 'really,' and predicate adjective phrases, as they are not easily classfiable in terms of information structure or thematic properties.[7]

The corpus contains 580 total sentences. As shown in Table 2, a large majority of them have only one tagged major constituent other than the verb. There are sentences in the source texts with only a verb, but they are not included in the corpus since we are mostly interested in relative order. Our corpus includes both main and embedded clauses; while clausal embedding has clear word order effects in some languages, we have not noticed any clear effects in Cherokee, but leave it as an open question to be confirmed in later work.[8]

[7] Some instances of *nogwu* 'now' remain in the corpus if they are contentful time expressions, as determined by the sentence context or the English translation provided in the cited works.

[8] There are also methodological difficulties in annotating a distinction between main and embedded clauses in the corpus, due to the absence in Cherokee of complementizers or verbal inflection that signal clausal subordination. While there are intonational cues to clausal subordination in the language (Uchihara 2013), they are not available from the transcribed texts.

One major constituent	410
Two major constituents	140
Three major constituents	23
Four major constituents	7
Total number of sentences	580
Total number of major constituents	787

TABLE 2: Number of non-verbal major constituents per clause in the corpus

In addition, each major constituent is tagged for word order values PREVERBAL, POSTVERBAL, and DISCONTINUOUS. We use PREVERBAL and POSTVERBAL as the main dependent variable in the analysis, and discuss the relatively uncommon discontinuous NP examples in Section 6.

In the next subsections, we introduce the potential word order factors that we examine, and their annotation procedures. Factors that were not found to have significant or reliable word order effects in the corpus (length and animacy), and potential factors that were not are discussed in Section 3.5. Section 4 presents the results of the mixed-effects logistic regression analysis, and further evidence for the independence of referential accessibility and thematic role as conditioning factors in Cherokee word order.

3.2 *Referential accessibility*

While there are are many proposed classifications of referential accessibility and related notions, we largely adopt the annotation scheme in Dipper, Götze, and Skopeteas (2007). This tagset is based on the classifications of *assumed familiarity* proposed by Prince (1981; 1992), defined by the listener's ability to identify the referent of an expression in the context of the discourse, as most likely assumed by the speaker. The benefit of this is that the primary tags can be indentified with relative confidence from observable properties of the narrative: what has been evoked in preceding text. This has been shown to facilitate agreement across annotators (Nissim et al. 2004). We use four tags: NEW, GIVEN, ACCESSIBLE, and NONREFERENTIAL.

NEW items are nominal expressions that are being introduced to the discourse for the first time, whose referents are not likely to be inferrable from general knowledge, or entities in the existing discourse. These are illustrated with the first few sentences of the narrative "Spearfinger." Each sentence introduces the main characters of the narrative for the first time.

(8) a.

ᏴᎩ	**ᎤᏩᏍᎩ**	**ᏚᏙᎡ**	**Ꭿ**	**ᎠᎨ**
Yvgi	u-wasgi	d-u-do?-e	h-e	[**age**].
Spear	3-finger	DST-3-named-REPP	live-REPP	woman

'There was a woman named Spearfinger.'

b.

ᏅᎩ	**ᎢᏯᏂ**	**ᎠᏂᏃᎭᎵᏙ**	**ᎤᎾᏂᏍᎨ**
[**Nvgi**	**iyani**	**a-ni-nohalido**]	u-n-anigis-e.
Four	number	3-PL-hunt	3-PL-leave-REPP

'Four hunters left (went hunting).'

c. ᏧᎾᏓᎵ ᏚᎾᏘᏅᏎ
[**j-u-n-adali**] d-u-n-atinvs-e.
DST-3-PL-spouse DST-3-PL-take.along-REPP
'They took along their spouses.' (Feeling et al. 2017: 62)

GIVEN items refer to entities that have been explicitly mentioned in the preceding discourse. We illustrate this with an example from "Spearfinger" that occurs after the sentences shown previously in (8). *Sgina yvgi uwasgih* 'that Spearfinger' is tagged as given since its referent (the woman named Spearfinger) has been previously referred to explicitly.

(9) ᎨᏴ ᏗᎦᏙᎨ ᏍᎩᎾ ᏴᎩ ᎤᏩᏍᎩ
geyv di-g-ado-g-e [**sgina** **yvgi** **u-wasgih**].
over.there DST-3-stand-PROG-REPP that spear 3-finger
'Spearfinger stood over in the distance.' (Feeling et al. 2017: 63)

The following example from "Water Beast" contains two discourse-given expressions. At this point in the narrative two men have already seen a bull-like animal in a river and are watching its movement. The "hole in the water" is a location where the animal previously surfaced. As seen in this example and in (9), It is common, though not obligatory, for given expressions to occur with demonstratives like *nasgina* 'that' and *nahna* 'there.'

(10) ᎾᏍᎩᎾ ᏩᎦ ᏧᎧᏅᏍᏕᏅ ᎾᏏ ᏩᏔᎴᏒ ᎠᎹᏱ ᎢᎤᏕᏘᏁ
[**nasgina** **wahg** **jukanvsden**] [**nahna** **watalesv** **ama-y**] i?-u-detin-e.
that cow bull there hole water-in again-3-dive-REPP
'The bull dived back in the water.' (Feeling et al. 2017: 110)

Our annotation scheme further distinguishes given items based on how recently they were last mentioned. Entities that are explicitly referred to or are implied thematic participants of the preceding clause are tagged as GIVEN-ACTIVE. Given entities that are last mentioned prior to the previous sentence are tagged as GIVEN-INACTIVE. The latter tag is a proxy measure in some ways for shifted topics, which show in some languages a greater tendency to occur early in the clause (Bader 2020).[9]

ACCESSIBLE entities have not been explicitly mentioned in the discourse, but the identity of their referents can be inferred by the hearer from either a relationship with a discourse-given entity, or from general knowledge about the world.[10] Intuitively, they have an intermediate status between new and given entities. This includes expressions whose referents are inferrable from a part-whole relation, subset relation, or superset relation with a given entity. Example (11) in "The invisible companion black fox" occurs in the narrative after the narrator has previously described stopping his car to pick someone up. The door of the car has not been previously mentioned, but its referent

[9] Because the narratives in Feeling et al. (2017) and Montgomery-Anderson (2015) are not transcribed with punctuation, and the placement of clause boundaries is in some cases ambiguous, we rely on the authors' English translations of the text to determine sentence boundaries, in order to distinguish between active versus inactive items.

[10] Our original corpus tags distinguish between three subtypes of accessible items, using the ACC-AGGREGATE, ACC-INFERRABLE, ACC-GENERAL tags of Dipper et al. (2007). However, the relatively rarity of ACC-AGGREGATE and ACC-GENERAL constituents in the corpus prevents us from identifying whether these subclasses pattern distinctly. It should also be noted that annotators are less likely to agree on these subclasses (Nissim et al. 2004).

is accessible from its part-whole relation with the discourse-given car. In example (12) from "Throw it home" the narrator has described a baseball game taking place. The existence of the pitcher is inferrable as being a necessary participant of the game.

(11) ᎤᎾᏦᏗ ᎤᎬᏩᏟ ᏓᏳᏍᏚᎢᏒᏃ ᎦᎶᎯᏍᏗ
u-na-jo-di ugvwahli d-a-yusdu?is-v=hno [**galohisdi**]
PL-3-open-INF purpose DST-3-open-EXP=CN door
'He opened the door to get in.' (Feeling et al. 2017: 35)

(12) ᏄᎵᏍᏘᏙᏏᏰᎥ ᎤᏕᎩ
n-u-lsgwidosiy-e?-v [**u-de-g**]
LAT-3-contort-EXPP 3-pitch-PROG
'The pitcher contorted.' (Feeling et al. 2017: 217)

Other items are accessible if their referents are known to the hearer as part of general, shared knowledge about the world. This includes expressions like *svnoyi ehi nvda* 'moon' in (13).

(13) ᏒᏃᏱ ᎡᎯ ᏅᏓ ᎥᏍᏊ ᎢᎦᎯ ᎤᏘᏍᏛᎢ
[**Svnoyi** **eh-i** **nvda**] vsgwu igahi u-tisd-v?i
Evening be-AG sun also brightly 3-shine-EXP
'The moon was shining brightly.' (Feeling et al. 2017: 14)

Finally, we use the NONREFERENTIAL tag for expressions of several types that do not refer to an identifiable entity in the discourse. This includes thematic arguments that do not clearly designate an entity. In example (14) from "Cat Meowing," the demonstrative pronoun and relative clause are both thematically related to the main verb *to happen*, but these expressions refer to events, rather than identifiable entities.

(14) ᎥᏍᎩᏃ ᏄᎵᏍᏔᏂᏙᎸ ᎯᎠ ᏥᏥᏃᎮᎭ
[**Vsgi=no**] n-u-lstani-dol-v [**hi?a** **ji-ji-noheh-a**]
This=CN SPEC-3-happen-around-EXP this REL-1-live-PRES.
'This is what happened in this story.' (Feeling et al. 2017; 23)

A second class of nonreferential expressions denote property or kind readings, rather than a set of individuals or objects in the world. In example (15) from "Spearfinger," *ajilvye* 'fire' and *jigoya* 'bug' refer to generic entities, rather than specific instances of them. In example (16) from "Rabbit and Buzzard," *uhnvwisgi* 'doctor (lit. one who treats)' refers not to a specific being, but a generic referent, like *any doctor*.

(15) ᎠᏥᎸᏰᏃ ᏱᎪᏔᎾ ᏥᎪᏯ ᏣᏔᏍᎩᏍᎪ
[**Ajilvye=no**] yi-g-otan-a [**jigoya**] j-atasgis-g-o
Fire=CN IRR-3-build.fire-PRES bug REL-explode-PROG-HAB
'the way a bug explodes when it is thrown in a fire.' (Feeling et al. 2017: 67)

(16) ᏃᏭᏃ ᎤᏂᏯᎴ ᎤᏅᏫᏍᎩ
nowu=no u-ni-hyal-e [**u-hnvwis-g-i**]
Now=CN 3-PL-search-REPP 3-treat-PROG-AG
'So then they searched for a doctor (lit. 'one who treats').' (Feeling et al. 2017: 142)

Finally, we use the nonreferential tag for predicate nominals like (17), and expressions like (18) that denote names, but do not themselves refer to an individual. These types of items are also annotated with the PRED-OBJ thematic role, discussed in Section 2.2.

(17) ᏍᎩᏅ ᎠᏑᏂᏙ ᎨᎲ
sgi=hnv [**asuhnidoh**] ge-hv.
that=CN fisher be-EXPP
'He was a fisher.' (Feeling et al. 2017: 215)

(18) ᎠᏂ ᏍᏊ ᏚᏙᎥ ᎠᎩᏥ ᏥᎨᏒ
[**An**] sgwu d-u-do?-v agi-ji ji-ges-v.
Ann also DST-3-be.named-EXP 1.POS-mother REL-be-EXP
"My mother was also named Ann." (Feeling et al. 2017: 84)

Having now defined the key tags for referential accessibility in our corpus, we now turn to their quantitative effects on word order preferences in Cherokee. First, we observe that there is a clear overall trend for all expressions to occur preverbally, across all levels of referential accessibility: 404 (71%) of the 569 tagged constituents precede the verb, and preverbal order is most common within each specification. As expected from previous studies, new information shows a higher preverbal preference (81%) than both accessible (71%) and given information items (60%). We do not find a clear difference in patterning between active and inactive given items. While we did not have a clear expectation based on previous works about the patterning of nonreferential items, we find that they show a relatively high preference for preverbal placement (86%). We discuss statistical significance in greater detail in Section 4.1.

	Accessible	Given-active	Given-inactive	New	Nonreferential	Total
Postverbal	29	64	34	24	14	165
Preverbal	70	95	55	100	84	404

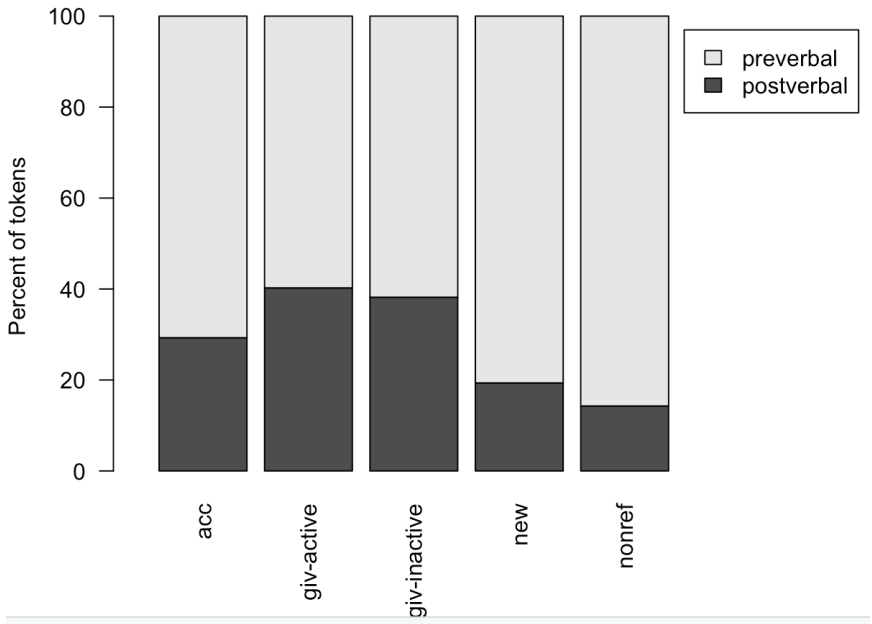

Figure 1: Referential accessibility and placement relative to verbs

3.3 *Contrast*

Broadly, we understand contrast to involve structures where an entity is evoked, to the exclusion of potential alternatives in a contextually relevant set (Vallduví and Vilkuna 1998; Neeleman et al. 2007; Molnár 2002; Aissen to appear). Following Repp (2016; 2010), one can also define different types and/or degrees of contrast. In order to maximize our confidence in correctly identifying contrasted items from written text alone, we employed the CONTRAST tag only for expressions that have an *explicit alternative set*, as defined by Repp. Specifically, we use the tag only for expressions that are part of a set explicitly mentioned in a preceding part of the narrative (ex. *four women were in the forest, one of them …*), or if an entity is explicitly compared with an alternative (ex. *one man ran, the other one stayed*). See (3) above for one example. All other entities were tagged with the value NO CONTRAST. We did not attempt to annotate instances of contrast that depend on implicit alternative sets that are not directly evoked in the narrative, as these are very difficult to objectively identify from textual properties alone (any entitity can plausibly be construed as being a member of some group, even for instance 'the set of people in a story').

Despite the relatively small number of entities that bear contrast with a contextually explicit alternative set, the property appears to be a fairly robust predictor of preverbal placement (22/23 = 96%). This distinction is also found to be statistically significant in the regression model presented in Section 4.

	Contrast	No contrast
Postverbal	1	132
Preverbal	22	268

TABLE 3: Contrast and placement relative to verbs

3.4 *Thematic role*

We examine the word order effects in our corpus of seven thematic roles, which we define below. This includes the argument thematic roles associated with verbs that describe events or states (AGENT, THEME), verbs related to psychological states (EXPERIENCER, STIMULUS) and roles involved in copular predicates (PREDICATE-SUBJECT, PREDICATE-OBJECT). In addition, we tagged the most common adjunct roles TIME and DYNAMIC LOCATION and STATIC LOCATION.[11] While the vast majority of major constituents with a thematic role tag are also tagged for information structure propereties, we do not use information structure tags for time and abstract location arguments, as they do not generally participate in the same system of reference as the argument thematic roles.

AGENTS are entities that initiate or cause an event denoted by a verbal predicate. As shown in the examples below, these can correspond to sole arguments of certain intransitive verbs, or subject arguments of some transitive predicates.[12]

(19)	ᎡᎵᏏᏃ	ᏃᏊ	ᎪᎯ	ᎢᏴ	ᎤᏂᏨ
	[**E-lisi=hno**]	nogwu	kohi	iyv	u-hnej-v
	1.POS-grandmother=CN	then	after	a.while	3-speak-EXP

‘After a while, my grandmother spoke.’ (Feeling et al. 2017: 26)

(20)	ᎥᏍᎩᏃ	ᎢᏹᎵᏍᏔᏂᏙᎳ	ᎠᎾᏓᏍᏕᎵᏍᎪ	ᏴᏫ	ᏧᎾᏍᏗ
	Vsgi=hno	yi-n-u-lstanidol-a,	a-n-adasdelis-g-o	[**yvwi**	**j-u-n-asdi**]
	That=CN	IRR-SPEC-3-happen-PRES	3-PL-help-PROG-HAB	people	DST-3-PL-little

‘If it happens, the little people help (you).’ (Feeling et al. 2017: 43)

[11] While languages make use of a larger inventory of thematic relations than the ones that we discuss in this paper (such as source and instrument adjuncts), these were not included in the quantitative analysis due to their very rare occurrence within the corpus.

[12] The quantiative analyses below exclude agents of verbs like *say* and *tell* when they have a clausal complement, as they do not appear to be representative of the patterning of agents more generally. Specifically, there is a unique tendency in these contexts for the agent argument to occur clause-finally, resembling quotative inversion structures in languages like English. An example of this is shown in sentence (i) from “Spearfinger.”

(i)	Ꮩ	ᏥᏈᏯ	ᏓᏛᏗ	ᎠᏂᎨ	ᎤᏛᏁ	ᏌᏊ	ᎠᏍᎦᏯ
	Do	jigwiya	d-a-tvdi	a-ni-ge	u-dvn-e	[**sagwu**	**asgaya**]
	Really	too.many	DST-3-do.away	3-PL-woman	3-say-REPP	one	man

‘‘She is just killing too many women,’ said one man.’ (Feeling et al. 2017: 66-67)

(21)	ᎾᏍᎩᏃ	ᎯᎠ	ᎠᏂᏔᎵ	ᎠᏂᏍᎦᏯ	ᎠᏂᎦᏪᎯᎮ
	Nasgi=hno	[**hi?a**	**a-ni-ta?li**	**a-ni-sgaya**]	a-ni-gawehih-e
	That=CN	this	3-PL-two	3-PL-man	3-PL-paddle-REPP
	‘Two men were paddling.’ (Feeling et al. 2017: 109)				

THEMES (a.k.a. patients) are entities that undergo a process or change of state. We use this tag for several types of items (see Sorace 2000 for additional discussion of subclasses of themes). In some examples, these are object arguments of transitive predicates, as shown in examples (22) and (23).

(22)	Ᏺ	ᎩᎳ	ᎠᎹ	ᎢᏙᏗᏍᎪ
	No	kil	[**am**]	ji-todis-g-o
	Then	until	water	1-heat.water-PROG-HAB
	‘Then I heat some water.’ (Feeling et al. 2017: 129)			

(23)	ᏴᏫ	ᏧᎾᏍᏗᎢ	ᏓᏂᏠᏎᎲᎢ
	[**Yvwi**	**j-u-n-sdi?i**]	d-a-ni-hloseh-v?i
	People	DST-3-PL-little	DST-3-PL-blame-EXP
	‘They blamed the little people.’ (Feeling et al. 2017: 42)		

We also tag the sole arguments of three types of intransitive predicates as themes. This includes *change-of-location predicates* (go, come, arrive, etc.), *change-of-state predicates* (fall, die, break, etc.), and *continuation-of-state predicates* (live, lie, stand, etc.). For example, in example (24) *jiyu* ‘canoe’ undergoes a change of state, and *anisgay* ‘men’ undergoes a change of location. Example (25) shows a theme argument of a change-of-location predicate, and (26) shows a theme of a change-of-state predicate.

(24)	ᏚᏟᏆᏗᏁᎴ	ᏥᏳ	ᏚᏂᎬᏤᏃ	ᎠᏂᏍᎦᏯ
	d-u-hlihgwadinel-e	[**jiyu**].	d-u-ni-gvje=hno	[**a-ni-sgay**].
	DST-3-turn.over-REPP	canoe	DST-3-PL-fall.in=CN	3-PL-man
	‘The canoe turned over, and the men fell (into the water).’ (Feeling et al. 2017: 111)			

(25)	ᎥᏍᎩᏃ	ᏄᎾᏛᏁᎳ	…
	vsgi=no	n-u-n-advnel-a	
	That=CN	SPEC-3-PL-do-PRES	
	‘When they did that … ’		
	ᏭᏅᎪᏣᏊ	ᏦᏍᏓᏓᏅᏟ	
	w-u-n-vsgoj-v=gwu	[**j-osd-adanvdli**]	
	TR-3-PL-go.out-EXPP=DT	DST-1.DUAL.EXCL-brother	
	‘my brother just went out.’ (Feeling et al. 2017: 101-102)		

(26)	ᏒᏃᏱ	ᎠᏟᎢᎵᏒ	ᎤᏲᎰᏎ	Ꮎ	ᎤᏛᏐᏅ	ᏧᏢᎬᎢ
	svnoyi	ahli?ilisv	u-yohus-e	[**na**	**utvsohnv**	**j-u-dlv-g-v?i**].
	midnight	time	3-die-REPP	that	old.man	REL-3-sick-PROG-EXP
	“The old man who had been sick died that night.” (Feeling et al. 2017: 26)					

GOAL arguments express the recipient or beneficiary of an action, which typically occur with ditransitive predicates that can also take an agent and theme arguments. An example is given in (27). Expressions that only denote the endpoint in space of an event are labeled instead as DYNAMIC LOCATION expressions, described later in this section. Due to their rarity within the corpus, goal arguments are not examined in the quantitative analyses in Sections 3 and 4. However, they are relevant to the discussion in Section 5 of word order in sentences with more than one non-verbal major constituent.

(27)	ᏏᎸ	ᏫᏗᎢᏅᏍ	ᎡᏥ	ᎯᎠᎾ	ᎠᏂᎪᎳ
	silv	wi-di-ji-nvhs	[**e-ji**]	hi?a=na	a-ni-gola
	first	TR-DST-PL-go	1.POS-mother	this=CN	3-PL-perch

“First, I will take this perch to my mother.’ (Feeling et al. 2017: 216)

For predicates that express perceptual or psychological states, we used the tags STIMULUS for entities that the cause of a mental state or sensory perception, and EXPERIENCER for the entities that experience the corresponding state. In (28), *jiistvvna* ‘crawdad’ is an experiencer argument, and *kaniitaʔv* ‘tail’ is a stimulus.

(28)	ᎦᏂᏔᏛ	ᏭᎪᎮ	ᏥᏍᏛᎾ
	ka-niitaʔtv	w-u-kooh-e	**jiistvvna**
	3-tail	TRN-3-see-REPP	crawdad

‘The crawdad saw the wolf’s tail.’ (Montgomery-Anderson 2008: 552)

We use two other thematic role tags for the arguments of equative verbal predicates, primarily *be* or *be called/named*. We use the tag PREDICATE SUBJECT (abbreviated as PRED-SUB) for subjects of these predicates. Informally, these are the “modified” items in the attribution or identity relation.

(29)	ᎠᏂ	ᏍᏊ	ᏚᏙᎥ	**ᎠᎩᏥ**	**ᏥᎨᏒ**
	An	sgwu	d-u-doʔ-v	[**agi-ji**	**ji-ges-v**]
	Ann	also	DST-3-be.named-EXP	1.POS-mother	REL-be-EXP

‘My mother was also named Ann.’ (Feeling et al. 2017: 84)

(30)	ᏍᎩᏅ	ᎠᏑᏂᏙ	ᎨᎲ
	[**Sgi=hnv**]	asuhnidoh	ge-hv
	That=CN	fisher	be-EXPP

‘He was a fisher.’ (Feeling et al. 2017: 215)

On the flip side, we use the tag PREDICATE OBJECT (PRED-OBJ) for the object argument of the same predicates. Informally, these express an attribute of the entity denoted by the subject. These can be illustrated with the preceding two sentences, repeated below as (31)-(32).

(31)	**ᎠᏂ**	**ᏍᏊ**	ᏚᏙᎥ	ᎠᎩᏥ	ᏥᎨᏒ
	[**An**	**sgwu**]	d-u-doʔ-v	agi-ji	ji-ges-v
	Ann	also	DST-3-be.named-EXP	1.POS-mother	REL-be-EXP

‘My mother was also named Ann.’ (Feeling et al. 2017: 84)

(32) **ᏍᎩᏅ** **ᎠᏑᏂᏙ** ᎨᎲ
Sgi=hnv [**asuhnidoh**] ge-hv
That=CN fisher be-EXPP
'He was a fisher.' (Feeling et al. 2017: 215)

Looking at all items in the corpus with an argument thematic role tag, we again see an overall preference for items of each type to precede verbs, but with several notable differences among different thematic role values. First, we find that PRED-OBJ is the only type of item in the corpus that occurs without exception in a preverbal position. This is consistent with the acceptability judgment studies by Scancarelli (1987) and Akkuş (2018), who describe this as the only inviolable ordering restriction in copular structures (PRED-SUB items can occur in any order relative to verbs and PRED-OBJ items). Second, we find that agents are more likely to show preverbal placement (78%) than themes (64%). This is consistent with cross-linguistic tendencies, as well as the description of Pulte & Feeling (1975). Section 4 presents further evidence that the distinction is statistically significant in the corpus.

	Agent	Experiencer	Pred-obj	Pred-sub	Stimulus	Theme
postverbal	18	5	0	3	10	88
preverbal	64	16	23	18	16	156

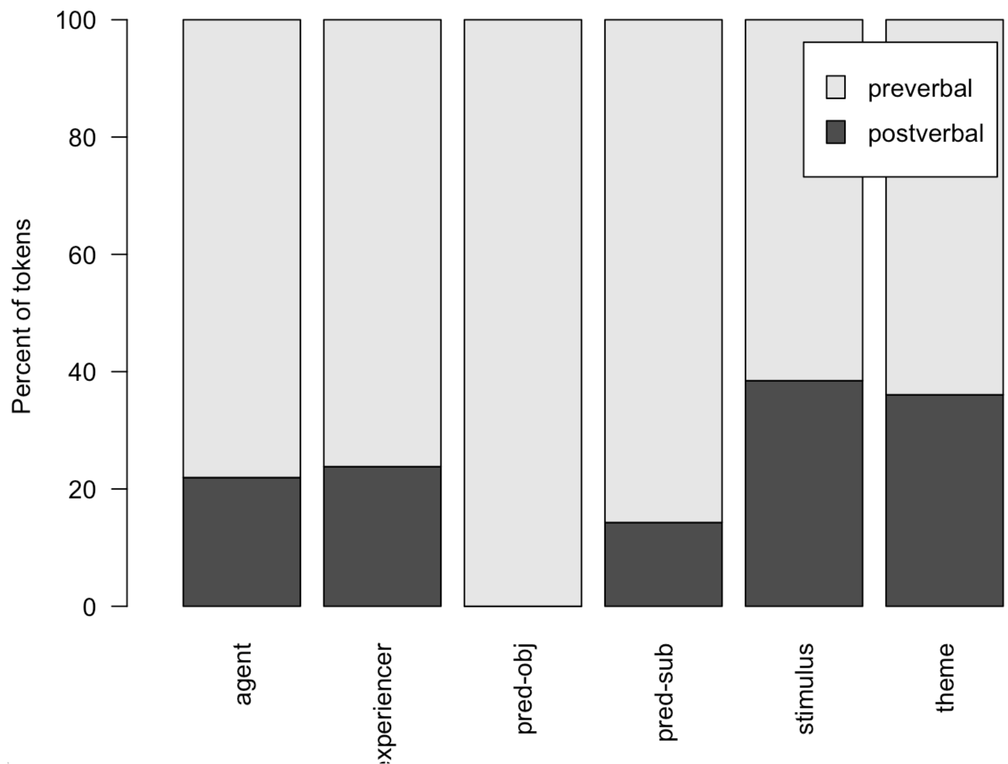

FIGURE 2: Placement relative to verbs for argument thematic roles

Finally, examine the word order patterning of several adjunct thematic roles. TIME, STATIC LOCATION, and DYNAMIC LOCATION. Time items include standalone adverbials like *hleg* '(for a)

while' in example (33). This tag is also used on adverbial clauses that specify the time of a main clause event, as in example (34).

(33) **ᏝᏕᏫ** **ᎠᏎ** **ᎠᎾᏂᏢᏯᏬᏞ** **ᎩᏟ**
[**hleg=hnv**] ase a-n-anhdlvs-g-e gitli.
while-CN maybe 3-PL-lie.down-PROG-REPP dog
'The dogs would lie down for a while.' (Feeling et al. 2017: 81)

(34) **ᎩᏆ** **ᏀᏢᏆ** **ᏴᎾ** **ᎾᎥ** **ᏂᏚᎾᏓᏁ**
[**Kilo** **y-u-dlvj-a**] yvwi na?v ni-d-u-n-adal-v ...
Someone IRR-3-get.sick-PRES people near SPEC-DST-3-PL-apart-EXPP
'When someone got sick, people in the neighboring area...'

ᏓᎾᏓᏩᏍᎯᏙᏯ
d-a-n-ada-watvh-idoh-v
DST-3-PL-REFL-visit.around-EXPP
'...would visit.' (Feeling et al. 2017: 23)

Static location expressions refer to the delimiting spatial location of an event or description, as in the two examples below.

(35) **ᏂᏀ** **ᎤᏯᏗ** **ᎤᎾᏦᏕ**
[**jiyu** **usdi**] u-n-ajod-e
canoe small 3-PL-be.in-REPP
'They were in a small canoe.' (Feeling et al. 2017: 109)

(36) **ᏫᏎ** **ᏯᎬᏏ** **ᎾᎥ** **ᏒᏬ**
[**ge=hnv** **oaks-i** **na?v**] e-h-e …
there=CN Oaks-LOC near 3-live-REPP
'Near the town of Oaks, there lived …'

ᎠᏕᏴᎱᏞ **ᎤᎵᏯᏕᏯᏍᎤ** **ᏧᏙᎢᏓ**
a-gayvlige ulsgasd j-u-do?id-a
3-old.woman Ulsgasd REL-3-name-REPP
'an old woman named Ulsgasd.' (Feeling et al. 2017: 78)

In contrast, dynamic location expressions express the direction of movement or path of movement of an action.

(37) **ᏦᎩ** **ᎠᏕᏗ** **ᎠᏂᏕᏬᏍᏬᏒ**
[**Jog** **akti**] a-ni-gawehih-e
Upstream toward 3-PL-paddle-REPP
'They were paddling upstream.' (Feeling et al. 2017: 109)

(38) ᎯᏓᏁᎭ ᏩᏩᏗᎾᏒ
[**Didanelv**] w-awadinvs-v
Home TR-throw-EXP
'I threw (it) towards home.' (Feeling et al. 2017: 218)

As shown in the tables below, static location and time expressions show a strong preference for preverbal placement (87% preverbal and 91.2% preverbal, respectively), greater than all argument thematic roles except for predicate objects (categorically preverbal). This is consistent with results to be presented in Section 5, which suggest that static location and time expressions typically precede all other constituents within a clause. In contrast, the distribution of dynamic location items (70% preverbal) more closely resembles that of the argument expressions discussed above.

	Dynamic location	Static location	Time
postverbal	27	6	12
preverbal	64	41	125

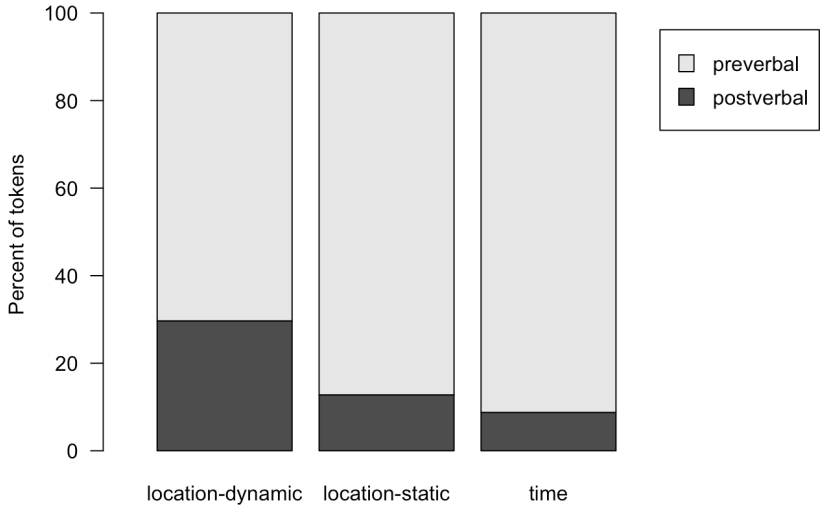

Figure 3: Placement relative to verbs for adjunct thematic roles

3.5 *Other potential factors*

This section summarizes potential word factors that were tagged in the corpus, but ultimately not found to have significant effects in the regression model. We also briefly discuss potential word order factors that were not investigated in the study.

3.5.1 *Animacy*

The animacy of nominal arguments has been found to be an independent conditioning factor in some word order patterns (Brody 1984; Rosenbach 2005). While effects of animacy on clausal word order in Cherokee have not been previously described, animacy distinctions do influence verbal agreement in some circumstances. First, some agreement prefixes in the set A class express the combination of a first-person inclusive or second-person agent and an animate third-person theme. Second, on some transitive verbs, animacy conditions the choice between prefix forms (set

A vs. set B) when there are two third-person arguments; the set A form occurs with an animate agent and inanimate theme, while the "inverse" set B prefix occurs with an inanimate agent and animate theme. The occurrence of inverse forms, however, is not known to correlate with word order preferences among nominal arguments and verbs. Nonetheless, given the influence of animacy on some aspects of Cherokee grammar, it is reasonable to consider the possibility that the property has independent word order effects. Animacy is also worth investigation as a potential confound with thematic roles, as agent arguments most often refer to animate beings, and inanimate-referring expressions are more likely to be theme arguments.

We tagged all nominal expressions in the corpus for animacy properties, excluding location and time arguments. These items were not tagged for animacy because they do not seem to have comparable referentiality properties as argument nominals; as discussed in Section 4.3, they have a strong tendency to systematically precede all other nominal expressions. Expressions referring to humans, animals, and mythical beings were tagged as ANIMATE and all other entities as INANIMATE.

3.5.2 *Length*

In some language patterns, longer constituents show a greater propensity to occur in a peripheral positions of the clause, with language-particular variation in which edge is preferred (Behagel 1909; Hawkins 1994). Constituent length is also a potential confound for referential accessibility; one may expect constituents with discourse-new referents to be longer on average than constituents with discourse-given referents, which are more likely to be realized as personal or demonstrative pronouns (relatively short in length, generally).

To investigate possible effects of length on word order, all verbs and major constituents in the corpus were coded for the absolute number of characters in the romanized Cherokee transcriptions provided by Feeling et al. (2017) and Montgomery-Anderson (2008). Given that romanized Cherokee is largely faithful to IPA pronunciation (most characters correspond to one IPA segment – except *ch*, *tl*, *ts*, *kw*), we consider this to be a reasonable approximation of phonological segmental length. In the statistical analysis, we tested two types of measures as predictor variables for the placement of non-verbal constituents: the absolute length of the constituent, and the difference in length between each the constituent and the length of the main verb of the clause.

3.5.3 *Potential factors not investigated*

We briefly mention other grammatical categories or properties that were not directly investigated, due to the difficulty of identifying them in our corpus. First, we did not directly investigate the effects of aboutness topichood (Reinhart 1982) on word order, as it is generally difficult to identify topics from textual properties alone. It is relevant to note that framesetting elements like time and location expressions, considered a type of topic in some theories (Chafe 1976), were tagged and investigated (see Section 5.3). Another information-structural category that we did not directly investigate is narrow focus, which we understand as arising in sentences where only one constituent represents non-presupposed, or non-backgrounded information. Some proposed subtypes of focus overlap with properties that were tagged, but without an exact correspondence. For example, the CONTRAST tag can be used with on constituents that would be described as bearing contrastive focus. However, contrast can also occur on non-focused items, like contrastive topics (Vallduví and Vilkuna 1998; Neeleman et al. 2007; Molnár 2002; Aissen to

appear). Similarly, while discourse NEW items can sometimes pattern as foci, we cannot assume that all new-information constituents are necessarily focused. Generally, the various types of focus are difficult to identify in monologue narratives of the sort in our corpus, and are more reliably observed and distinguished in in dialogues with questions, or by using traditional elicitation methods (Aissen to appear).

Finally, as the works in our corpus do not have associated audio files, our work has not investigated the prosodic and/or intonational correlates of word order choices in Cherokee. We are also not currently aware of existing formal description of the intonational correlates of information-structural marking in the language. These critical questions must be left for future work.

4 Word order factors in Cherokee

In this section, we first present the results of a logistic regression model, which we use to identify the properties that are the most robust and independent predictors of word order preferences in the Cherokee corpus. We then present a closer examination of the cumulative interaction of information-strutural and thematic properties in word order preferences, in relation to prior claims about Cherokee grammar.

4.1 *Regression model results*

We used a mixed-effect logistic regression model to evaluate the reliability of the annotated grammatical properties as conditioning factors on Cherokee word order in the corpus. Modeling was done with the statistics software R, using the glmer( ) function of the lme4 package (Bates, Maechler, and Bolker 2013). Within the model, we used two word order values PREVERBAL, POSTVERBAL as the dependent variables, with referential accessibility, contrast, thematic role, animacy, and the two length measures (absolute constituent length, difference in length between the constituent and the verb) as independent variables. The numeric length variables were centered and standardized to improve comparisons with the other non-numeric variables. Finally, speaker identity was used as a random variable.

Several types of expressions are omitted from the input to the presented regression model for the following reasons. TIME expressions are not included as there is not a clear way to annotate them for a referential accessibility value. The presented model also omits items with the thematic roles PREDICATE OBJECT, PREDICATE SUBJECT, EXPERIENCER, and STATIC LOCATION, as the models that include them always produce small coefficients less than one-half times their standard error. Recall that predicate objects precede verbs without exception in the corpus, so we maintain that they are a robust predictor of word order despite their omission from the regression analysis (independent variables with categorical effects are not reliably captured in regression models due to high standard error estimates).

The model that we present below was obtained after a nested model comparison. At each step, we removed the factor with the lowest ratio of its coefficient over its standard error (using absolute values), as long as none of its individual sub-values (i.e. individual thematic roles, referential accessibility values) had a statistically significant effect.

Before turning to the results of the model below, we briefly explain how to interpret it. The reference categories (the intercept) consists of items tagged as NON-REFERENTIAL, AGENT, and CONTRAST, the categories with the greatest propensity for preverbal placement. The word order

effects of each information-structure and thematic tag are shown in the rows below. For each tag, a positive coefficient in the Estimate column indicates a greater likelihood of preverbal placement, and a negative coefficient indicates a greater likelihood of postverbal placement than the reference category. Variables identified by the model as statistically significant predictors are shown with asterisks (*) at the end of some rows; the number of asterisks corresponds to *p*-value thresholds, such that more asterisks indicate a lower *p*-value (greater confidence in the predictive power of the variable).

420 major constituents 9 speakers

FACTOR	ESTIMATE	STD. ERROR	Z VALUE	PR(>|Z|)	
(Intercept)	2.5321	0.4746	5.336	9.51e-08	***
REFERENTIAL ACCESSIBILITY FACTORS					
accessible	-1.2011	0.4598	-2.612	0.008991	**
given-active	-1.6286	0.4150	-3.924	8.70e-05	***
given-inactive	-1.5711	0.4463	-3.520	0.000431	***
new	-0.3130	0.4581	-0.683	0.494358	
THEMATIC FACTORS					
location-dynamic	-0.6168	0.3818	-1.615	0.106254	
stimulus	-1.2218	0.5206	-2.347	0.018939	*
theme	-0.9631	0.3188	-3.022	0.002515	**
CONTRAST					
no contrast	-2.1769	1.0601	2.053	0.040034	*

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

RANDOM EFFECTS

Speaker (intercept) Variance: 0 Standard Deviation: 0

LIKELIHOOD AND DEVIANCE

AIC	BIC	logLik	deviance	df.resid
507.1	547.5	-243.5	487.1	410

TABLE 4: Regression modeling estimates

First, we note that the only significant factors in model are properties related to referential accessibility, contrast, and thematic role. Distinctions in animacy and constituent length (both absolute length, and the difference in length relative to the verb) were not found to be significant predictors, with animacy having the least predictive power. While this result does not let us conclude that these factors play no role in the ordering of major constituents in the clause (they could potentially show a significant effect in a larger corpus), it suggests that these factors play a relatively minor role in this domain of Cherokee grammar. On the other hand, the factors that were

found to be statistically significant are likely to play a larger role in determining clausal word order in the language.

The model finds that features related to referential accessibility, contrast, and thematic role are statistically significant predictors of preverbal versus postverbal ordering of major constituents. Looking at information structure properties, the model does not find a significant difference between non-referential constituents (the reference category) and new information, but all types of discourse-accessible and discourse-given items show a statistically significant greater propensity for postverbal placement relative to non-referential constituents. As for thematic properties, the model finds a statistically significant difference between agents versus stimulus and theme arguments. Finally, items that bear contrast are likelier to precede verbs than items that do not bear contrast.[13]

To verify whether sentences with more non-verbal major constituents follow distinct word-order restrictions, We repeated the same procedure on a subset of the corpus that includes only sentences with one item per clause (roughly 70% of major constituents in the full corpus). This yielded the same inventory of statistically significant predictors, with the exception of thematic stimuli (likely due to its relatively small number of examples). This result suggests that longer clauses do not follow dramatically different word order principles from those with only one major constituent.

The regression model suggests several important findings about Cherokee grammar. First, it confirms the claims of Pulte and Feeling (1975) and King (1975) that thematic properties play a role in clausal word order in Cherokee, albeit in a probabilistic, rather than categorical way. Specifically, our results suggest that both thematic distinctions and information-structure properties interact cumulatively in determining the preferred placement of individual constituents within the clause. The following subsection presents other measurements within the corpus that reflect the cumulative interaction of properties related to thematic structure and referential accessibility.

4.2 *Cumulative effects of thematic structure and referential accessibility*

First, we note that even when we restrict our attention to items of the same thematic role, we see similar effects of information-structural differences on the placement of nominal constituents. Figure 4 below shows the attested placement of theme nominals only, for each information structure specification. We observe the same trends that we identified based on all nouns in the corpus (see Section 3.4); New and nonreferential items are the most likely to precede verbs, and discourse-given items the least likely to precede verbs.

[13] We considered an alternative model that includes interaction terms between these independent variables, and found one significant interaction term THEME * GIVEN-INACTIVE, suggesting that these properties combined have a super-additive effect on the probability of postverbal placement.

	Accessible	Given-active	Given-inactive	New	Nonreferential	Total
postverbal	13	38	20	13	6	90
preverbal	19	41	16	47	37	160

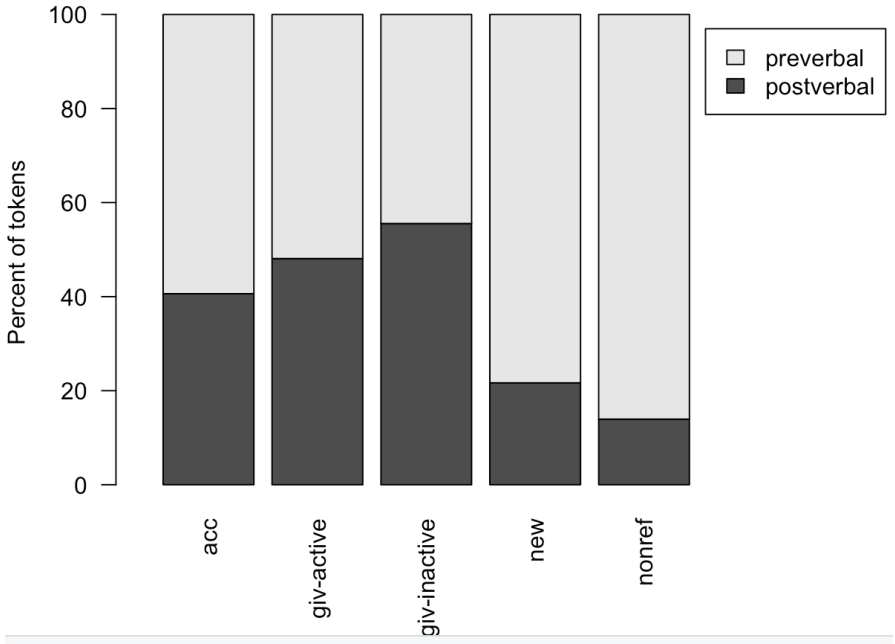

Figure 4: Referential accessibility and placement relative to verbs, for theme arguments only

We find similar effects of referential accessibility when we examine only agent nominals, shown below in Figure 5. While the data is relatively limited (there are fewer agents than themes in the corpus), we again find that new items are more likely to precede verbs than accessible and given items. Somewhat surprisingly, we find a relatively large difference in patterning between active given vs. inactive given items. While it is plausible for languages to show a greater preference for placing inactive given items (which are often shifted topics) early in the clause, the pattern is unexpected given that this effect is not observed in the full corpus (which includes items of all thematic roles). Due to the relatively small number of examples that we have for agents in our corpus, we remain agnostic on the significance of this pattern.

	Accessible	Given-active	Given-inactive	new	nonreferential	Total
Postverbal	4	10	1	1	0	16
Preverbal	13	21	17	13	3	67

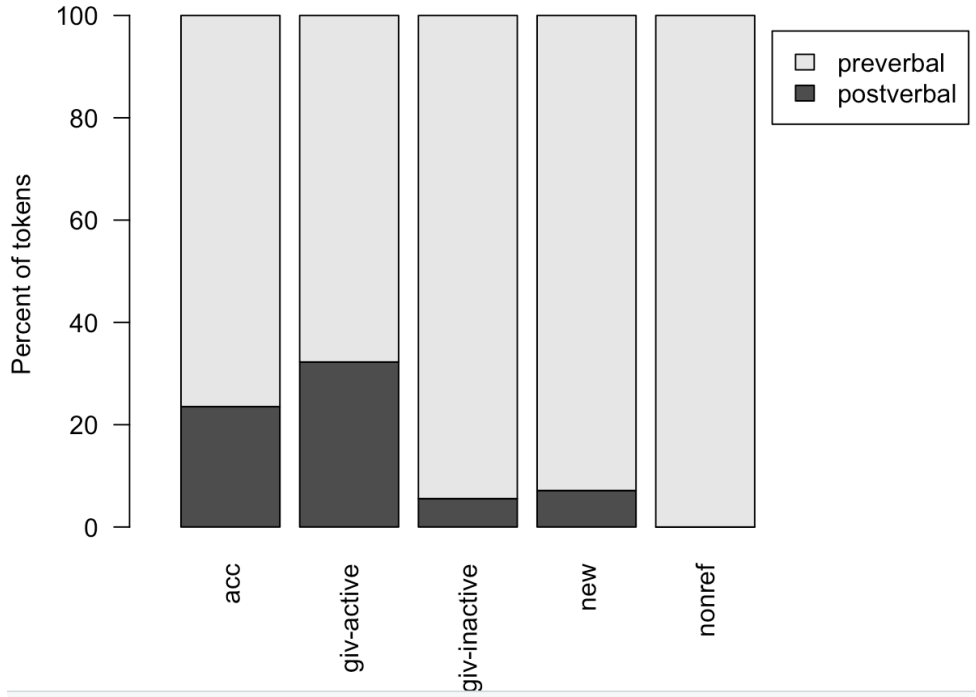

Figure 5: Referential accessibility and placement relative to verbs, for agent arguments only

It is also informative to compare this pattern to the behavior of agents in Figure 4. The key observation is that theme arguments of all information structure values are less likely to occur in a preverbal position than agents that have the same information structure property. For example, accessible themes are 59% preverbal while accessible agents are 24% preverbal.

The cumulative interactions among the ordering preferences associated with information-structural and thematic properties can also be observed in the cross pair table below. Each pair of rows and columns compares two properties with a relatively high difference in their ordering preferences; the rows compare new versus given information, and the columns compare agents versus themes. Nouns that refer to agents are uniformly more likely to precede verbs than those that refer to themes (compare the left column to right column). Nouns that refer to new entities are more likely to precede verbs than nouns that are known entities (compare top row to bottom row). A noun that has two properties favoring preverbal placement (agent and new information) shows a greater likelihood of preceding the verb than nouns with only one such property (given agents and new themes).

	Noun is agent	*Noun is theme*
Noun is new	**92%** preverbal (12/13)	**73%** preverbal (36/49)
Noun is given	**76%** preverbal (37/49)	**50%** preverbal (57/114)

Table 5: Percentage of Cherokee corpus sentences with *noun* > *verb* order, depending on properties of the noun.

Overall, the results of the model support our claim that multiple types of grammatical properties can contribute, probabilistically but systematically, to word order preferences in Cherokee. These results let us conclusively revisit the the claim by Mithun (1992: 58) that "the order of constituents does not reflect their syntactic functions at all" in languages like Cherokee. If one reasonably takes syntactic functions to include thematic structure, we would expect to find no effect of thematic role on word order preferences. The findings in our corpus analysis cast strong doubt on this claim. We suggest that there is no strict macroparameter that differentiates languages that rely on information structure versus thematic role for clausal word order. Rather, the syntactic component of all language grammars can access both types of properties, with potentially fine-grained variation across languages in the extent to which they determine word order preferences in various domains of structure. It is worth emphasizing that the claim that thematic roles partly determine Cherokee clausal word order is *not* equivalent to the idea that there is a single “basic” word order from which all others are derived (as argued against by Mithun 1992). The crucial claim is that the mental grammar of the Cherokee speaker can access both thematic and information-structural properties when determining how to order a set of items in a clause.

5. Word order in longer clauses

Thus far, we have discussed grammatical properties that influence the order of certain major constituents relative to verbs. This section specifically examines clauses that that contain more than one major constituent other than the verb, focusing on the relative order of these items. While the portion of the corpus that consists of sentences with more than one non-verbal item (170 sentences; 30% of the corpus) is too small for a robust quantitative analysis using all of the word order predictors we have identified, we can nonetheless observe several word tendencies that are highly consistent with the results from the previous section.

5.1 *Placement of major constituents relative to verbs*

There is a tendency for expressions of all types to precede verbs, abstracting away for the moment from other grammatical properties. Among the 170 clauses in the corpus that contain two or more major constituents other than the verb, 67 clauses (40%) contain only preverbal items, 100 clauses (59%) contain one postverbal item and at least one preverbal item, and 3 clauses (2%) contain two postverbal items.

	Only preverbal items	One postverbal item, one or more preverbal items	Two postverbal items	Total
numbers	67	100	3	170
percent	40%	59%	2%	100%

Table 6: Word order in clauses with more than one non-verbal major constituent

While it relatively easy to find sentences with multiple preverbal constituents, sentences with multiple postverbal items are quite rare in comparison. Put another way, we see that verbs in Cherokee are almost invariably either the last or second-to-last item in their clause. The pattern is largely expected, based on what we observe in clauses with only one non-verbal constituent. In clauses with only one major constituent other than the verb, 24.6 percent of these expressions occur in a postverbal position; the predicted combined probability of finding two postverbal constituents in a clause with two such items is 6 percent, quite close to what is observed (2 percent). This suggests that ordering preferences on major constituents in Cherokee does not depend highly on the number of major constituents per clause.

It is noteworthy that in all three of the sentences in the corpus with two postverbal items, all postverbal constituents express given information. Two of the sentences (39) and (40) contain a postverbal theme and goal of a ditransitive predicate, while sentence (41) has a postverbal agent and theme. With the exception of the agent in the last example, these are the same types of items that are most likely follow verbs in shorter sentences.

(39) ᏃᏭᏃ ᏣᎵᏟᎭᏭᏴ ᎠᎷ Ꮓ ...
nown j-a-ljihawyv am no
then REL-3-start.to.boil water then
'When the water is just beginning to boil ...

ᏕᎦᏑᏴᏍᎪ ᏗᎵ ᎦᏚ
de-gasuyvs-g-o [tili] [gadu].
DST-1.mix.in-PROG-HAB chestnut bread
'I mix the chestnuts into the batter (bread).' (Feeling et al. 2017: 129)

(40) ᏏᎸ ᏫᏗᎢᏅᏍ ᎡᏥ ᎯᎠᎾ ᎠᏂᎪᎳ
[silv] wi-di-ji-nvhs [e-ji] [hi?a=na a-ni-gola].
first TR-DST-PL-go 1.POS-mother this=CN 3-PL-perch
'First, I will take this perch to my mother.' (Feeling et al. 2017: 216)

(41) ᏄᎾᏛᏁᎴ ᎠᏂᏐ ᏓᎦᏏ ᏚᏃᏎᎸ
N-uun-tvvneel-e [anii-so?] [taks t-u-hnooseel-v].
PRT-3B-do-REPP 3A.PL-other turtle DST-3B-tell-EXPP
'The others did what the turtle told them.' (Montgomery-Anderson 2008: 564)

5.2 *Relative ordering of agent and theme arguments*

To evaluate whether thematic role affects ordering preferences among nominal constituents, it is useful to examine sentences in which all argument nominal constituents have the same level of referential accessibility (we examine the patterning of time and location adjuncts separately in the next subsection). Our corpus contains 10 sentences of this type, which contain an agent and a theme (it is quite uncommon in general for copular predicates to have a subject and object of the same information status, and we do not find any examples in the corpus). While we again find some variability in how these items are ordered, there is an apparent preference for agent arguments to precede themes.

	Agent precedes theme	Theme precedes agent	Total
numbers	8	2	10
percent	80%	20%	100%

Table 7: Relative orders of nominal expressions with identical referential accessibility tags

This dominant trend is illustrated with the examples (42)-(44) below. In each sentence, all nominal constituents have the same referential accessibility status. The ordering tendencies appear to hold independently of whether the verb occurs between both arguments as in (42) and (43), or whether both arguments precede the verb as in (44).

(42) *Agent > V > theme*. All nominals are given-inactive.

ᎡᏙᏓ **ᏥᎨᏒ** ᏃᏊ **ᏗᏱᎬ** **ᎦᎸᎳᏗ** **ᎢᏗᏢ**
[e-doda ji-ges-v] nogwu didluhgv galvladi ididlv ...
1.POS-father REL-be-EXP now tree up toward

ᏫᏚᎵᏐᏍᏔᏅ **ᎤᏒᏕᎾ** **ᎦᎶᏇ**
wi-d-u-lisostan-v [usvdena galogwe]
TR-DST-3-aim-EXP big-barrel gun
'My dad aimed the shotgun up toward the branch.' (Feeling et al. 2017: 15)

(43) *Agent > V > theme.* All nominals are new.

ᏓᎦᏏ ᏚᎩᏴ ᏥᏍᏚ
[taks] t-uu-khiiy-v [jiist]
Turtle DST-3B-beat.in.race-EXP rabbit
'The turtle beat the rabbit.' (Montgomery-Anderson 2008: 561)

(44) *Agent > theme > V.* All nominals are accessible.

ᏌᏊᏃ **ᏧᎾᏔᎾ** **ᎠᏫ** ᏓᎯᎮ
[sagwu=no] [j-u-n-atana ahwi] d-a-hih-e.
one=CN DST-3-PL-big deer DST-3-kill-REPP
'One (of the men) killed big deer.' (Feeling et al. 2017: 53)

5.3 *Placement of location and time expressions*

Here, we examine the positions of time and location expressions, relative to other items in the clause.[14] In brief, we find that time expressions show the most robust tendency to precede all other expressions in the clause (80% are clause-initial). Static location expressions show a lighter tendency for early placement in the clause (56% clause-initial), followed by dynamic location expressions (30% clause-initial). This is consistent with the cross-linguistic tendency for FRAME-SETTING items, which delimit the context in which an event ocurs (as most time and static location expressions do), to occur in clause-initial positions and precede argument expressions (Speyer 2008; Wolfe 2015). In contrast, dynamic location expressions, which often have a greater selectional relation with event-denoting verbs, do not show the same pattern.

As shown in Table 8, a large majority of time expressions (80%) occur as the first item in their clause. They occur with substantially less frequency in clause-medial positions (14%), and even more rarely in clause-final positions (6%). All clause-medial constituents in these examples precede the main verb; recall from Section 5.1 that there are very few sentences with more than one postverbal constituent.

	Clause-initial preverbal	Clause-medial preverbal	Clause-final postverbal	Total
Numbers	51	9	4	64
Percent	80%	14%	6%	100%

Table 8: Position of time adverbials in longer sentences

In some cases, the position of time elements seems to be conditioned by scopal differences, as illustrated in the sentence in (45). The clause-initial expression *sagwu-hno iyuwakdi* 'one time' modifies a full proposition, wherease the clause-medial time expression *usv* '(at) night' modifies a subordinate proposition 'walking home.' Similarly, in (46) the postverbal time expression *hlega* '(for a) while' modifies only the subordinate proposition 'become silent.' However, this explanation does not seem to apply in all cases; the time expression *kohi iyv* 'after a while' in (47) appears to modify the full proposition, even though it occurs in a clause-medial position.

(45) ᎿᏩZ TᏀGAᏗ RᎮᏏ RSSZ
[**sagwu=hno** **iyuwakdi**] e-lisi e-dudu=hno ...
one=CN time 1.POS-grandmother 1.POS-grandmother=CN
'One time, my grandmother and my grandfather ...'

ᎤᏫᎾ TDᎾT4 ᏧᏁ10ᏉR TᏗᏕ
[**usv**] i?-a-n-a?is-e j-u-nenvsv ididla
night ITR-3-PL-walk-REPP DST-3-home toward
'... were walking home at night.' (Feeling et al. 2017: 23)

[14] Here, we do not take information structure features directly into account, as time and location items are often difficult to classify in terms of the givenness-related properties that characterize argument nominal expressions.

(46) ᏦᏍᏓᏓᏅᏟᏃ ᎡᏝᏪ ᏄᎵᏍᏔᏅ ᏞᎦ
j-osd-adanvdli=no ehlawe n-u-listan-v [**hlega**]
DST-1.DUAL.EXCL-brother=CN silent SPEC-3-become-EXPP while
'My brother became silent for a while.' (Feeling et al. 2017: 102-103)

(47) ᎡᎵᏏᏃ ᏃᏊ ᎪᎯ ᎢᏴ ᎤᏂᏨ
E-lisi=hno nogwu [**kohi iyv**] u-hnej-v
1.POS-grandmother=CN then after a.while 3-speak-EXP
'After a while, my grandmother spoke.' (Feeling et al. 2017: 26)

Static location adverbials are also most likely to occur at the beginning of the clause, as shown in Table 9 below. However, they are somewhat likelier than time adverbials to occur in a clause-medial or clause final position. It is relevant to note, however, that three of the five clause-medial items in the corpus are preceded only by a time adverbial. In contrast, none of the clause-initial static location expressions are followed by a time expression. Thus, 66 percent of static location adverbials precede all non-time expressions and the verb in their clause.

	Clause-initial preverbal	Clause-medial preverbal	Clause-final postverbal	Total
Numbers	15	5	7	27
Percent	56%	19%	26%	100%

Table 9: Position of static location adverbials in sentences with more than one major constituent

The observations so far suggest that there is a strong tendency in Cherokee for time and static location expressions (as frame-setting items) to precede all other expressions in the clause, and for time expressions to precede static location expressions when both items are present: *time* > *static location* > *all other expressions*. This ordering tendency is exemplified by the sentence in (48), which contains a clause-initial time expression followed by a static location expression. Again, there are some exceptions to the dominant pattern; example (49) is a sentence with a postverbal static location expression.

(48) ᎥᎾᎾᏛᏍᏊ ᎾᎥ ᎤᏙᎯᏎ ᏥᏍᏛᎾ
vnawtvv=skwu [**na?v**] uu-athohis-e jiistvvna
right.then=DT near 3-whoop-REPP crawdad
'Right then beside him the crawdad whooped.' (Montgomery-Anderson 2008: 553)

(49) ᏝᎴ ᎥᏍᏊ ᎠᎾᎦᎵᏍᎩᏊ ᏱᏂᏙᎬᏁ
tla-le vsgwu a-n-galis-g-i-gwu y-i-ni-dogvn-e ...
NEG-and also 3-PL-flow-PROG-AG-just IRR-1-PL-have-REPP

ᎣᏥᏁᎸ ᎦᎵᏦᏕ
[**oj-inel-v** **galijode**]
1.PL.EXCL-live-EXPP house
'We didn't even have electricity in our house.' (Feeling et al. 2017: 99)

Finally, dynamic location expressions show the greatest degree of variability in placement, and occur much less frequently in a clause-initial position (30% of examples). In this respect, their distribution more closely resembles that of thematic argument expressions (eg. agents, themes, stimuli) than that of frame-setting expressions (time and static location).

	Clause-initial preverbal	Clause-medial preverbal	Clause-final postverbal	Total
Numbers	9	9	12	30
Percent	30%	30%	40%	100%

Table 10: Position of dynamic location adverbials in sentences with more than one major constituent

5.4 *Effects of referential accessibility on the relative order of nominal expressions*

Finally, we examine clauses that contain multiple constituents with distinct information structure tags. For this comparison, we abstract away from the thematic roles of the individual constituents, though we acknowledge that they likely have some effects on how items in these sentences are ordered. The table below shows the number of attested orderings for each pair of information structure features. While it is difficult to draw robust conclusions from the limited number of relevant sentences in the corpus, and we find variability in the ordering of each feature pair, we see two trends that are generally consistent with those that we have previously identified. First, there is a tendency for new items to precede given items, as consistent with observation that new items are more likely to precede verbs than given items. Second, we see that the relative order of accessible and given items is highly flexible, as consistent with the previously observed similarity in how they are ordered relative to verbs.

New precedes given	Given precedes new
7	1
New precedes accessible	Accessible precedes new
2	1
New precedes nonreferential	Nonreferential precedes new
1	2
Accessible precedes given	Given precedes accessible
7	5
Given preceedes nonreferential	Nonreferential precedes given
5	2
Nonreferential before accessible	Accessible before nonreferential
3	1

TABLE 11: Relative orders of nominal expressions with different referential accessibility tags

6. Discontinuous nominal expressions

Our corpus contains eight sentences with a discontinuous nominal expression, in which words that are interpreted as being part of a single nominal expression, are "split up" and occur on separate sides of the verb. All of these sentences are provided below, with bold text used to designate the content of each discontinuous expression. While one might consider the possibility that these sentences include two distinct nominal expressions in an appositive or parenthetical relation, such an interpretation does not seem plausible for most of the examples. Although Williams (1996) notes that discontinuous expressions can arise with possessors and possessor nouns, the broader pattern in Cherokee has not to our knowledge been previously discussed. A similar pattern is documented for its Northern Iroquoian relative Mohawk (Baker 1996).

Although the corpus examples do not allow us to conclusively know which types of noun-phrase-internal items can be "split," there are some noteworthy tendencies. For example, we posit that the types of items that occur in preverbal and postverbal positions conform to the language's general word order preferences related to information structure and thematic role. In each example in our corpus, the postverbal portion of the nominal expression includes the head noun, which in each case denotes a given or accessible entity. The preverbal portion of the nominal expression is most typically a quantifier (*jigwiya* 'too many,' *sagwu* 'one (of),' *ayehliyv* 'half (of)'), and in one case an adjective (55). All of these items express a subset relation between the referent of the full expression and the set of entities denoted by the postverbal head noun. Finally, we see a tendency for discontinuous expressions to bear a thematic role that has a relatively high propensity for postverbal placement; examples (50)-(54) include a theme argument and examples (55)-(56) have a predicate subject. There is only one example of a discontinuous nominal expression that is an agent (57).

(50) **ᏧᏘᏯ** ᎾᎾᎵᏍᏗᎲ **ᏴᏫ**
Tsgwiya n-a-n-alsdih-v **yvwi**.
Too.many SPEC-3-PL-begin.to.be-EXP people
'It happened to too many people.' (Feeling et al. 2017: 80)

(51) **ᏦᏊ** ᏚᎾᏘᏅᏎ **ᏧᎾᏓᎵ**
ko=gwu d-u-n-atinvs-e **j-u-n-adali**.
three=DT DST-3-PL-take-REPP DST-3-PL-spouse
'They took three of their wives.' (Feeling et al. 2017: 63-64)

(52) **ᏥᏘᏯ** ᏓᏛᏗ **ᎠᏂᎨ**
Jigwiya d-a-tvdi **a-ni-ge**.
Too.many DST-3-do.away 3-PL-woman
'(She) killed too many women.' (Feeling et al. 2017: 66-67)

(53) **ᏌᏊ** ᎣᎩᎷᎶᏤ **ᎠᏁᏦᎥᏍᎩ**
Sagwu ogi-luloch-e **a-hno?ejo?vsg-i**.
One they.and.I-lack-REPP 3-play.PROG-NOM
'We are missing one player.' (Feeling et al. 2017: 216)

(54) ᎤᎿᏅ ᏍᏈᏍᏙᏒ ᏚᏂᎪᎮ
uuhna=nv **skwisto-sv** t-uu-nii-kooh-e?i ...
there=CN a.lot-INT DST-3B-PL-see-REPP

ᎠᏣᏗ ᏓᏄᎩ ᎢᏳᏍᏗ
aja?ti **taahnuuk** **iyuust**.
fish gar like
"There they saw a lot of gar-like fish." (Montgomery-Anderson 2008: 559)

(55) Ꮎ ᏓᎦᏏ ᎤᏍᎦᏃᎵ ᎨᎲ ᎠᏟᏙ
na taks **uu-skanool** keeh-v **a-thliito**.
that turtle 3B-slow be-EXPP 3A-run
'The turtle was a slow runner.' (Montgomery-Anderson 2008: 561)

(56) ᎠᎹᏱᏅ **ᎠᏰᏟᏴ** ᎨᏎ **ᎾᏍᎩ** **Ꮎ** **ᎠᏔᎴᏒ**
Ama-yi=hnv **ayehliyv** ges-e **nasgi** **na** **atalesv**.
water-in=CN half be-repP there that hole
'Half of that hole was in the water.' (Feeling et al. 2017: 110)

(57) **ᏄᏓᎴ** ᏂᎦᏪᏍᎬ **ᏴᏫ**
Nudale ni-g-awes-g-v **yvwi**.
Different LAT-3-utter-prog-EXP person
'A different person was speaking.' (Feeling et al. 2017: 83)

7. Conclusion

In this paper, we used an annotated corpus analysis to identify several novel generalizations about clausal word order preferences in Cherokee. We have shown that word order in the language is principally determined by a variety of information-structural and thematic properties. These ordering principles are observed in the likelihood with which each factor influences the order of major constituents relative to verbs, and in preferences on the relative order among nominal and adverbial phrases.

We believe that the generalizations that we have identified based on naturalistic corpus data should be used as a basis for new pedagogical materials on word order in Cherokee. Specifically, we claim that language-learners can benefit from understanding the most common types of word orders, or the most likely associations between particular grammatical properties and positions in the clause, all while acknowledging the relatively high degree of flexibility that the grammar allows.

In addition, our results have key implications for several questions in syntactic theory. First, they suggest that word order preferences in languages with non-configurational and/or polysynthetic properties cannot be entirely attributed to information structure, and that thematic properties may play a more significant word order role in such languages than previously assumed. It further suggests that it is not productive to classify thematic structure and information structure as belonging to fundamentally different modules, such as "grammar" vs. "discourse." More broadly, our results suggest that the influence of any particular property on clausal word order in a language may be probabilistic, rather than "all or nothing." While we will not propose an

implementation of such an approach in this work, we note that the observed patterns of optionality and cumulativity in Cherokee appear to be quite compatible with stochastic, weighted constraint models of grammatical computation (Goldwater and Johnson 2003; Smith and Pater 2020), and formal syntactic theories that adopt these systems (Murphy 2017; Hsu 2021; Müller et al. 2022).

We leave open the possibility that there are more fine-grained distinctions in word order effects among the properties that we have examined (for instance, among subtypes of accessible items, or subtypes of themes), which may emerge from analyzing a larger corpus. We acknowledge that there could well be other types of grammatical properties that influence clausal word order in Cherokee, including prosodic or segmental phonological properties, other information structure properties related to topic and focus, nominal properties like quantification, and clause-level properties related to (irrealis) mood, negation, or tense and aspect. Investigation of these potential factors will require further study using a more extensive or differently-structured corpus, traditional elicitation methods, or some combination of both. We believe, however, that the qualitative and quantitative generalizations that we have identified in this work provide a new groundwork for such projects.

Finally, our corpus analysis would not have been possible without the availability of existing morphologically segmented and translated Cherokee texts. We would like to highlight the importance of efforts to add to these resources in publicly accessible forms, such as the Digital Archive of American Indian Languages Preservation and Perseverance (Bourns 2019) for Cherokee. These materials can be of essential value for language documentation and linguistic analysis, as well as for machine translation or other natural language processing tasks (Zhang et al. 2020a; 2020b).

References

AISSEN, JUDITH. To appear. Documenting topic and focus. *Language Documentation and Conservation*.

AKKUS, FARUK. 2018. Copular constructions and clausal syntax in Cherokee. *Proceedings of the Workshop on the Structure and Constituency of Languages of the Americas 21, University of British Columbia Working Papers in Linguistics 46*, ed. by Megan Keough, Natalie Weber, Andrei Anghelescu, Sihwei Chen, Erin Guntly, Khia Johnson, Daniel Reisinger, and Oksana Tkachman.

BADER, MARKUS. 2020. Objects in the German prefield: a view from language production. *Rethinking verb second*, ed. by Rebecca Woods and Sam Wolfe, 15–39. Oxford: Oxford University Press.

BADER, MARKUS, and JANA HÄUSSLER. 2010. Word order in German: a corpus study. *Lingua* 120.717–762.

BAKER, MARK C. 1996. The polysynthesis parameter. Oxford: Oxford University Press.

BATES, DOUGLAS.; MARTIN MAECHLER.; and BEN BOLKER. 2013. lme4: Linear mixed-effects models using 'Eigen' and S4. R package. https://cran.r-project.org/web/packages/lme4/

BEGHELLI, FILIPPO. 1996. Cherokee clause structure. Cherokee papers from UCLA, ed. by Filippo Beghelli, Barbara Blankenship, Michael Dukes, Edward S. Flemming, Pamela Munro, Brian Potter, Robert S. Williams, and Richard Wright, 105–114. Los Angeles: Department of Linguistics, University of California Los Angeles.

BEHAGEL, OTTO. 1909. Beziehungen zwischen Umfang und Reihenfolge von Satz gliedern. *Indogermanische Forschungen* 25.110–42.

BOURNS, JEFFREY. 2019. Cherokee syllabary texts: digital documentation and linguistic description. *2nd Conference on Language, Data and Knowledge (LDK 2019)*, ed. by Maria Eskevich, Gerard de Melo, Christian Fäth, John P. McCrae, Paul Buitelaar, Christian Chiarcos, Bettina Klimek, and Milan Dojchinovski, 18:1-18:6. Dagstuhl: Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

BRODY, JILL. 1984. Some problems with the concept of basic word order. *Linguistics* 22.711–736.

CHAFE, WALLACE. 1976. Givenness, contrastiveness, definiteness, subjects, topics and points of view. *Subject and topic*, ed. by Charles Li, 25–55. New York: Academic Press.

COOK, WILLIAM HINTON. 1979. A grammar of North Carolina Cherokee. Yale University.

DIPPER, STEFANIE; MICHAEL GÖTZE; and STAVROS SKOPETEAS (eds.) 2007. *Information structure in cross-linguistic corpora: Annotation guidelines for phonology, morphology, syntax, semantics, and information structure (ISIS Working Papers of the SFB 632)*. Potsdam: Universitätsverlag.

DRYER, MATTHEW S. 2013. Order of Subject, Object and Verb. *The World Atlas of Language Structures Online.*, ed. by Matthew S. Dryer and Martin Haspelmath. Leipzig: Max Planck Institute for Evolutionary Anthropology. http://wals.info/chapter/81.

ELLSIEPEN, EMILIA, and MARKUS BADER. 2018. Constraints on argument linearization in German. *Glossa: a journal of general linguistics* 3.6. 1-36.

FEELING, DURBIN, and WILLIAM PULTE. 1975. *Cherokee-English dictionary*. Talequah: Cherokee Nation of Oklahoma.

FEELING, DURBIN; WILLIAM PULTE; and GREGORY PULTE. 2017. *Cherokee narratives: a linguistic study*. Norman: University of Oklahoma Press.

FORTESCUE, MICHAEL, MARIANNE MITHUN, and NICHOLAS EVANS (eds.) 2017. *The Oxford handbook of polysynthesis*. Oxford: Oxford University Press.

FREY, BENJAMIN. 2020. "Data is nice:" Theoretical and pedagogical implications of an Eastern Cherokee corpus. *Collaborative approaches to the challenge of language documentation and conservation: Selected papers from the 2018 Symposium on American Indian Languages (SAIL)*, ed. by Wilson de Lima Silva and Katherine Riestenberg, 38–53. Honolulu: University of Hawai'i Press.

GOLDWATER, SHARON, and MARK JOHNSON. 2003. Learning OT constraint rankings using a maximum entropy model. *Proceedings of the Workshop on Variation within Optimality Theory*, ed. by Jennifer Spenader, Anders Eriksson, and Osten Dahl, 111–120. Stockholm: Stockholm University.

GUNDEL, JEANETTE K. 1988. Universals of topic-comment structure. *Studies in syntactic typology*, ed. by Michael Hammond, Edith A. Moravcsik, and Jessica Wirth. Amsterdam: John Benjamins.

HALE, KEN. 1983. Warlpiri and the grammar of non-configurational languages. *Natural Language & Linguistic Theory* 1.5–47.

HAWKINS, JOHN A. 1994. *A performance theory of order and constituency*. Cambridge: Cambridge University Press.

HSU, BRIAN. 2021. Harmonic Grammar in phrasal movement: an account of probe competition and blocking. *NELS 51: Proceedings of the 51st Annual Meeting of the North East Linguistic Society*, ed. by Alessa Farinella and Angelica Hill, 237–250. Amherst, MA: GLSA.

KING, DUANE HAROLD. 1975. A grammar and dictionary of the Cherokee language. Ph.D

dissertation, University of Georgia.
MITHUN, MARIANNE. 1992. Is basic word order universal? *Pragmatics of word order flexibility*, ed. by Doris L. Payne, 15–62. Amsterdam: John Benjamins.
MITHUN, MARIANNE. 1995. Morphological and prosodic forces shaping word order. *Word order in discourse*, ed. by Pamela A. Downing and Michael Noonan, 387–423. Amsterdam/Philadelphia: John Benjamins.
MITHUN, MARIANNE. 2017. The Iroquoian language family. *The Cambridge handbook of linguistic typology*, ed. by Alexandra Y. Aikhenvald, 747–781. Cambridge University Press.
MONTGOMERY-ANDERSON, BRAD. 2008. A reference grammar of Oklahoma Cherokee. Ph.D dissertation, University of Kansas.
MONTGOMERY-ANDERSON, BRAD. 2015. *Cherokee Reference Grammar*. Norman: University of Oklahoma Press.
MÜLLER, GEREON; JOHANNES ENGLISCH; and ANDREAS OPITZ. 2022. Extraction from NP, frequency, and Minimalist Gradient Harmonic Grammar. *Linguistics* 60.1619–1662
MURPHY, ANDREW. 2017. Cumulativity in syntactic derivations. Ph.D dissertation, Universität Leipzig.
NISSIM, MALVINA; SHIPRA DINGARE; JEAN CARLETTA; and MARK STEEDMAN. 2004. An annotation scheme for information status in dialogue. *Proceedings of the Fourth Language Resources and Evaluation Conference (LREC)*. Lisbon: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2004/pdf/638.pdf.
REINHART, TANYA. 1982. *Pragmatics and linguistics: An analysis of sentence topic*. Bloomington, IN: Indiana University Linguistics Club.
REPP, SOPHIE. 2010. Defining 'contrast' as an information-structural notion in grammar. *Lingua* 120.1333–1345.
REPP, SOPHIE. 2016. Contrast: dissecting an elusive information-structural notion and its role in grammar. *The Oxford Handbook of Information Structure*, ed. by Caroline Féry and Shinichiro Ishihara. Oxford: Oxford University Press.
PAYNE, DORIS L. 1987. Information Structuring in Papago Narrative Discourse. *Language* 63.783–804.
PULTE, WILLIAM, and DURBIN FEELING. 1975. Outline of Cherokee grammar. *Cherokee-English dictionary*, ed. by Durbin Feeling, 235–355. Talequah: Cherokee Nation of Oklahoma.
PRINCE, ELLEN F. 1981. Toward a taxonomy of given-new information. *Radical Pragmatics*, ed. by Peter Cole, 223–256. New York: Academic Press.
SCANCAFELLI, JANINE. 1986. Pragmatic Roles in Cherokee Grammar. *Proceedings of the Twelfth Annual Meeting of the Berkeley Linguistics Society*, ed. by Vassiliki Nikiforidou, Mary VanClay, Mary Niepokuj, and Deborah Feder, 224–234. Berkeley: Berkeley Linguistics Society.
SZMRECSANYI, BENEDIKT; JASON GRAFMILLER; JOAN BRESNAN; ANETTE ROSENBACH; SALI TAGLIAMONTE; and SIMON TODD. 2017. Spoken syntax in a comparative perspective: the dative and genitive alternation in varieties of English. *Glossa: a journal of general linguistics* 2.86:1–27.
SMITH, BRIAN W., and JOE PATER. 2020. French schwa and gradient cumulativity. *Glossa: a journal of general linguistics* 5.24. 1-33.
TONHAUSER, JUDITH, and ERIKA COLIJN. 2010. Word order In Paraguayan Guaraní. *International Journal of American Linguistics* 76.255–288.
UCHIHARA, HIROTO. 2013. Tone and accent in Oklahoma Cherokee. Ph.D dissertation, The

University at Buffalo, State University of New York.
UCHIHARA, HIROTO. 2014. Cherokee noun incorporation revsited. *International Journal of American Linguistics* 80.5–38.
VERHOEVEN, ELISABETH. 2014. Thematic Asymmetries Do Matter ! A Corpus Study of German Word Order. *Journal of Germanic Linguistics* 27.45–104.
WILLIAMS, ROBERT S. 1996. Cherokee possession and the status of *-jeeli*. *Cherokee papers from UCLA*, ed. by Pamela Munro, 97–104. Los Angeles: University of California, Los Angeles.
ZHANG, SHIYUE; BENJAMIN FREY; and MOHIT BANSAL. 2020a. ChrEnTranslate : Cherokee-English machine translation demo with quality estimation and corrective feedback. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, 272–279. Associtaion for Computational Linguistics.
ZHANG, SHIYUE; BENJAMIN FREY; and MOHIT BANSAL. 2020b. ChrEn: Cherokee-English machine translation for endangered language revitalization. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 577–595.