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

Brian Hsu (University of North Carolina at Chapel Hill), Benjamin Frey (University of Nort Carolina at Asheville)

Abstract: Although Cherokee is known to show highly flexible clausal word order, the principles that predict speakers' preferences among possible orders are not extensively described. This article presents a new description of the grammatical properties that predict clausal word order in spoken Cherokee, based on a corpus study of word order variation. Our results show that the placement of nominal expressions relative to verbs, and the relative order of nominal expressions within a clause, are determined in a probabilistic way by the cumulative interaction of several factors: REFERENTIAL ACCESSIBILITY, CONTRAST, and THEMATIC ROLE. The findings suggest that thematic properties may have a greater word order role than generally assumed in languages with nonconfigurational and/or polysynthetic properties.[*]

[*] Brian Hsu is a specialist in formal linguistics working with the second author to enhance descriptions of Cherokee grammar, with an eye toward pedagogical applications. Ben Frey is a citizen of the Eastern Band of Cherokee Indians who has worked toward language reclamation in his community and in conjunction with Cherokee Nation and United Keetoowah Band citizens since 2003. He currently teaches the Cherokee language at the University of North Carolina at Asheville and is working to create more pedagogical materials for second language learners of Cherokee.

We are grateful to Jolie Hiers and Erin Humphreys for their assistance in annotation. For helpful discussion at various stages of the work, we thank Raúl Aranovich, David Mora-Marín, Elliott Moreton, Emily Morgan, Katya Pertsova, Caitlin Smith, Jennifer Smith, George Walkden, three anonymous reviewers, and two editors of *Language.* Portions of this work were supported by a UNC Institute for the Arts and Humanities Faculty Fellowship. We would like to acknowledge the tireless efforts of language advocates in the Eastern Band of Cherokee Indians, the Cherokee Nation, and the United Keetoowah Band of Cherokee Indians. From fluent speaking elders to adult and child immersion participants, to those diligently attending community language classes, this effort relies on the community. You are doing the work necessary to carry our language into the next generation. ᎢᏨᎦᎵᎡᎵᏤᏂ.

1. INTRODUCTION. This research report 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). Following Dryer 1997, we illustrate the general flexibility of clausal word order in the language with the possible orders of transitive verbs and agent arguments in examples 1a-b, and of transitive verbs and theme arguments in examples 1c-d. Both types of arguments can either precede or follow verbs, and each type of order is found in spontaneous speech (Pulte & Feeling 1975, Scancarelli 1986, 1987, Beghelli 1996, Montgomery-Anderson 2008, 2015, Akkuş 2018). Like many languages with rich verbal agreement, Cherokee does not require all thematic arguments to be realized as a distinct nominal expression. Examples 1a-b lack an overt nominal theme, while 1c-d lack an overt nominal agent.[1]

(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 paper is to better understand which grammatical properties determine Cherokee speakers’ preferences among possible orders, and how these factors interact with each other. All prior works on Cherokee grammar describe clausal word order as largely influenced by factors related to information structure and discourse context, such as the referential accessibility and contrastive value of constituents (Pulte & Feeling 1975, King 1975, Cook 1979, Scancarelli 1987, Montgomery-Anderson 2015). The extent to which ordering is determined by other properties is less certain. Pulte and Feeling (1975) and King (1975) posit a tendency for noun placement to be determined by grammatical function, such that subjects preferentially precede objects. Montgomery-Anderson (2015: 310) suggests a possible effect of animacy, known to condition clausal word order in some languages (Brody 1984).

In this paper, we present a quantitative analysis of word order variation in Cherokee, using an annotated corpus of spoken narratives. This allows us to evaluate prior descriptions in a novel way and to test the effects of factors not yet considered, yielding a newly comprehensive description of the factors that shape clausal word order in the language. Although the corpus methodology and statistical methods that we use are not novel in themselves, they have rarely been used in the investigation of this type of pattern (clausal word order in a polysynthetic language), or in documentation work on similar languages (Indigenous and/or endangered, and underresourced). Our study makes several descriptive, theoretical, and methodological contributions, which we preview below.

In terms of descriptive generalizations about Cherokee grammar, we confirm previous observations about the effects of referential accessibility and contrast on word order, but find that their effects are PROBABILISTIC, rather than categorical. Second, we show that the thematic roles of nominal constituents also contribute to word order preferences; thematic agents are likelier to occur early in the clause than theme arguments. Our results suggest that previous descriptions using the grammatical function terms subject and object can be stated purely in terms of thematic roles. Third, we find that these information-structural and thematic properties are stronger predictors of clausal word order than noun animacy and phonological length. We use a logistic regression model to show that the relevant information-structural and thematic properties are statistically significant, independent factors that interact CUMULATIVELY to influence word order probabilities.

Our findings have several implications for broader areas of research related to optionality in syntax. Broadly, our results add to evidence that word order in language can be determined by the interaction of multiple properties, whose preferences may conflict, and that the effect of individual factors can be probabilistic (Payne 1987, Bresnan et al. 2001, Manning 2003, Rosenbach 2005, Benor and Levy 2006, Szmrecsányi & Hinrichs 2008, Bresnan & Ford 2010, Schoenmakers et al. 2021, see Grafmiller et al. 2018 for a recent overview). Our results also offer new evidence that information-structural and thematic properties can interact in flexible word order patterns (Bader & Häussler 2010, Verhoeven 2014, Ellsiepen & Bader 2018).

As a methodological contribution, our results suggest that annotated corpus methods can play a useful role in documentating typologically similar languages with variable word order, as they can uncover generalizations that are difficult to obtain from traditional elicitation methods (Tonhauser & Colijn 2010). This is particularly important, given that similar patterns of flexible word order are common in Indigenous languages of North America and Oceania (Dryer 2013). Similar corpus analyses can also facilitate the creation of pedagogically-oriented materials and grammars, crucial for revitalizing endangered languages like Cherokee. For example, it is easier to design instructional materials on flexible word order patterns if one can identify the most frequent patterns, and the properties that condition them (Frey 2020).

The paper is organized as follows. Section 2 provides an overview of clausal word order in Cherokee, and its prior descriptions. Section 3 describes the corpus and our annotation procedures for the tested word order predictors. Section 4 presents quantitative evidence that referential accessibility, contrast, and thematic roles independently contribute to word order preferences in the corpus. Section 5 discusses the effects of these factors on the relative order of nominal elements in longer sentences. Section 6 concludes the paper.

2. Language background and previous observations on word order. Cherokee is the only surviving member of the Southern branch of the Iroquoian language family (Julian 2010). Prior to forced removal in 1838 on the infamous Trail of Tears, the language had three main dialects: Overhill, Underhill, and Kituwah; also known as the Middle dialect. By the early 1900s, the Underhill dialect had fallen into dormancy while the Overhill and Middle dialects persisted (Mithun 1999:418–419, Montgomery-Anderson 2008:304). The language's vitality was severely impacted by the Federal Boarding School period, wherein Indigenous children from all over North America were taken away from their communities and placed in residential schools meant to "civilize" them. This practice was based in large part on the philosophy of Col. Richard H. Pratt, articulated as "kill the Indian, save the man". Children were beaten for speaking their languages instead of English and were subject to having their mouths washed out with soap (Duncan 1998).

Today, the combined impact of the boarding school period and a growing tide of increased interconnection with non-Cherokees in social, institutional, and economic domains (Frey 2013) has rendered the language endangered. In 2019, all three governments of Cherokee people passed a joint resolution to declare the language in a state of emergency (McKie 2019). Today there are approximately 2,000 first-language speakers remaining, the majority of whom are over the age of 65. The language is not being widely passed on in the home, but immersion schools exist in Oklahoma and North Carolina to educate K-12 students through the medium of Cherokee. Adult immersion programs and more conventional language classes augment these efforts, but generally lack the robust selection of pedagogical materials available to more commonly-taught languages.

The Cherokee language shows many morphosyntactic properties that are common among polysynthetic languages (for an overview, see chapters in Fortescue et al. 2017). In addition to the aforementioned flexibility in word order, all verbs contain a pronominal agreement prefix that inflects for properties of at least one thematic argument, including distinctions of person, number, and clusivity (Scancarelli 1987, Montgomery-Anderson 2015). It is relevant to note that Cherokee does not have productive noun incorporation, though it likely did at a historic stage (Uchihara 2014). We thus expect sentences in Cherokee to contain more nonpronominal nominal expressions overall than equivalent sentences in languages with productive noun incorporation, including its Northern Iroquoian relatives.

All previous descriptions of Cherokee agree that word order is conditioned to a large extent by factors related to information structure and/or discourse context. We present two types of patterns, based on similar examples from Scancarelli 1986, 1987. First, there are effects of what we call REFERENTIAL ACCESSIBILITY (Kuno 1972, Haviland & Clark 1974, Chafe 1976, 1994, Prince 1981). Nominal expressions that denote entities that are brand new to the discourse at hand tend to occur near the beginning of the clause, 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. ᎴᎦᏅ 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)

Second, phrases that express CONTRAST also tend to occur early in the clause. 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 & Vilkuna 1998, Neeleman et al. 2007, Molnár 2002, Aissen 2023).[2] This is shown in the excerpt in (3): the first sentence 3a 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.

(3) a. ᏫᎦᏥᎪᎥ ᎠᏂᏔᎵ ᎠᏂᏍᎦᏯ ᏩᎾᎢᏒᎢ.
wi-g-ajigo?-v a-ni-ta?li a-ni-sgaya w-a-n-a?isv?-i.
TR-1-see-EXPP 3-PL-two 3-PL-man tr-3-PL-walk-AG
‘I saw two men walking.’

b. ᏌᏊᏃ ᎩᎳᏊ ᎢᏴᏓ ᏥᏲᎵᏨ ...
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. ᎤᎾᎵᎪᏒᎢᏍᎩᏂ Ꮭ ᏱᏥᏲᎵᏤᎢ.
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; 1987) proposes that Cherokee follows the NEWSWORTHINESS PRINCIPLE (Mithun 1992, 1995), quoted below in (4), with our added roman numerals.[3]

(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)

Here, ‘syntactic function’ refers to the distinction between subjects and objects. The essential idea is that there are no ordering principles in these languages that refer specifically to subjects (i.e. the sole argument of an intransitive verb, the more prominent thematic argument of a transitive verb) or objects (the less prominent thematic argument of a transitive verb). Rather, the placement of a constituent is determined by a range of properties that contribute to its newsworthiness. Some of these properties are familiar information-structure classifications; (i) refers in part 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 & Vilkuna 1998, Neeleman et al. 2007).

It is likely that other types of properties contribute to the relative ordering of constituents, at least in Cherokee. This is apparent in sentences with multiple constituents that 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 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).

(5) Ꮎ ᎠᎾᎵᏍᏕᎵᏍᎩ ᏫᏚᏂᏍᎦᏤ 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-REPPnow dog
'The helpers called off the dogs/called the dogs back.' (Feeling et al. 2017: 83)

(6) ᏄᎵᏍᏔᏂᏙᎸ 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.' (Montgomery-Anderson 2008: 561)

(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)

Pulte and Feeling (1975) and (King 1975), noting similar patterns, ascribe them to a subject-object asymmetry in word order, such that subjects preferentially precede objects. In this paper, we propose that it can instead be ascribed to a difference in THEMATIC ROLES, such that thematic agent arguments have a propensity to precede theme arguments, all else being equal. In brief, this is because the ordering tendencies of individual nominal constituents do not depend on the number of constituents in the sentence (i.e. the placement of themes does not depend on whether it is the object in a transitive sentence, or the sole subject argument).

At this point, we note that it may not be possible to fully understand the language's word order principles in a traditional approach that relies on the qualitative examination of individual sentences, even when their context in a conversation is considered. First, because every nonverbal major constituent is specified for many types of properties, even beyond those that we have mentioned, it is difficult to isolate which property among them is responsible for an observed word order, and confounds can be hard to detect. Second, some word order principles are probabilistic; grammars can follow formal principles that tolerate occasional, but genuine exceptions (Bresnan et al. 2001, Manning 2003, Rosenbach 2005, Benor & Levy 2006, Szmrecsányi & Hinrichs 2008, Bresnan & Ford 2010, Schoenmakers et al. 2021). Finally, without a large quantity of examples, it is difficult to know how multiple word order factors interact, for example, whether some factors are more robust predictors of word order than others.

The rest of the paper presents a quantitative corpus analysis aimed at identifying the grammatical properties that best predict clausal word order in spoken Cherokee (see Tonhauser & Colijn 2010 for a similar approach to Guaraní). Crucially, this method allows us to evaluate the effects of a broader range of factors (including animacy, phonological length, and other thematic contrasts), which would otherwise be difficult to study. We also use a mixed-effects 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 view of how they interact.

Before presenting our corpus, we note that our descriptive model of Cherokee grammar shares some components with the newsworthiness principle, but differs in other ways. We maintain its key idea that the placement of a constituent is determined by a range of properties that it has, without reference to subject- or object-hood. However, we will not attempt to describe clausal word order as being determined by comparing the relative newsworthiness of each constituent. This is for a practical reason: it is difficult to annotate verbs and nominal constituents for the same types of properties. Because nouns typically refer to entities while verbs refer to predicates, verbs cannot be clearly characterized in terms of referential accessibility or contrast. And because many verbal predicates must cooccur with nominal arguments, it is difficult to independently evaluate their pertinence in a sentence. Rather, our model focuses on properties of nominal expressions only, 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. Our model does not annotate properties of verbs themselves, aside from phonological length. Finally, there are other types of factors that could in principle contribute to newsworthiness in that framework, for example topicality or surprisal, which we do not examine.

As a final caveat, our results should not be taken as an exhaustive description of factors that condition word order in Cherokee. Because the corpus consists of narratives spoken by individual speakers, rather than conversations, we cannot investigate information-structural properties beyond referential accessibility and contrast (discussed in 3.5). As the works in our corpus do not have associated audio files, and we are not aware of prior descriptions of intonational correlates of information structure in Cherokee, we have not investigated the role of phonological factors other than constituent length. These may ultimately be important, as word order choices in language can be influenced by segmental phonological restrictions (Szmrecsányi & Hinrichs 2008, Shih & Zuraw 2017) and lexical stress or tone (Shih et al. 2015), and information-structural properties may be correlated with particular prosodic structures (Bresnan & Mchombo 1987; Mithun 1996). We also do not investigate variation based on dialectal or sociolinguistic factors (Bresnan & Ford 2010, Szmrecsányi et al. 2017). Nonetheless, given the endangered and underresourced state of the language, we believe it is important to proceed from the currently available collections of transcribed spoken Cherokee, while inviting critical future work in these areas. Ultimately, we are confident that our study has identified the most salient properties that influence Cherokee word order, given their frequent occurrence and ease of identification in natural speech.

3. The Cherokee corpus and its annotated features

3.1. OVERVIEW OF THE CORPUS. Our annotated corpus consists of recorded narratives that have been published in Montgomery-Anderson 2008, 2015 and Feeling et al. 2017, the largest 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, assuming 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.[4] As listed below in Table 1, 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 and one historical narrative.[5] Given the small current inventory of transcribed and morphologically segmented texts in the language, we acknowledge that we cannot rule out possible effects of genre or formality on our results. Nonetheless, we proceed with the assumption that this sample contains a reasonable representation of the language's main clausal word order principles.

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. We exclude filler particles like *nogwu* 'now', nonthematic 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.[6]

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. Note that these counts include both argument and adjunct constituents. There are sentences in the source texts with only a verb, but they are not included in the corpus, given our interest 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 in Cherokee, but leave the question open for later work.[7]

Clauses with one major constituent	415
Clauses with two major constituents	139
Clauses with three major constituents	22
Clauses with four major constituents	75
Total number of clauses	581
Total number of major constituents	779

TABLE 2. Number of nonverbal major constituents per clause in the corpus

Each major constituent is tagged for the word order values PREVERBAL or POSTVERBAL, which are the main dependent variables in the quantitative analyses.[8]

The next subsections introduce the word order factors that we examine, and their annotation procedures. Factors that were ultimately not found to be significant predictors in the statistical analysis (length and animacy), and potential factors that were not examined, are discussed in Section 3.5.

3.2. REFERENTIAL ACCESSIBILITY. While there 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* in 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 annotations can be made 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 inferable from general knowledge or a relationship with an already-mentioned entity. 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 Ꭾ 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 above. *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 next example from ‘Water Beast’ contains two discourse-given expressions. At this point in the narrative, two men have 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 required, for given expressions to occur with demonstratives like *nasgina* ‘that’ and *nahna* ‘there’. Independent, nonprefixal pronouns referring to all persons (first-, second-, third-) are tagged as given, though the occurrence of independent first- and second-person pronouns is relatively rare in the corpus.

(10) **ᎾᏍᎩᎾ ᏩᎦ ᏧᎧᏅᏍᏕᏅ ᎾᎿ 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 before 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 inferable from a part-whole relation, subset relation, or superset relation with a given entity.[11] Example (11) occurs after the narrator has described stopping his car to pick someone up. The car’s door has not previously been mentioned, but its referent is accessible from its part-whole relation with the discourse-given car. Example (12) occurs in another narrative after the narrator has described a baseball game taking place. The existence of the pitcher is inferable 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-EXPP=CN door
‘He opened the door to get in.’ (Feeling et al. 2017: 35)

(12) ᏄᎵᏍᏈᏙᏏᏰᎥ 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-EXPP
‘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), the demonstrative pronoun and relative clause are both thematically related to the main verb *to happen*, but these expressions refer to events, rather than 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-EXPP 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), *ajilvye* ‘fire’ and *jigoya* ‘bug’ refer to generic entities, rather than specific instances of them. In example (16), *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 *asuhnidoh* ‘fisher’ in (17), and expressions like *Ann* in (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 3.4.

(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	ᏍᏚ	ᏚᏙᎥ	ᎠᎩᏥ	ᏥᎨᏒ.
	[**An**]	sgwu	d-u-do?-v	agi-ji	ji-ges-v.
	Ann	also	DST-3-be.named-EXPP	1.POS-mother	REL-be-EXPP

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

Having defined the key tags for referential accessibility in our corpus, we turn to their quantitative effects on word order preferences in Cherokee. First, expressions of all levels of referential accessibility tend to precede verbs: 398 (70%) of the 567 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 (80%) than accessible (72%), active given (58%) and inactive given information items (59%). There is no 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, these show the highest preference for preverbal placement (85%). We discuss statistical significance in greater detail in Section 4.1.

	Nonreferential	New	Accessible	Given-active	Given-inactive	Total
Preverbal	85	98	72	91	52	398
Postverbal	15	24	28	66	36	169

TABLE 3. Placement of major constituents relative to verbs, by referential accessibility value

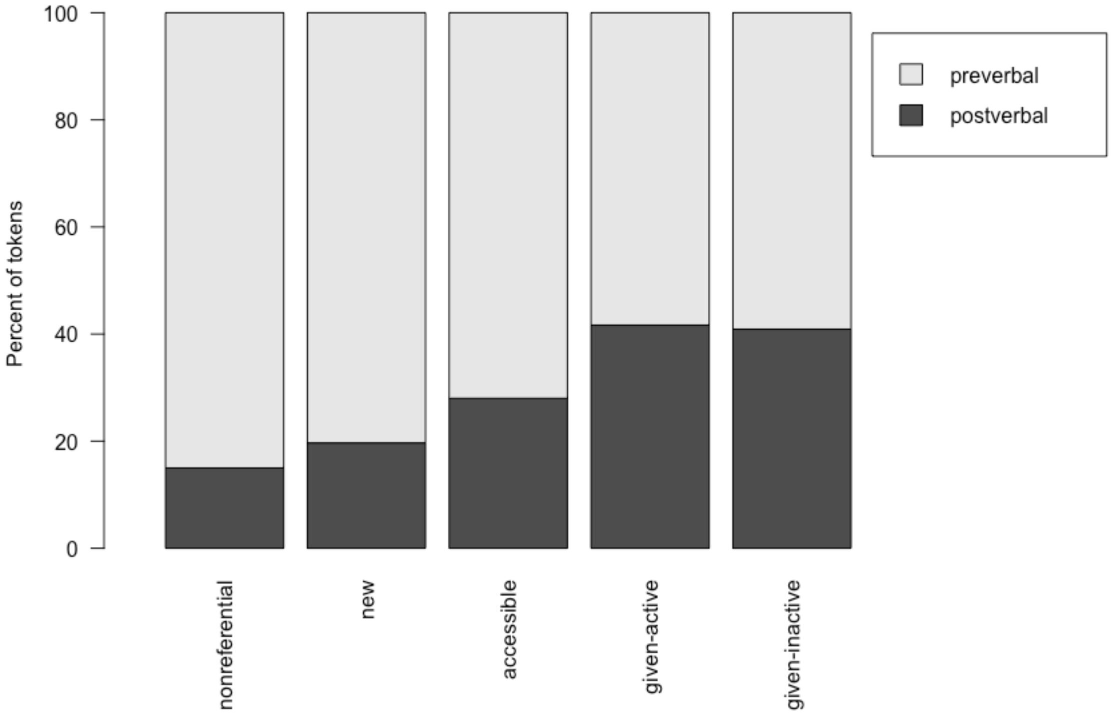

FIGURE 1. Percentage of preverbal vs. postverbal placement, by referential accessibility value

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 & Vilkuna 1998, Molnár 2002, Neeleman et al. 2007, Aissen 2023). Following these authors, we take contrast to be a distinct property from topic and focus (see Section 3.5). In order to maximize our confidence in correctly identifying contrasted items from transcribed speech alone, we employed the CONTRAST tag only for expressions that have an explicit alternative set, as defined by Repp (2010; 2016). Specifically, we use the tag only for expressions that are part of a set explicitly mentioned in a preceding part of the narrative, and are explicitly compared with an alternative (ex. *Two men saw the bull. One man ran, the other one stayed*). 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, as these are difficult to objectively identify from textual properties alone (any entity can plausibly be construed as 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 is a robust predictor of preverbal placement (18/20 = 90%). This effect is also statistically significant in the regression model presented in Section 4.

	Contrast	No contrast
Preverbal	18	534
Postverbal	2	180

TABLE 4. Placement of major constituents relative to verbs, by contrast value

3.4. THEMATIC ROLE. We examine the word order effects in the corpus of nine thematic roles. This includes argument 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, DYNAMIC LOCATION, and STATIC LOCATION.[12] While the vast majority of constituents with a thematic role tag are also tagged for referential accessibility and contrast, we do not use referential accessibility tags for time and abstract location expressions, as they do not generally participate in the same system of reference as items with argument thematic roles. We determine thematic role labels solely on semantic criteria (how the entity participates in the event or state denoted by the verb), independently of properties related to grammatical function (the number of arguments per verb, and their relative configuration). We comment below on specific cases where these criteria diverge.

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.[13]

(19) ᎡᎵᏏᏃ ᏃᏊ ᎪᎯ ᎢᏴ ᎤᏁᏨ.
[**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 ᏱᏄᎵᏍᏔᏂᏙᎳ 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)

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 (21) and (22). As shown in (23), the tag is also used for the sole argument of transitive predicates in the valency-reducing object focus structure; which suppresses a more prominent argument (typically an agent).

(21) Ꮓ ᎩᎳ ᎠᎹ ᏥᏙᏗᏍᎪ.
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)

(22) ᏴᏫ ᏧᎾᏍᏗᎢ ᏓᏂᏰᏎᎲᎢ.

[**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)

(23) ᎠᏥᎴ **ᏥᏍᏕ**.

Aji-l-e [**jisd**].

3.SG.O-kill-REPP rabbit

‘The rabbit was killed.’ (Feeling et al. 2017: 144)

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 … ’

ᏭᏅᎪᏨᏊ ᏦᏍᏓᏓᏅᏟ.

w-u-n-vgoj-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	ᎤᏲᎱᏎ	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)

As a caveat, there is likely more fine-grained variation involving themes that we do not examine. For instance, some predicates can assign combinations of agent-like and theme-like properties (Dowty 1991). In the absence of known diagnostics in Cherokee for additional distinctions, however, we rely primarily on whether the predicate denotes a change of state, change of location, or continuation of state, to label items as themes. We leave open the possibility that ordering preferences may ultimately be sensitive to finer degrees of themeness.

For predicates that express perceptual or psychological states, we used the tags STIMULUS for entities that cause a mental state or sensory perception, and EXPERIENCER for entities that experience the corresponding state. In (27), *jiistvvna* ‘crawdad’ is an experiencer argument, and *kaniitaʔv* ‘tail’ is a stimulus.

(27)	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 the ‘modified’ constituent in the attribution or identity relation. This tag is used for *agi-ji ji-ges-v* ‘my mother (honorific)’ in (28) and *sgi* ‘he’ in (29). On the flip side, the tag PREDICATE OBJECT (PRED-OBJ) is used for constituent that denotes ‘attribute’ being ascribed to the other argument. This tag used used for *An* ‘Ann’ in (28) and *asuhnidoh* ‘fisher’ in (29).

(28) **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)

(29) **ᏍᎩᏅ** **ᎠᏑᏂᏙ** ᎨᎲ.
[**Sgi]**=hnv [**asuhnidoh**] ge-hv.
That=CN fisher be-EXPP
‘And 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 these constituents to precede verbs, but with notable differences across thematic role values. First, we find that PRED-OBJ is the only type of item in the corpus that precedes verbs without exception. This is consistent with acceptability judgments reported 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, agents are more likely to show preverbal placement (78%) than themes (64%). This is consistent with crosslinguistic tendencies, as well as the description by Pulte and Feeling (1975). Section 4 presents further evidence that the distinction is statistically significant in the corpus.

	Pred-obj	Pred-sub	Agent	Experiencer	Theme	Stimulus	Total
preverbal	23	18	64	15	156	16	293
postverbal	0	3	18	5	88	10	124

TABLE 5. Placement of major constituents relative to verbs, by thematic role

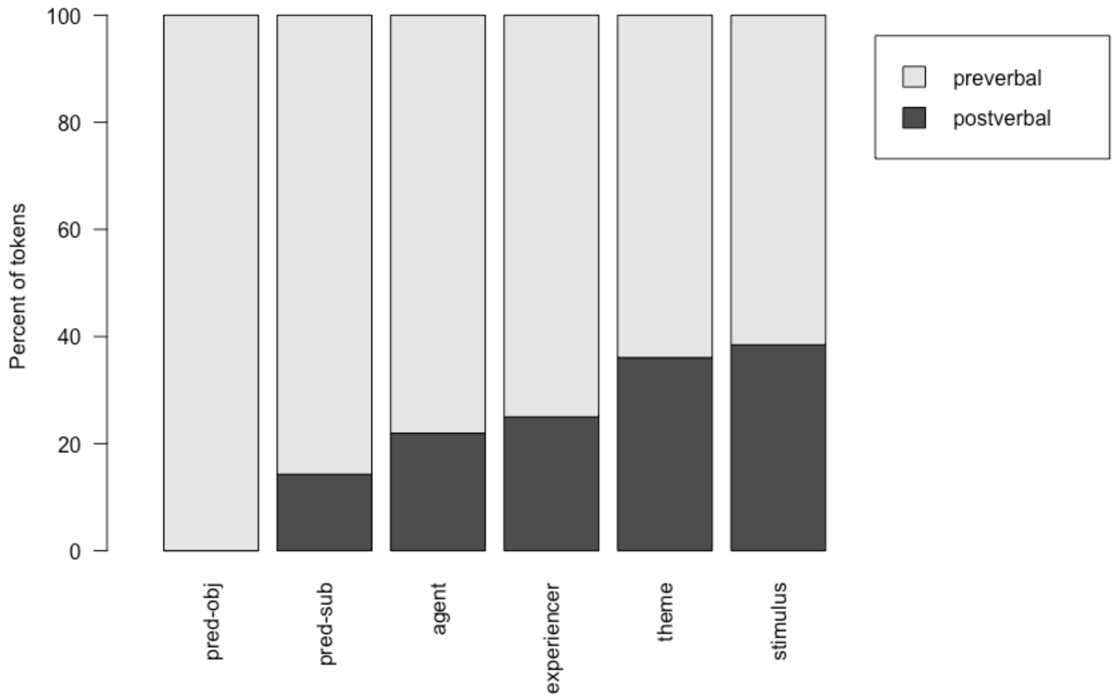

FIGURE 2. Percentage of preverbal vs. postverbal placement of arguments, by thematic role

Finally, we examine the patterning of several adjunct thematic roles: TIME, STATIC LOCATION, and DYNAMIC LOCATION. Time items include standalone adverbials like *hleg* '(for a) while' in (30). This tag is also used on adverbial clauses that denote the time of a main clause event, as in (31).

(30)	**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)

(31)	**ᎩᎶ**	**ᏳᏢᏣ**	ᏴᏫ	ᎾᎥ	ᏂᏚᎾᏓᎸ …
	[**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...'

ᏓᎾᏓᏩᏛᎯᏙᎲ.
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.

(32) ᏥᏳ ᎤᏍᏗ 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)

(33) ᎨᏅ ᎣᎬᏏ ᎾᎥ ᎡᎮ …
[**ge=hnv** **oaks-i** **na?v**] e-h-e …
there=CN Oaks-LOC near 3-live-REPP
‘Near the town of Oaks, there lived …’

ᎠᎦᏴᎵᎨ ᎤᎵᏍᎦᏍᏛ 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.

(34) ᏦᎩ ᎠᎦᏘ ᎠᏂᎦᏪᎯᎮ.
[**Jog** **akti**] a-ni-gawehih-e.
Upstream toward 3-PL-paddle-REPP
‘They were paddling upstream.’ (Feeling et al. 2017: 109)

(35) ᏗᏓᏁᎸ ᏩᏩᏗᏅᏒ.

[**Didanelv**] w-awadinvs-v.

Home TR-throw-EXP

‘I threw (it) towards home.’ (Feeling et al. 2017: 218)

As shown in the table below, static location and time expressions show a strong preference for preverbal placement (87% and 91.2%, respectively), greater than all argument thematic roles except for predicate objects (categorically preverbal). In Section 5, we show that static location and time expressions typically precede all other constituents in their 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
preverbal	64	41	125
postverbal	27	6	12

TABLE 6. Placement of major constituents with adjunct thematic roles

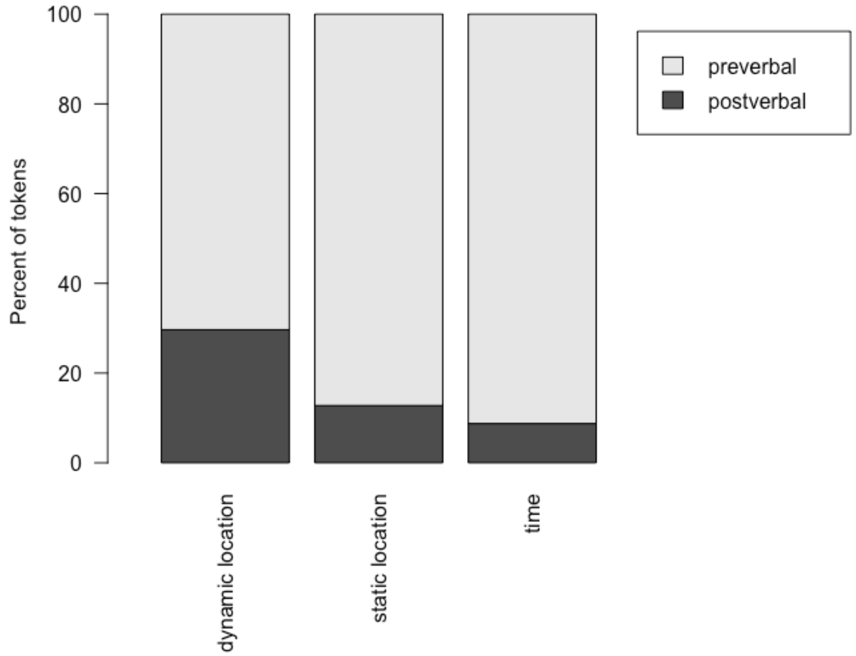

FIGURE 3. Percentage of preverbal vs. postverbal placement of adjuncts, by thematic role

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 information-structural factors that were not investigated.

ANIMACY. The animacy of nominal arguments has been identified as a conditioning factor in a range of morphosyntactic phenomena (see Toosarvandani 2023 for recent discussion), including word order restrictions (Brody 1984, Rosenbach 2005). Although animacy not known to have clear effects on clausal word order in Cherokee, it is a reasonable possibility to investigate, given the influence of animacy in several domains of agreement morphology (Scancarelli 1987, Montgomery-Anderson 2015). 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. Finally, animacy distinctions are marked on predicative adjectives that take set A agreement prefixes.

Animacy is also worth investigation as a potential confound with thematic roles, as agent arguments typically 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 expressions (as they do not generally have comparable referentiality properties). Expressions referring to humans, animals, and mythical beings are tagged as ANIMATE and all other entities as INANIMATE.

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 discourse-new constituents to be longer on average than discourse-given constituents, which are more likely to be realized as relatively short personal or demonstrative pronouns.

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 in 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 nonverbal constituents: the absolute length of the constituent, and its relative length (the length of the constituent, minus the length of the main verb in its clause). We cannot rule out possible effects of other measures of structural complexity (in terms of structural nodes, phrase types, etc.), but note that such measures would likely be quite correlated with segmental length.

INFORMATION-STRUCTURAL FACTORS NOT INVESTIGATED. We briefly mention other properties that are not investigated, due to the difficulty of identifying them in our corpus. First, we do not examine effects of aboutness topichood (Reinhart 1982) on word order, as it is generally difficult to identify topics from textual properties alone. However, we note that framesetting elements like time and static location expressions, sometimes considered a type of topic (Chafe 1976), are tagged and investigated (see Section 5.3). Another information-structural category that we do not directly examine is narrow focus, which we understand as arising in sentences where only one constituent represents nonpresupposed information. Some subtypes of focus overlap with properties that are tagged, but without an exact correspondence. For example, while discourse-new items sometimes pattern as foci, we cannot assume that all new-information constituents are focused. Generally, foci are difficult to find in monologue narratives, and are more reliably identifiable in dialogues with questions, or by using traditional elicitation methods (Aissen 2023). We reiterate that topic and focus are distinct notions from contrast; topics and foci need not be contrastive, and items that bear contrast can be either topics or foci (Vallduví & Vilkuna 1998, Molnár 2002, Neeleman et al. 2007, Aissen 2023).

4. PREDICTIVE WORD ORDER FACTORS IN CHEROKEE

4.1. REGRESSION MODEL. We used a mixed-effects 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 et al. 2013). Within the model, we used two word order values PREVERBAL, POSTVERBAL as the dependent variables, and considered thematic role, referential accessibility, conntrast, animacy, absolute length, and relative length (difference in segmental length between the constituent and the verb) as independent variables. The numeric length variables were centered and standardized to improve comparisons with the other, nonnumeric variables.

Expressions corresponding to certain thematic roles are omitted from the input to the regression model for the following reasons. STATIC LOCATION and TIME expressions are omitted, as they are the only uniformly adjunct-like roles, and TIME expressions cannot be annotated for a referential accessibility value (these items are examined separately in Section 5.4). Expressions with the PREDICATE OBJECT role are excluded, since this property categorically predicts preverbal placement; independent variables with categorical effects are not reliably captured in regression models due to high standard error estimates.

In order to quantify collinearity between factors, we first calculated Variance Inflation Factors (VIF) in a regression model that contains all of the tagged factors, using the vif( ) function of the car package (Fox & Weisberg 2019). The top row of Table 7 shows VIF scores for each factor in a model that contains all variables. Unsurprisingly, the two constituent length measures show severe collinearity (a VIF score greater than 5), as relative length is calculated on the basis of absolute length. As shown in the two rows below, models that remove either of the two length variables show generally lower VIF scores for all factors, all within moderate levels. We continue modeling using only absolute length as a tested variable, since it results in somewhat lower VIF scores overall.

	Thematic role	Animacy	Contrast	Referential accessibility	Length (absolute)	Length (relative)
VIF (Base model)	1.833718	1.763333	1.040403	1.263739	7.591688	7.667757
VIF (Base – absolute length	1.692396	1.715940	1.041117	1.254867	-	1.099354
VIF (Base – relative length)	1.687423	1.708414	1.040920	1.237969	1.087684	-

TABLE 7. Variance inflation factors for models with and without length factors

The model that we present below was obtained after a nested model comparison. The maximal model included thematic role, animacy, contrast, referential accessibility, absolute lenth, and all possible two-way interaction terms as fixed effects. Speaker identity was included as a random effect.[14] Models with different combinations of fixed effects were compared and selected using the likelihood ratio test, using the anova( ) function in R. The final model below is the one that contains the fewest variables, while yielding a significantly better fit compared to models with any of its variables removed (p = 0.05 or less). This model includes only referential accessibility, contrast, and thematic role as predictive factors. Neither absolute constituent length nor animacy, nor any interaction terms, are in the final model. We cannot conclude that these factors do not influence the ordering of major constituents Cherokee clauses, as they could show significant effects in a larger corpus. However, it suggests that any role that they play in determining clausal word order in Cherokee is less detectable.

We briefly explain how to interpret the results below. The reference categories (the intercept) consists of items tagged as NONREFERENTIAL, 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. A greater absolute value of the coefficient in the

Estimate column indicates a greater likelihood of postverbal placement for items with that tag, relative to the reference category. Predictors identified as statistically significant are shown with asterisks (*); the number of asterisks corresponds to *p*-value thresholds, such that more asterisks indicate a lower *p*-value (the probability that the observed numbers could arise by chance).

463 major constituents 9 speakers

Factor	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	4.6194	1.1636	3.970	7.19e-05	***
Referential accessibility factors					
accessible	-1.0649	0.4428	-2.405	0.01618	*
given-active	-1.5941	0.3973	-4.013	6.00e-05	***
given-inactive	-1.4755	0.4337	-3.402	0.00067	***
new	-0.2970	0.4407	-0.674	0.50035	
Thematic factors					
experiencer	-0.2801	0.6256	-0.448	0.65438	
location-dynamic	-0.6485	0.3830	-1.693	0.09043	.
pred-sub	-0.3238	0.7276	0.445	0.65629	
stimulus	-1.2224	0.5209	-2.347	0.01893	*
theme	-0.9736	0.3190	-3.052	0.00227	**
Contrast					
no contrast	-2.1633	1.0590	-2.043	0.04108 *	

Significance codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Random effects

Speaker (intercept) Variance: 0.0166 Standard Deviation: 0.1288

Likelihood and deviance

AIC	BIC	logLik	deviance	df.resid
551.7	601.4	-263.9	527.7	451

Table 8. Regression modeling estimates

The results confirm the general expectation that discourse new items are more likely to precede verbs than discourse-given items. A post-hoc Tukey's test finds significant differences in postverbal vs. preverbal placement between new information and active given items (*p*=0.0004), and between new information and inactive given items (*p*=0.0153). As noted previously, we did not have an expectation on the patterning of nonreferential items, but find that they pattern very similarly to new information. A pairwise Tukey's test finds a significant difference between nonreferential and given-active items (*p*=0.0006), and between nonreferential and given-inactive (*p*=0.0060). Turning to differences among thematic roles, a post-hoc analysis finds only the difference between agent and theme arguments to be significant (*p*=0.0276).

To verify whether sentences with more than one nonverbal major constituent follow distinct word-order restrictions, We repeated this procedure on a subset of the corpus that includes only sentences with one item per clause (roughly 70% of items in the full corpus). This yielded the same inventory of statistically significant predictors in the regression model, 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 fact that thematic distinctions, contrast, and referential accessibility are found to be significant word order factors by the model suggests that these properties interact cumulatively in determining the placement of individual constituents in the clause. That is, individual constituents with multiple factors that favor preverbal placement should show a greater total tendency for preverbal placement than constituents that have only one such property. The next subsection presents other measurements 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, 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 placement of theme nominals only, in each referential accessibility category. 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.

	Nonreferential	New	Accessible	Given-active	Given-inactive	Total
Preverbal	36	43	19	40	14	152
Postverbal	5	14	12	36	20	87

TABLE 9. Placement of theme arguments relative to verbs, by referential accessibility value

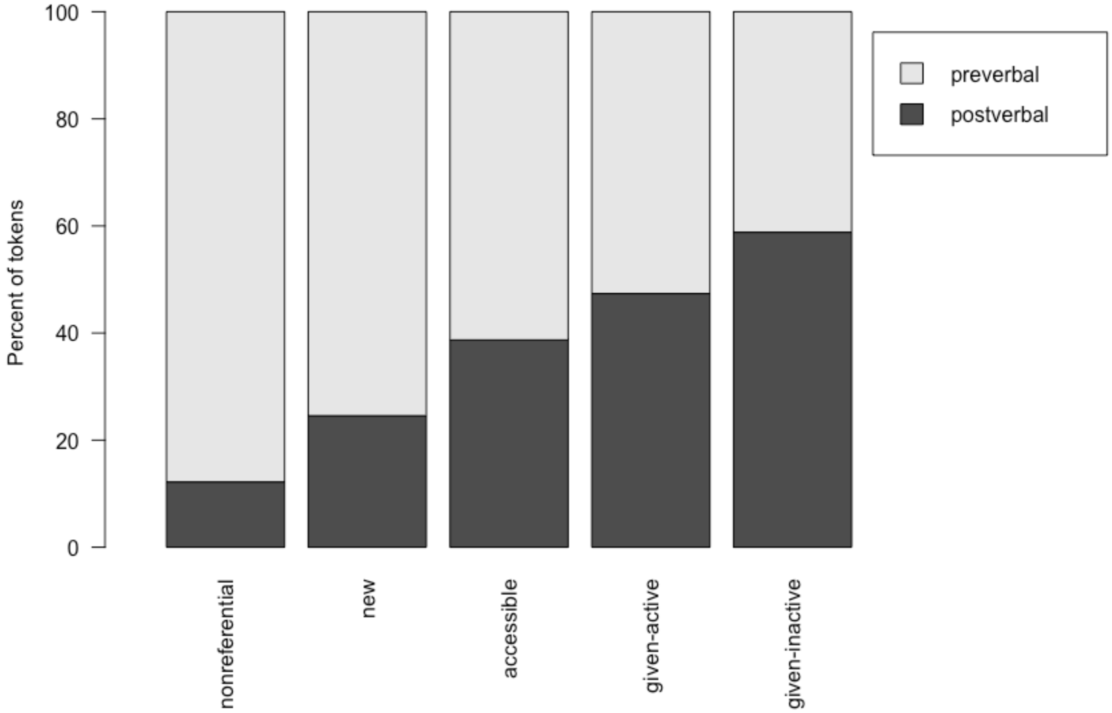

FIGURE 4. Percentage of preverbal vs. postverbal placement of theme arguments, by referential accessibility value

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

	Nonreferential	New	Accessible	Given-active	Given-inactive	Total
Preverbal	3	13	13	18	15	62
Postverbal	1	1	4	10	1	17

TABLE 10. Placement of agent arguments relative to verbs, by referential accessibility value

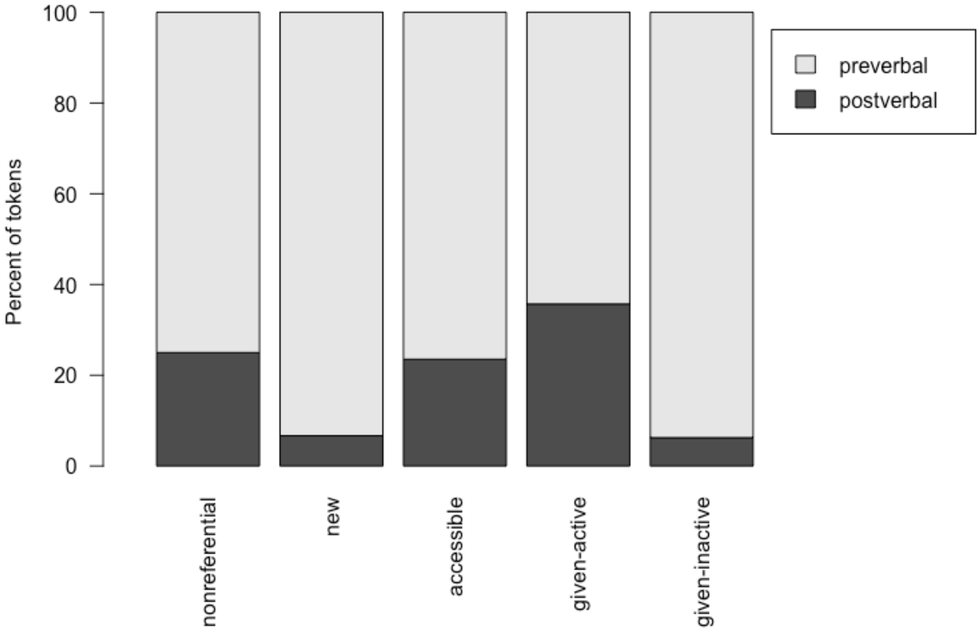

FIGURE 5. Percentage of preverbal vs. postverbal placement of agent arguments, by referential accessibility value

It is also informative to compare this pattern to the behavior of agents in Figure 4, and observe that theme arguments of all referential accessibility values are less likely to occur in a preverbal position than agents with the same referential accessibility. For example, accessible themes are 61% preverbal while accessible agents are 76% preverbal.

The cumulative interaction of referential accessibility and thematic role in predicting word order is also 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; this graph makes no distinction between active and inactive items. The columns compare agents versus themes. Agent arguments are uniformly more likely to precede verbs than theme arguments (compare left column to right column). Arguments that refer to new entities are more likely to precede verbs than arguments that are discourse-given (compare top row to bottom row). Items that have two properties favoring preverbal placement (agent and new information) are more likely to precede the verb than nouns with only one such property (given agents and new themes).

	Noun is agent	*Noun is theme*
Noun is new	**93%** preverbal (13/14)	**77%** preverbal (43/56)
Noun is given	**74%** preverbal (32/43)	**44%** preverbal (44/100)

TABLE 11. Percentage of preverbal placement, by select thematic role and referential accessibility values

Overall, the results support the idea that multiple types of grammatical properties contribute, probabilistically but systematically, to word order preferences in Cherokee. More broadly, they 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 in the extent to which they influence word order in various domains of structure (Payne 1987:801–802, Hale 1992, Baker 2006). Our result is also consistent with research on language production in which ordering preferences on nominal constituents arise from their relative CONCEPTUAL ACCESSIBILITY, as determined by a confluence of factors that include information structure and thematic structure (Bock & Warren 1985, Prat-Sala & Branigan 2000, Christianson & Ferreira 2005).[15]

5. WORD ORDER IN LONGER CLAUSES. Thus far, we have discussed properties that influence the order of certain constituents relative to verbs. This section examines clauses with more than one major constituent other than the verb, focusing on their relative order. While the portion of the corpus made of sentences of this type (170 sentences; 30% of the corpus) is too small for a robust quantitative analysis, we can nonetheless observe several trends 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 now from other 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 12. Distribution of nominal and adverbial expressions in clauses with more than one nonverbal major constituent

While it is easy to find sentences with multiple preverbal constituents, sentences with multiple postverbal items are relatively rare. Put another way, verbs in Cherokee are almost always 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 nonverbal constituent, where 24.7 percent (100/405) of tagged constituents follow the verb; 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 do not depend highly on the number of major constituents per clause.

It is noteworthy that in the three sentences in the corpus with two postverbal items, all postverbal constituents express given information. Sentence (36) has a theme and a goal; sentence (37) has a theme and a beneficiary or dynamic location; sentence (38) 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.

(36) ᏃᏭᏃ ᏣᎵᏟᎭᏭᏴ ᎠᎹ 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)

(37) ᏏᎸ ᏫᏗᏥᏅᏍ 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)

(38) ᏄᎾᏛᏁᎴ ᎠᏂᏐ ᏝᎦᏏ 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 ORDER 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 5.4). Our corpus contains 10 sentences of this type, which contain an agent and a theme (it is unusual for copular predicates to have a subject and object of the same information status, and we do not find any examples). While there is again some variability, there is an apparent preference for agents to precede themes.

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

TABLE 13. Relative orders of agents and themes with identical referential accessibility values

This dominant trend is illustrated with examples (39)-(41) below. In each sentence, all nominal constituents have the same referential accessibility status. The ordering tendency appears to hold independently of whether the verb occurs between both arguments as in (39) and (40), or whether both arguments precede the verb as in (41).

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

ᎡᏙᏓ ᏥᎨᏒ ᏃᏊ **ᏗᏡᎬ 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)

(40) 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)

(41) Agent > theme > V. All nominals are accessible.

ᏌᏇᏃ 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. EFFECTS OF REFERENTIAL ACCESSIBILITY ON RELATIVE ORDER OF NOMINAL EXPRESSIONS. Here, we examine clauses that contain multiple constituents with distinct referential accessibility tags. For this comparison, we abstract away from the thematic roles of each constituent, even though they are likely to independently affect how these constituents are ordered. Table 14 shows the number of attested orderings for each pair of referential accessibility features. While it is difficult to draw robust conclusions from the small number of examples, we observe trends that are consistent with our previous findings. First, new items very often precede given items, as consistent with observation that new items are more likely to precede verbs than given items. Second, the relative order of accessible and given items is highly flexible, as consistent with the 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 precedes nonreferential	Nonreferential precedes given
5	2
Nonreferential before accessible	Accessible before nonreferential
3	1

TABLE 14. Relative orders of nominal expressions with different referential accessibility tags

5.4. PLACEMENT OF LOCATION AND TIME EXPRESSIONS. We now examine the positions of time and location expressions, relative to other items in the clause. In brief, time expressions and static location expressions have a robust tendency to precede all other expressions in the clause. This is consistent with the crosslinguistic tendency for FRAME-SETTING items, which delimit the context in which an event ocurs, to occur in clause-initial positions and precede argument expressions (Speyer 2008, Wolfe 2015). In contrast, dynamic location expressions, which often have a stronger selectional relation with event-denoting verbs, occur later in the clause in a more argument-like distribution.

As shown in Table 15, a large majority of time expressions (80%) are the first item in their clause. They occur much less frequently 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 15. Position of time adverbials in sentences with more than one major constituent

In some cases, the position of time elements seems to be conditioned by scopal differences, as illustrated in sentence (42). 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'. In (43), 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 (44) appears to modify the full proposition, even though it occurs in a clause-medial position.

(42) ᏌᏊᏃ ᎢᏳᏩᎪᏗ ᎡᎵᏏ ᎡᏚᏚᏃ …
[**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 ...'

ᎤᏒ ᎢᎠᎾᎢᏎ ᏧᏁᏅᏒ ᎢᏗᏝ.
[**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: 43)

(43)	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)

(44)	ᎡᎵᏏᏃ	ᏃᏊ	ᎪᎯ	ᎢᏴ	ᎤᏁᏨ.
	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 (56%), as shown in Table 16. However, they are somewhat likelier than time adverbials to occur in a clause-medial or clause final position. It is relevant, 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, 67 percent (18/27) of static location adverbials precede all nontime 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 16. 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 to precede static location when both items are present: *time > static location > all other expressions*. This particular ordering is exemplified in (45).

(45) **ᎥᎾᏫᏛᏍᏚ** **Ꮎi** 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)

In contrast, dynamic location expressions show greater variability in placement, and are much less frequently clause-initial (30% of examples). In this way, their distribution more closely resembles that of thematic arguments (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 17. Position of dynamic location adverbials in sentences with more than one major constituent

6. CONCLUSION. In this paper, we used an annotated corpus analysis to identify several novel generalizations about clausal word order in Cherokee. We have shown that word order in the language is principally determined by several information-structural and thematic properties. These factors probabilistically influence both the order of major constituents relative to verbs, and ordering relations among nominal and adverbial phrases.

We believe that these generalizations drawn from naturalistic corpus data should be used as a basis for new pedagogical materials on word order in Cherokee. Specifically, language-learners may benefit from understanding the most common types of word orders, and the most likely associations between predictive grammatical properties and positions in the clause, all while acknowledging the flexibility that the grammar allows.

We briefly mention some broader theoretical implications of our findings. First, they suggest that word order preferences in languages with nonconfigurational 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. Similarly, it may not always be productive to classify thematic structure and information structure as belonging to fundamentally different modules, such as 'grammar' vs. 'discourse.' Finally, 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 & Johnson 2003, Smith & 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 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. There could be other predictive word order factors, including prosodic or segmental phonological restrictions, topic and focus, nominal properties like quantification, and clause-level properties related to negation, mood, tense, or aspect. These types of investigations would require a more extensive or differently-structured corpus, traditional elicitation methods, or a combination of approaches. We believe, however, that the generalizations that we have identified here provide a new groundwork for such projects.

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 invaluable 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. 2023. Documenting topic and focus. *Key topics in language documentation and description. Language Documentation & Conservation Special Publication no. 26*, ed. by Peter Jenks and Lev Michael, 11−57. Honolulu: University of Hawai'i Press.

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, 1−16. Vancouver: University of British Columbia.

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.

BAKER, MARK C. 2006. On zero agreement and polysynthesis. *Arguments and agreement*, ed. by Peter Ackema, Patrick Brandt, Maaike Schoorlemmer, and Fred Weermann, 74:289–320. 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.

BENOR, SARAH BUNIN; and ROGER LEVY. 2006. The chicken or the egg? A probabilistic analysis of english binomials. *Language* 82.233–278.

BOCK, J. KATHRYN, and RICHARD K. WARREN. 1985. Conceptual accessibility and syntactic structure in sentence formulation. *Cognition* 21.47–67.

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.

BRESNAN, JOAN; SHIPRA DINGARE; and CHRISTOPHER D. MANNING. 2001. Soft constraints mirror hard constraints: voice and person in English and Lummi. *Proceedings of the LFG '01 Conference*, ed. by Miriam Butt and Tracy Holloway King. Stanford: CSLI.

BRESNAN, JOAN, and MARILYN FORD. 2010. Predicting syntax: processing dative constructions in American and Australian varieties of English. *Language* 86.168–213.

BRESNAN, JOAN, and SAM A. MCHOMBO. 1987. Topic, pronoun, and agreement in Chicheŵa. *Language* 63.741–782.

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.

CHAFE, WALLACE. 1994. *Discourse, consciousness, and time*. Chicago: University of Chicago Press.

CHRISTIANSON, KIEL, and FERNANDA FERREIRA. 2005. Conceptual accessibility and sentence production in a free word order language (Odawa). *Cognition* 98.105–135.

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.

DOWTY, DAVID. 1991. Thematic proto-roles and argument selection. *Language* 67.547–619.

DRYER, MATTHEW S. 1997. On the six-way word order typology. *Studies in Language* 21.69–103.

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.

DUNCAN, BARBARA R. 1998. Living stories of the Cherokee. Chapel Hill, NC: University of North Carolina Press.

ELLSIEPEN, EMILIA, and MARKUS BADER. 2018. Constraints on argument linearization in German. *Glossa: a journal of general linguistics* 3.6.1–36.

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.

FOX, JOHN, and SANFORD WEISBERG. 2019. An R companion to applied regression, third edition. Thousand Oaks, CA: Sage.

FREY, BENJAMIN. 2013. Toward a general theory of language shift: A case study in Wisconsin German and North Carolina Cherokee. Ph.D dissertation, University of Wisconsin-Madison.

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.

GRAFMILLER, JASON; BENEDIKT SZMRECSÁNYI; and MELANIE RÖTHLISBERGER. 2018. General introduction: a comparative perspective on probabilistic variation in grammar. *Glossa: a journal of general linguistics* 3.94. 1–10.

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. 1992. Basic word order in two "free word order" languages. In Payne 1992, 63–82.

HAVILAND, SUSAN, and HERBERT CLARK. 1974. What's new? Aquiring new information as a process in comprehension. *Journal of Verbal Learning and Verbal Behavior* 13.512–521.

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.

JULIAN, CHARLES. 2010. A history of the Iroquoian languages. Ph.D dissertation, University of Manitoba.

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

KUNO, SUSUMU. 1972. Functional sentence perspective: a case study from Japanese and English. *Linguistic Inquiry* 3.269–320.

MANNING, CHRISTOPHER D. 2003. Probabilistic syntax. *Probabilistic linguistics*, ed. by Rens Bod, Jennifer Hay, and Stefanie Jannedy, 289–341. Cambridge, MA: MIT Press.

MCKIE, SCOTT. 2019. Tri-Council declares state of emergency for Cherokee language. *Cherokee One Feather* article, 27 June 2019. Online: https://theonefeather.com/2019/06/27/tri-council-declares-state-of-emergency-for-cherokee-language/.

MITHUN, MARIANNE. 1992. Is basic word order universal? In Payne 1992, 15–62.

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. 1996. Prosodic cues to accessibility. *Reference and referent accessibility*, ed. by Thorstein Fretheim and Jeanette K. Gundel, 223–234. Amsterdam/Philadelphia: John Benjamins. doi:10.1075/pbns.38.13mit.

MITHUN, MARIANNE. 2017. The Iroquoian language family. *The Cambridge handbook of linguistic typology*, ed. by Alexandra Y. Aikhenvald, 747–781. Cambridge University Press.

MOLNÁR, VALÉRIA. 2002. Contrast from a contrastive perspective. *Information structure in a cross-linguistic perspective*, ed. by Hilde Hallelgård, Stig Johansson, Bergljot Behrens, and Cathrine Fabricius-Hansen, 147-161. Amsterdam/New York: Rodopi.

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.

NEELEMAN, AD; ELENA TITOV; HANS VAN DE KOOT; and REIKO VERMEULEN. 2007. A syntactic typology of topic, focus and contrast. *Alternatives to Cartography*, ed. by Jeroen van Craenenbroeck, 15–52. Berlin: Walter de Gruyter.

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.

PAYNE, DORIS L. 1992. *Pragmatics of word order flexibility*. Amsterdam: John Benjamins.

PULTE, WILLIAM, and DURBIN FEELING. 1975. *Outline of Cherokee grammar with Cherokee-English dictionary*. Tahlequah: Cherokee Nation of Oklahoma.

PRAT-SALA, MERCÈ, and HOLLY P. BRANIGAN. 2000. Discourse constraints on syntactic processing in language production: A cross-linguistic study in English and Spanish. *Journal of Memory and Language* 42.168–182.

PRINCE, ELLEN F. 1981. Toward a taxonomy of given-new information. *Radical pragmatics*, ed. by Peter Cole, 223–256. New York: Academic Press.

PRINCE, ELLEN F. 1992. The ZPG letter: subjects, definiteness, and information status. *Discourse description: Diverse analyses of a fund-raising text*, ed. by William Mann and Sandra Thompson, 295–325. Philadelphia: John Benjamins.

ROSENBACH, ANETTE. 2005. Animacy versus weight as determinants of grammatical variation in English. *Language* 81.613–644.

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.

SCANCARELLI, JANINE. 1987. Grammatical relations and verb agreement in Cherokee. Ph.D dissertation, University of California, Los Angeles.

SCHOENMAKERS, GERT-JAN; MARJOLEIN POORTVLIET; and JEANNETTE SCHAEFFER. 2021. Topicality and anaphoricity in Dutch scrambling. *Natural Language & Linguistic Theory* 40:541–571.

SHIH, STEPHANIE; JASON GRAFMILLER; RICHARD FUTRELL; and JOAN BRESNAN. 2015. Rhythm's role in genitive construction choice in spoken English. *Rhythm in cognition and grammar: a Germanic perspective,* ed. by Ralf Vogel and Ruben Vijver, 207–234. Berlin: De Gruyter.

SHIH, STEPHANIE S, and KIE ZURAW. 2017. Phonological conditions on variable adjective and noun word order in Tagalog. *Language* 93.e317–e352.

SORACE, ANTONELLA. 2000. Gradients in auxiliary selection with intransitive verbs. *Language* 76.859–890.

SPEYER, AUGUSTIN. 2008. German vorfeld-filling as constraint interaction. *Constraints in discourse*, ed. by Anton Benz and Peter Kühnlein, 267–290. Amsterdam: John Benjamins.

SZMRECSÁNYI, 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.

SZMRECSÁNYI, BENEDIKT, and LARS HINRICHS. 2008. Probabilistic determinants of genitive variation in spoken and written English: A multivariate comparison across time, space, and genres. *The dynamics of linguistic variation: Corpus evidence on English past and present*, ed. by Tertu Nevalainen, Irma Taavtsainen, Paivi Pahta, and Minna Korhonen, 291–309.

Amsterdam: John Benjamins.

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.

TOOSARVANDANI, MAZIAR. 2023. The interpretation and grammatical representation of animacy. *Language* 99.760–808.

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.

VALLDUVÍ, ERIC, and MARIA VILKUNA. 1998. On rheme and kontrast. *The Limits of Syntax, Syntax and Semantics 29*, ed. by Peter W. Culicover and Louise McNally, 79–108. New York: Academic Press.

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.

WOLFE, SAM. 2015. The nature of Old Spanish verb second reconsidered. *Lingua* 164:132–155.

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*, ed. by Heng Ji, Jong C. Park, and Rui Xia, 272–279. Online: Association 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 (EMNLP)*, ed. by Bonnie Webber, Trevor Cohn, Yulan He, Yang Liu, 577–595. Online: Association for Computational Linguistics.

[1] We use the following glosses in the Cherokee examples (see Pulte and 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, ITR = iterative, LAT = lateral movement, LOC = locative, NEG = negation, NONF = nonfinite, O = object focus, 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.

[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] Mithun (1995; 2017) suggests that all Iroquoian languages fall into the class of pragmatically based languages, and 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 highly similar in their word order principles.

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

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

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

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

[8] The corpus contains eight nominal expressions that are apparently *discontinuous*: words interpreted as part of a single expression occur on separate sides of the verb. As exemplified below, the head noun in all such cases is postverbal, while its preverbal associate is often a numeral or quantifier (which typically precede head nouns in nondiscontinuous expressions). The full expression in most often a theme argument (five of eight examples). While Williams (1996) notes that possessors and possessum nouns can occur in a similar configuration, 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). We leave a detailed investigation of these structures, and how they relate to our main findings, for future work.

(i) **ᏦᏊ** ᏚᎾᏗᏅᏎ **ᏧᎾᏓᎵ**.
jo=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)

(ii) **ᏥᏇᏯ** ᏓᏛᏗ **ᎠᏂᏇ**.
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)

[9] Because the source texts are not transcribed with punctuation, and the placement of clause boundaries is sometimes ambiguous, we rely on the authors’ English translations to determine sentence boundaries, in order to distinguish between active versus inactive items.

[10] Although the term ‘accessible’ is from Chafe (1994), the definition here more closely resembles the concept of ‘inferable’ in Prince (1981; 1992). In brief, Chafe uses the term more broadly to the mental retrievability of a referent, potentially influenced by factors beyond logical inferences that the listener can make.

[11] Our annotation scheme does not distinguish subclasses of accessible items, due to the relative rarity of expressions that denote entities known through general knowledge or a superset relation with given entities. Annotators are also less likely to agree on these subclasses (Nissim et al. 2004).

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

[13] Our quantitative analyses 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 tendency in these contexts for the agent argument to occur clause-finally, resembling quotative inversion structures in languages like English. An example is shown here in (iii).

(iii)						
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)

[14] Random slope terms based on speaker identity were omitted; no model with random slopes converged, likely due to the low numbers of tagged items per speaker.

[15] One caveat is that in these approaches, discourse-given items are classified as having greater conceptual accessibility than discourse-new items, while thematic agents have a greater conceptual accessibility than themes. In Cherokee, however, both agents and discourse-new items tend to occur early in the clause. One possible explanation for this difference is that the pressure to place new information early (grounded in communicative pressure) can in some languages overcome preferences related to ease of retrieval (grounded in processing pressure).