

THE DISTRIBUTION OF RELATIVE CLAUSES IN KAZAKH CONVERSATIONS

by

Akyl Akanov

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Arts in

Eurasian Studies

at

NAZARBAYEV UNIVERSITY -
SCHOOL OF SCIENCES AND HUMANITIES

2024

THESIS APPROVAL FORM

NAZARBAYEV UNIVERSITY
SCHOOL OF HUMANITIES AND SOCIAL SCIENCES

THE DISTRIBUTION OF RELATIVE CLAUSES IN KAZAKH CONVERSATIONS

BY

Akyl Akanov

NU Student Number: 201765957

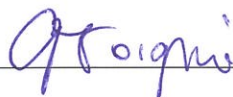
APPROVED

BY

Dr. GIORGIA TROIANI, POSTDOCTORAL SCHOLAR

ON

The 29 of April, 2024



Signature of Principal Thesis Adviser

In Agreement with Thesis Advisory Committee
Second Adviser: Dr. Andrey Filchenko, Professor/Vice Dean for Academic Affairs
External Reader: Dr. Sandra Auderset, Postdoctoral Researcher

Abstract

In conversation, speakers need to track the referents they introduce, establishing their identity and thereby building common ground with the hearer. This communicative need can be achieved through relative clauses (RCs), among other linguistic means. While most studies have traditionally focused on the formal syntactic properties of RCs, either in a language-specific or cross-linguistic perspective, other studies in usage-based model of language have focused on the variation of RCs in naturally occurring discourse, including conversations. These studies suggest that the distribution of RCs in naturally occurring discourse is affected by a number of linguistic, cognitive, and discourse-related factors such as word order, information flow, markedness, humanness status of the referent in the head noun phrase, and functions of RCs, among others.

Under the framework of Discourse and Grammar, I focus on relative clause constructions in the Kazakh language and explore the skewed distributional patterns of RCs as influenced by a number of linguistic, cognitive, and discourse-related factors that govern communication. Through the analysis of approximately 300 minutes of naturally occurring informal conversations from the Multimedia Corpus of Modern Spoken Kazakh, I have found that the distribution of relative clauses in Kazakh conversations exhibits statistically significant skews. I argue that these skewed patterns are best predicted by the interplay of the semantic factor of Humanness, the cognitive factor of Information Status as well as the grammatical factor of Function of the RC.

The findings support the view that discourse is always driven by the communicative goals of interactants, and that, consequently, grammar is a crystallization of such recurrent linguistic behavior. As such, this work corroborates the importance of studying linguistic structures in their ‘social habitat’ — everyday social interactions. Most importantly, this study contributes to a holistic representation of Kazakh, whose grammatical descriptions, as of now,

are mostly based on introspection, written language, and idealized language use, with only few works analyzing spoken data from fieldwork interviews.

Table of Contents

<u>1.</u>	Introduction.....	1
<u>1.1.</u>	Framework	1
<u>1.2.</u>	The role of conversational data	2
<u>1.3.</u>	The Kazakh language	4
<u>1.4.</u>	Relative Clauses	5
<u>2.</u>	Data & Methods	10
<u>2.1.</u>	Data and Variables	10
<u>2.2.</u>	Methodology	17
<u>3.</u>	Results & Discussion.....	21
<u>3.1.</u>	Frequency Distributions of FUNCTION OF HEAD NP IN MATRIX CLAUSE & FUNCTION OF HEAD NP IN RC	21
<u>3.2.</u>	A multivariate analysis of HEAD-RC COMBINATIONS and POSITION	28
<u>3.3.</u>	Pairwise comparisons between variables	46
<u>3.4.</u>	A note on the use of the Russian relative pronoun <i>kotor-</i> in Kazakh RCs.....	63
<u>3.5.</u>	Summary of results	63
<u>4.</u>	Conclusion	64
<u>5.</u>	References.....	67
<u>6.</u>	Appendices.....	73

1. Introduction

1.1. Framework

My work focuses on relative clauses in the Kazakh language and explores the skewed distributional patterns of relative clause constructions as influenced by a number of linguistic and cognitive factors that constrain communication. This topic is analyzed through the framework of Discourse and Grammar, a functional approach to language which relates grammatical structure to discourse structure based on the assumption that language form is motivated by cognitive and communicative demands necessitated by the need to produce coherent discourse (Ochs, Schegloff & Thompson, 1996: 10; Couper-Kuhlen & Selting 2017: 4). This means that discourse is not a random collection of utterances and is always driven by the communicative goals of interactants, and that, consequently, “grammar codes (best) what speakers do most [in discourse]” (Du Bois 1987: 811). It also argues that there is regular feedback between grammar and discourse, with discourse making a selective use of grammar, creating specific discourse patterns, and grammar, in turn, being created by the grammaticization of these discourse patterns (Ariel 2009). Narrative and written data were primary forms of discourse analyzed under this framework, but with the growing interest in real-time data, scholars began focusing their attention on social interactions, relating grammar to the sequential organization of talk, thus giving rise to an overarching framework of Interaction and Grammar (Ochs, Schegloff & Thompson 1996: 11). As such, Interaction and Grammar seeks to understand the relationship between grammar and one particular form of discourse — everyday social interactions. My research is grounded in the analysis of language use in social interactions and, therefore, falls within the scope of Interaction and Grammar research as well. In this framework, language emerges as an entity co-constructed by speakers in social interactions, implying that linguistic structures are not only resources for

carrying out interactions but also a product of these very interactions (Ochs, Schegloff & Thompson 1996: 38). Scholars in this field have looked at interactions where speakers were performing social actions such as requests, compliments, complaints etc. in order to understand how the sequential implementation of these actions shapes the grammatical forms being deployed; likewise, scholars have also focused on particular grammatical forms to investigate their interactional role in conversations (Taleghani-Nikazm 2006: 7). Since social interactions are a primary form of discourse analyzed under this framework, linguists have relied on the recordings of naturally occurring conversations for analysis (Ford & Thompson 1996: 136). Here the term *naturally occurring* refers to speech events that take place for the speakers' social goals and are consequential for their lives (Troiani, Du Bois, & Filchenko, in press). Thus, I chose to focus on conversational data over other forms of discourse to capture the use of relative clauses as motivated by circumstances arising in naturally occurring interactions.

1.2. The role of conversational data

Since Discourse and Grammar assumes that grammar is a crystallization of frequent linguistic behavior, it is important to choose the right kind of data for the investigation of grammar in use. Linguistic data comes in written and spoken forms, each of which is comprised of a variety of genres. In this context, the term *genre* can best be understood in Fairclough's (1993: 138) terms: genre is "the use of language associated with a particular social activity." Conversations serve as a genre of discourse most suitable for such Discourse and Grammar research for the following reasons. First, conversation is a universal form of discourse, found in all cultures at every stage of human history (Schegloff 2015; Chafe 1994). Thus, it is a genre of spoken discourse available to all speakers. Second, everyday conversational exchanges are a medium through which children acquire their first language (Clark & Casillas 2015). This aspect is crucial because it allows researchers to observe how

linguistic structures are acquired, reinforced, and refined through repeated exposure and interaction. Third, the spontaneous and interactional nature of conversations lends them to the investigation of on-line processes involved in the production and management of conversations (Du Bois 2003: 52-53). This allows researchers to examine how speakers rapidly navigate linguistic choices, negotiate meaning, and adapt their language use in real-time, shedding light on the dynamic nature of grammar in use.

The focus on conversational data partially addresses the scarcity of scholarship on grammar in use within naturally occurring spoken discourse. It also confronts the issue that was common to linguists throughout much of the twentieth century called ‘written language bias’ (Linell 2004); it refers to the idea that descriptive concepts and categories that have been developed in linguistics were primarily informed by the analysis of written language (Linell 2004; Couper-Kuhlen & Setling 2017). This means that scholars have long relied on written-language-biased concepts to analyze and theorize phenomena in spoken discourse as well. However, written language cannot account for the phenomena in spoken discourse, including naturally occurring conversations. For instance, the data from Mohawk (Iroquoian) comprising extended bodies of everyday speech reveals unique grammatical patterns, challenging simplistic descriptions found in standard written pedagogical materials (Mithun 2015). For example, the particle *ne* in Mohawk functionally does not align neatly with the English *the* as it would initially seem when analyzing this particle in written Mohawk. While superficially similar, *ne* functions more as a reference to previously mentioned referents rather than indicating general identifiability. Mithun (2015) argues that this nuanced distinction becomes apparent only through the analysis of Mohawk spoken discourse. Linguists who specialize in documenting spoken languages with no written tradition also have to deal with this bias. For example, it is common for unwritten languages such as Yélí Dnye (East Papuan) to lack metalinguistic terms for certain social actions such as promises,

threats, offers and etc., which are usually present in languages with a very long written tradition such as English (Levinson 2012: 124). This means that even such metalinguistic concepts should be applied to unwritten languages with great attention and only on the basis of conversational data from these varieties.

1.3. The Kazakh language

Kazakh belongs to the South Kipchak group of the Kipchak branch of the Common Turkic subfamily of the Turkic language family. It is head-final and has an SOV word order. According to the Bureau of National Statistics of the Agency for Strategic Planning and Reforms of the Republic of Kazakhstan (2021), as of 2021, 80.1% of approximately 17 million census respondents reported fluency in Kazakh, of which 61.54% indicated that they used Kazakh in everyday life. Despite this considerable number of speakers and the language's vital status, Kazakh has received relatively limited attention in modern linguistic research, especially considering the abundance of available resources such as the available corpora: Almaty Corpus of Kazakh (Bazarbayeva et al. 2023), Kazakh Language Corpus (Makhambetov et al. 2013), the National Corpus of Kazakh Language (Zhubanov 2009), Kazakh Speech Corpus 2 (Mussakhojayeva et al. 2022), and the Multimedia Corpus of Modern Spoken Kazakh (Filchenko, Troiani, Du Bois & Sarseke et al. 2023).

The Central Asian region has seen intricate dynamics of linguistic contact due to trade networks and the Russian imperial colonization of Central Asian territories (Manz 2018). Traders along the Silk Road that passed through modern-day Kazakhstan were proficient in a variety of languages such as Persian (Iranic), Chinese (Sinitic), and Arabic (Semitic) (Sinor 1995) which implies a history of contact between them and the local Central Asian varieties, including Kazakh. The Russian imperial expansion also inevitably led to the prolonged contact of Russian with Kazakh, characterized by an unequal division of power and prestige between these two languages and their respective communities (Smagulova 2006). As such,

Kazakh serves as a fertile ground for investigating language contact and evolution both in this area and more generally.

Much of the existing scholarship on Kazakh linguistics, particularly grammars, including both reference and pedagogical grammars, tend to be prescriptive and anchored in highly standardized literary written language and introspection (Balakaev 1959; 1962; Amanzholov 1994; Zhanpeisov 2002; Zholshayeva 2016). Consequently, spoken Kazakh, with its distinct characteristics, has been largely overlooked, except for a few works analyzing fieldwork interviews¹ (Muhamedowa 2005; 2009). Everyday Kazakh conversations have not yet been the primary focus of any linguistic study. This research seeks to fill this gap and shed light on the intricacies of spoken Kazakh, contributing to a more holistic understanding of the language. Additionally, the majority of existing studies on syntactic constructions in conversation tend to focus on European languages, including German, French, English, Finnish, and Estonian (Günthner 2011a; 2011b; Imo 2011; Pekarek Doehler 2011; Hopper and Thompson 2008; Thompson 2002; Clift 2007; Keevallik 2011). While there are notable studies on Japanese as well (Higashiizumi 2011; Laury and Okamoto 2011; Suzuki 2011), its presence underscores the need for broader typological diversity in these studies.

1.4. Relative Clauses

1.4.1. Definition

Relative clauses are devices that aid in tracking referents that speakers introduce in discourse by establishing their identity and building a common ground with the hearer (Givón 1993: 108). In the Kazakh example in (1), the relative clause (RC), marked with square brackets, identifies the referent of the underlined head noun phrase (Head NP). Thus, RCs

¹ I do not consider fieldwork interviews as naturally occurring events because these are events that mainly happen for the goal of the researcher, not for the social goals of the language consultant. It is an activity that does not normally happen in language consultants' everyday life. Additionally, it has been shown that the interactional dynamics of participants in interviews differ significantly from that of naturally occurring conversations (Troiani, Du Bois & Filchenko, in press).

help delimit the reference of the Head NP by specifying the function it fulfills in the situation expressed in the RC (Andrews 2007: 206). In all the examples that follow, Head NPs are underlined while RCs are shown in square brackets.

- (1) Men [bir şäri-de tur-atın] xan-nıñ bala-sı edi-m.
 1SG one town-LOC live-PTCP khan-GEN child-3.POSS COP.PST-1SG
 ‘I was a child of a khan [who used to live in one town].’
 (modified from Ótött-Kovács 2015: 137)

In the Kazakh relative clause above, the verb in the RC is nominalized (i.e., a non-finite verb form). It is restricting the reference of the NP *xannıñ balası* ‘a child of a khan,’ functioning in exactly the same way as the finite RC in the English translation. In fact, according to Comrie (1989) and Givón (1993), English non-finite participial clauses can fulfill the same functions as English finite relative clauses, as illustrated in (2):

- (2) That woman [sitting at the end of the bar]... (Givón 1993: 113)

Constructions in which there is no verb such as prepositional phrases have also been included into the category of relative clauses as in (3).

- (3) The roof [on your summer cabin] is leaking (Givón 1993: 114)

For these reasons, I adopt a functional definition of relative clauses and will treat constructions with and without verbs alike as relative clauses as long as they modify the Head NP.

So far, I have defined the function of relative clauses as constructions that delimit the reference of an NP. However, it has been shown that relative clauses exhibit several functions in addition to identification. Below is an overview of the different functions RCs have been shown to fulfill.

1. IDENTIFICATION/RESTRICTION. As was mentioned before, relative clauses help single out or identify a referent from potential members of a class (Comrie 1989:138), as in (4).

- (4) The man [that I saw yesterday] left this morning (Comrie 1989: 138).

In (4), assuming that the identity of *the man* is not clear to the hearer yet, before the RC is uttered, *the man* could potentially refer to any man. The RC thus makes an *assertion* about *the man* which specifies *the man's* identity as a particular man that the speaker saw yesterday.

2. COMMENTARY/NON-RESTRICTION. Relative clauses can also serve “to give the hearer an added piece of information about an already identified entity, but not to identify that entity” (Comrie 1989: 138). If we assume that the hearer of the utterance in (4) already knows which man the speaker is talking about, then the RC in brackets is construed as an additional commentary about that man, not essential to establishing the identity of the man while still asserting a certain state of affairs related to the man.
3. GROUNDING. Relative clauses can also “ground a noun phrase,” i.e., “locate its referent in conversational space by relating it to a referent whose relevance is clear, that is, to a Given referent in the immediate context” (Fox and Thompson 1990: 300). In other words, every time an NP is uttered by one speaker, conversationalists need to make that NP relevant to what has been said prior to this moment by both parties or to any information or knowledge shared by them (and hence to any Given information). Among the syntactic resources that allow speakers to accomplish grounding such as *proposition-linking* or *main-clause grounding* (Fox and Thompson 1990: 300-301), relative clauses also fulfill this function. Take a look at example (5).

(5) *This man [who I have for linguistics] is really too much.*

(Fox & Thompson 1990: 301)

When *this man* was introduced, it had to be made relevant to any Given information that the speakers share; thus, the RC *who I have for linguistics* links *this man* to the Given referent who is the speaker himself, indicated by the pronoun *I*. Fox & Thompson (1990) argue that this is not equivalent to making an assertion about an NP as in previous examples.

When a referent has already been grounded by one relative clause or by some other means, another relative clause may then be used to make an assertion about it, as in (6).

(6) *There is a woman [in my class] [who is a nurse]* (Fox and Thompson 2007: 301).

The Head NP in (6) is a newly introduced referent that is not yet in the focal consciousness of the hearer, and its newness is signaled by the indefinite article. This referent is made germane to the conversation (i.e., grounded) by being related to the speaker via the prepositional phrase *in my class* (because *in my class* contains the pronoun *my* pertaining to the speaker), which can also be construed as a verbless relative clause, as discussed before. After the grounding is accomplished, the RC in brackets ‘characterizes’ the identity of the woman (Fox and Thompson 2007: 301), functioning similarly to RCs that give additional commentary.

In my study, I will take the function of RCs as a variable affecting the distribution of RCs in my data, and I will only consider restrictive and non-restrictive RCs because grounding overlaps with both of these functions.

1.4.2. Relative Clauses in Written Kazakh

Muhamedowa (2016: 39) writes: “[a]s our materials show, there is no difference in forming relative clauses between spoken and written Kazakh”. While claiming so, Muhamedowa (2016) does not indicate whether the examples of relative clauses she presents in the book come from either spoken or written form of Kazakh. In contrast, Ótött-Kovács’s (2015) description of relative clauses in Kazakh is based solely on published written texts. Below is the list of characteristics of RCs in written Kazakh according to Ótött-Kovács (2015):

- Kazakh RCs tend to be pre-nominal, i.e., they tend to precede the Head NPs they modify;

- In Kazakh RCs, the predicate of the relative clause is usually non-finite, formed by the use of participle suffixes *-GAn*, *-AtIn* or *-(A)r*.
- Kazakh RCs do not make use of any relative pronouns; instead, they use the ‘gap strategy’.

In languages with a ‘pronoun strategy’ this ‘gap’ is filled with a relative pronoun to explicitly signal the grammatical function of the relativized head. For instance, compare examples (7) and (8) from Russian and Kazakh, respectively:

(7) Knīga, [kotor-wyu ya proçita-l], izmeni-l-a moyu jžn’
 book REL-FEM.ACC 1SG read-PST.M change-PST-F my life
 ‘The book [that I read] changed my life.’ (constructed example)²

(8) [Men oqı-ğan] kitap ömir-im-di özger-t-ti.
 1SG read-PTCP book life-1SG.POSS-ACC change-CAUS-PST
 ‘The book [that I read] changed my life.’ (constructed example)

Notice that while the Russian RC in (7) is post-nominal and contains a finite verb, the Kazakh RC in (8) is pre-nominal and the verb in the RC is nominalized via the participial suffix *-ğan*. In both examples, ‘the book’ is the object of ‘read,’ a verb in the RC. The verb in the Russian RC encodes this information by the use of the accusative suffix on the relative pronoun, whereas the verb in the Kazakh RC does not. Object is not the only grammatical function that Kazakh does not mark in RCs; this is true for all other functions as well (subject, indirect object, genitive, adjunct).

Kazakh also allows ‘headless’ RCs, and they “refer to nonspecific head nouns [and] have the following structure: the relativized verb takes possessive and plural suffixes that could have otherwise been attached to the omitted head noun” (Muhamedowa 2016: 42). This is exemplified in example (9) which is an interrogative sentence with a headless relative clause.

² Constructed examples are examples that are possible to elicit.

- (9) [Qazaqstan-dı alğaş mekende-gen]-der kim-der?
 Kazakhstan-ACC first inhabit-PTCP-PL who-PL
 ‘Who were (the people) [who first inhabited Kazakhstan]?’
 (modified from Muhamedowa 2016: 42)

In (9), the nominalized verb in the RC is taking the plural suffix *-der*; which essentially stands for the omitted Head NP ‘people.’

2. Data & Methods

2.1. Data and Variables

The relative clauses examined in this study are taken from the Multimedia Corpus of Modern Spoken Kazakh (MCSKL) (Filchenko, Troiani, Du Bois & Sarseke et al. 2023) being assembled currently at Nazarbayev University, Astana, Kazakhstan. MCSKL is a first ever corpus dedicated to the documentation of naturally occurring Kazakh discourse. The recorded events primarily encompass conversations but also encompass some instances of ritual language use, museum excursions, and class lectures. As of 2023, some portion of the corpus also represents varieties of Kazakh spoken in other countries such as China (Xinjiang region) and Russia.

MCSKL corpus currently consists of about 150 hours of recordings and 70 hours of transcriptions with at least one level of annotation (IPA, translation, or segmentation at the level of intonation units). Of these hours, 20 hours are fully annotated with an annotation schema which includes: segmentation at the level of intonation units, orthographic transcription, phonetic transcription, morpheme-by-morpheme glossing, part-of-speech tagging, and translation into Russian and English. All the transcriptions and annotations have been done by a team of Kazakh-speaking linguists. The recordings are available either in audio or video format and are accompanied by their respective transcriptions, annotations, and metadata. The database is being updated every month.

The recordings have been transcribed under the Discourse-Functional Transcription framework (Du Bois 1983), with Intonation Units (IUs) representing the basic units of speech. Intonation Units are defined as “a stretch of speech uttered under a single coherent intonation contour” (Du Bois 1983: 47), which have also served as a basis for the Santa Barbara Corpus of Spoken American English (Du Bois et al. 2000-2005). These units have been shown to have an important function in organizing discourse both in cognitive and interactional terms (Fox & Thompson 1990; Troiani 2023). IUs are identified on prosodic grounds, with both native and non-native transcribers with proper training being able to perform segmentation accurately (Troiani 2023).

The length of the transcribed material that I included in my sample totaled approximately 5.25 hours (315 minutes). I manually analyzed the sample of transcripts to identify and extract IUs containing relative clause constructions that satisfy the definition of relative clauses given in Section 1.4.1. I manually coded each relative clause construction for each of the variables listed below in an Excel spreadsheet. A total of 214 relative clauses were collected and coded.

The following variables have been shown to affect the distribution of relative clauses in the literature which will be analyzed in this study using the data from Kazakh:

1. FUNCTION OF HEAD NP IN MATRIX CLAUSE, FUNCTION OF HEAD NP IN RC, and HEAD-RC COMBINATIONS

Following Fox and Thompson (1990), the relative clauses in my data were categorized according to the function of Head NP within the main clause and its function in the RC. Head NP functions in the matrix clause were Subject, Object, Existential Theme, i.e., subject of an existential clause, and Adjunct. Those Head NPs whose grammatical function was not clear, e.g., due to them not being situated in a main clause, and those that functioned as Predicate Nominals were categorized as Other, and hence will not be a focus

of this study. The functions that Head NPs had in the RC were Subject, Object, and Adjunct. Other functions such as Indirect Object or Possessor (Genitive) were not attested even though, in theory, these positions are also relativizable and are possible to elicit from speakers.³

The variable HEAD-RC COMBINATIONS is a combination of FUNCTION OF HEAD NP IN MATRIX CLAUSE and FUNCTION OF HEAD NP IN RC. Examples (10) to (21) illustrate various combinations of these functions of Head NPs attested in the data.⁴

- (10) [subject RC] subject head
 [Kel-gen] qonaq-tar, gostinica-ğa, jat-tı, besplatno.
 come-PTCP guest-PL hotel-DAT lay-PST for.free
 ‘The guests [who came] stayed at the hotel for free.’
- (11) [object RC] subject head
 [Kim-der, ayt-qan] prikol-dar-ı, ne et-pe-ytin bol-dı,
 who-PL say-PTCP joke-PL-3.POSS what do-NEG-PTCP be-PST
 öt-pe-ytin bol-dı
 pass-NEG-PTCP be-PST
 ‘The jokes [that those people said] will no longer be relevant.’
- (12) [adjunct RC] subject head
 Prosto, [bar-atın,] nemene-miz=de normal’niy, bol-uw kerek
 just go-PTCP thing-1PL.POSS=also normal be-INF need
 ‘It is just that the thing [to which we will go] needs be good as well.’
- (13) [subject RC] object head
 [elw mıñ dannie bar,] kod-tı jiber-e-di,
 fifty thousand data EXST code-ACC send-PRS-3
 ‘(They) send code [that has fifty thousand data].’
- (14) [object RC] object head
 se-ni qara-p otır-mız, [sol qara-p otır-ğan,] narse-ñ-di.
 2SG-ACC look-CVB AUX-1PL that look-CVB AUX-PTCP thing-2SG.POSS-ACC
 ‘We are looking at you, at your thing [that you are watching].’

³ I was suggested to do a field questionnaire in order to identify all the possible positions Kazakh can relativize on. However, due to time constraints, I did not do that; instead, I relied on introspection to identify these possible positions.

⁴ Note that punctuation signs in the transcripts signal intonation unit boundaries and encode the intonation of the contour: a period, ‘.’, represents a falling intonational contour, a comma ‘,’ – a continuing contour, and a question mark ‘?’ – a rising contour.

- (15) object head [adjunct RC]
karta aş de-p jatır eşçe. [Stipendiya aqşa awdar-atın]⁵
 card open say-CVB AUX also stipend money transfer-PTCP
 ‘They are also telling (us) to open (bank) cards [to which stipend money will be transferred].’
- (16) [subject RC] existential theme
 [tuwra Nılqı-ğa öt-ip ket-etin] köpir bar
 straight Nylqy-DAT pass-CVB AUX-PTCP bridge EXST
 ‘There is a bridge [which passes straight to Nylky].’
- (17) [object RC] existential theme
 [Al-ıp ket-etin], küim-der, bar arasında.
 take-CVB AUX-PTCP cloth-PL EXST among.them
 ‘Among them, there are clothes [that are to be taken away].’
- (18) [adjunct RC] existential theme – such a combination is not attested in the data.
 An English constructed through introspection would be:
 There is a box [in which I put my belongings].
- (19) [subject RC] adjunct head
 eşçe [qasımda otır-ğan] adam-dar-ğa sovet et-ip
 also near.me sit-PTCP person-PL-DAT advice do-CVB
 ‘I was also giving advice to people [who were sitting near me].’
- (20) [object RC] adjunct head
 komnat-ta [tan-ıtın] adam-men bol-ğan=da jaqsı=ğoy.
 room-LOC know-PTCP person-COM be-PTCP=also good=EMPH
 ‘You know, it is good to live in a room with a person [whom (you) are familiar with].’
- (21) [adjunct RC] adjunct head
 [banki tur-ğan] jer-ge qoy-a sal?
 jar stand-PTCP place-DAT put-CVB AUX
 ‘Just put (it) in the place [where the jars are].’

2. HUMANNES. Whether the Head NP or other NPs in the RC are human or non-human has been shown to affect the distribution of relative clauses (English: Fox and Thompson 1990; 2007; Korean: Kim and Shin 1994; Japanese: Collier-Sanuki 1990; Chinese: Pu 2007; Tao 2002). In English, non-human subject and object Head NPs were shown to pattern systematically with RCs that code specific grammatical

⁵ This is an example of a post-nominal relative clause.

functions (Fox and Thompson 1990). For example, non-human subject heads tended to occur mostly with relative clauses in which they functioned as objects:

(22) *the car [that she borrowed] had a low tire* (Fox and Thompson 1990: 303)

In (22), the car is the subject of the main clause but an object in the relative clause. In contrast, object heads which also functioned as objects in their RCs were not preponderant among non-human referents in English conversations.

Among human referents, NPs comprising a New piece of information that are formulated in non-specific terms tended to be introduced in the object slot of a transitive clause, while those that are formulated as specific tended to be introduced as subjects of *there is/are* or ‘existential’ constructions. Compare (23) and (24):

(23) New, specific human referent
there was a boy [that played the trombone] that he kind of knew (Fox & Thompson 1990: 311)

(24) New, non-specific human referent
and she hates anyone [who isn't a Catholic] (Fox & Thompson 1990: 311)

This is due to the prototypical associations that exist with each grammatical role (Fox & Thompson 1990: 311). The subject role is associated with specificity and definiteness (Givón 1979), and so the existential construction is still able to accommodate specific but indefinite human referents because indefinite referents are of interest to be discussed further. Conversely, the object role is associated with nonspecific referents, often grounded by proposition-linking and not discourse-deployable themselves. Instead, it is usually another referent in the relative clause that gets deployed in the unfolding discourse.

I adopt the binary categorization into human and non-human from Fox and Thompson (1990) for reasons of simplicity, even though Animacy Hierarchy as a wider concept has more nuance (Osten & Fraurud 1996).

3. INFORMATION STATUS OF HEAD NP. This is a concept from the theory of Information Flow (Chafe 1987). This factor has also been shown to influence the patterns in distributions of RCs in prior research (Fox and Thompson 1990; Pu 2007). The following two categories were used to classify the information status of Head NPs:
- GIVEN: A referent which is presumed to be in the interlocutor's focal consciousness, an Active concept/piece of information (Chafe 1987: 26). I evaluated those NPs in the data which have been the main topic of the conversation since their introduction as Given information in the analysis.
 - NEW: A referent which is presumed not to be in the interlocutor's immediate focal consciousness, an Inactive concept/piece of information (Chafe 1987: 31). I evaluated those NPs in the data which have not been previously introduced in the conversation as New information in the analysis.
4. FUNCTION OF RC. In my analysis, this variable is a binary categorical variable with the values 'restrictive' and 'non-restrictive.' I excluded the grounding function of RCs from the analysis since it overlaps with both these functions as explained in Section 1.4.1. In addition to this, almost all Head NPs are grounded by their RCs in the data, whether restrictive or non-restrictive. In the literature, the restrictivity of the RCs has been shown to affect the distribution of RCs. For example, a corpus-driven analysis of non-restrictive relative clauses in spoken English by Tao and McCarthy (2001) has found that non-restrictive *which*-clauses used in everyday conversations fall into three functional categories: evaluative clauses, expansion clauses, and affirmative clauses. The overwhelming majority (62%) were evaluative clauses, i.e., clauses where the speaker expressed their stance/opinion towards the message of the preceding utterance as in (25):

- (25) *It is like if they do not spend two pounds on children for Christmas, it is not enough, [which I think is silly but what's the way of things today].*

Tao and McCarthy (2001) claim that these clauses have a preferred syntactic configuration in discourse which can be schematized as “*which* + modal expressions (including discourse markers) + *is*”, as well as a preferred function in that they are evaluative.

5. POSITION. This variable refers to the relative position of an RC vis-à-vis the Head NP. It is also a categorical variable, with the values ‘headless,’ ‘post-nominal,’ and ‘pre-nominal.’ For example, Wang & Wu (2020) found that post-nominal RCs in Chinese—previously thought to be inadmissible in the syntax of Chinese—function mostly as afterthoughts, the use of which is driven by information structure in spoken discourse and word order.

Table 1 summarizes information on the variables analyzed in this study.

Variable	Variable Type	Values
HEAD-RC COMBINATION (FUNCTION OF HEAD NP IN MATRIX CLAUSE & FUNCTION OF HEAD NP IN RC)	categorical	subject-subject RC (s-s) ⁶
		subject-object RC (s-o)
		subject-adjunct RC (s-a)
		object-subject RC (o-s)
		object-object RC (o-o)
		object-adjunct RC (o-a)
		exst.theme-subject RC (e-s)
		exst.theme-object RC (e-o)
		adjunct-subject RC (a-s)
		adjunct-object RC (a-o)
		adjunct-adjunct RC (a-a)
FUNCTION OF HEAD NP IN RC	categorical	subject
		object
		adjunct
FUNCTION OF HEAD NP IN MATRIX CLAUSE	categorical	subject
		object
		adjunct
		existential theme

⁶ These small-letter notations in which the head role in the matrix is followed by its function in the RC will be used in the conditional inference tree plot in Figure 5 due to limited space in the plot.

		other
HUMANNESS	categorical	human
		non-human
INFORMATION STATUS OF HEAD NP	categorical	Given
		New
FUNCTION OF RC	categorical	restrictive
		non-restrictive
		pre-nominal
POSITION	categorical	post-nominal
		headless

Table 1. Summary of the information on the variables analyzed in this study.

2.2. Methodology

2.2.1. The Statistical Base of the Study

Language use is inherently fuzzy, characterized by probabilistic structures. Therefore, statistics serves as an optimal tool for uncovering meaningful patterns. Statistical analysis is indispensable especially when working with large datasets and corpora (Levshina 2015: 3). Historically, however, statistics was deemed unnecessary by some scholars due to the fundamental assumptions they held regarding language. For example, Bloomfield (1935: 37), the founder of the American structuralist school, wrote:

“Large groups of people make up all their utterances out of the same stock of lexical forms and grammatical constructions. A linguistic observer therefore can describe the speech-habits of a community without resorting to statistics”.

The view that people make up utterances from ‘the same stock of lexical forms and grammatical constructions’ implied that grammar is a set of clear-cut, discrete categories. The scholars who held this view regarded their native speaker knowledge as representing all the necessary information about the entire language. Such an approach to language, consequently, did not necessitate any use of statistics (Levshina 2015: 2). For instance, Chomsky (1957: 17) also downplayed the importance of statistics, claiming that “probabilistic models give no particular insight into some of the basic problems of syntactic structure.”

However, in recent years, quantitative methods have gained popularity in linguistics. While hybrid disciplines such as psycholinguistics, sociolinguistics, and computational linguistics have been relying on statistical techniques for quite some time, “it is only recently that the awareness of its importance has reached the more traditional areas of linguistics” (Levshina 2015: 1). This thesis will contribute to the body of quantitative studies on morphosyntactic variation, aligning with the broader movement toward statistical approaches in linguistic research.

In Sections 2.2.2 and 2.2.3, I will very briefly describe the statistical techniques I used for my analysis.

2.2.2. Measuring Associations Between Categorical Variables

The first statistical technique I employed is the measurement of associations between one or more categorical variables. A *variable* is some property of an object that varies and which can be measured or described (Levshina 2015: 16). For example, the POSITION of a relative clause, whether it is pre-nominal or post-nominal, is a variable. *Categorical variables* are two or more non-numeric categories that are mutually exclusive. POSITION is a categorical variable because it consists of only two possible, mutually exclusive categories or values: ‘pre-nominal’ and ‘post-nominal.’ In my dataset, all variables are categorical; I list and describe them in Section 2.2.4.

When we talk about associations between two categorical variables, we talk about one of the categorical variables – the *dependent* categorical variable – changing as a result of the influence of the other categorical variable – the *independent* categorical variable. For instance, we can hypothesize that the categorical variable POSITION may be dependent on the categorical variable INFORMATION STATUS, which encodes whether the Head NP modified by the RC is either ‘New’ or ‘Given,’ in terms of its discourse salience. In other words, we can hypothesize that whether the RC is pre-nominal or post-nominal may be dependent on the

discourse salience of the Head NP. In order to test this hypothesis, one should run a test of independence. For most of my analyses, I conducted a test of independence called ‘ χ^2 -test’ (‘chi-squared test’). The null hypothesis of the χ^2 -test is that there is no association between the two variables. The results of the test reveal whether the null hypothesis should be accepted or rejected. If the so-called *p*-value of the test is smaller than 0.05, the null hypothesis is rejected; if it is greater than 0.05, the null hypothesis is accepted. On one occasion in my analyses, I used the Fisher’s test of independence instead of the χ^2 -test due to one of the *expected frequency* values for an observation equaling to less than 5, which is one of the rules of thumb when identifying the right test of independence (Levshina 2015: 214). These expected frequencies are frequencies that one can expect for a variable under the null hypothesis, i.e., when this variable is independent of the other variable. The χ^2 -test bases its calculation on these *expected frequencies* as well as *observed frequencies* – actual frequencies observed in the data.

While the χ^2 -test helps confirm the possibility of statistically significant association between two categorical variables, it does not tell us about the direction of the relationship. On certain occasions where specifying the direction of the association was important, I calculated the *odds ratio*. *Odds* are the ratio that compares the chances of X and the chances of non-X (Levshina 2015: 208). *Odds ratio* is simply the ratio of these odds. For instance, odds ratio can help us answer the following question: what are the odds of observing post-nominal RCs *versus* pre-nominal RCs when the Head NP is a Given referent?

Lastly, as an additional test of the strength of the relationship between two variables, I also obtained the *Cramér’s V* scores, which ranges from 0 (no association) to 1 (perfect association) (Levshina 2015: 209).

In the Results and Discussion section of the thesis, I will indicate the values obtained for these tests for the reader’s reference for every association I discuss. All of my statistical

analyses were done using the statistical programming language called R (Posit team 2024). All the code I wrote to analyze my dataset is attached in the appendices to this work.

2.2.3. Random Forest and Conditional Inference Trees

The other two related statistical techniques I used in my analysis are random forests and conditional inference trees. These two methods were introduced to linguistic analysis by Tagliamonte & Baayen (2012) in their paper on the *was/were* variation in English. While the tests of independence outlined in section 2.2.2 assess the correlation between a single dependent and independent variable, these two techniques evaluate the association between a single dependent variable and multiple independent variables simultaneously (multivariate analysis), allowing the identification of complex relationships and patterns in the data. As was shown in Section 2.2.1, my dataset contains multiple variables; therefore, there is a need to take into account the possible interaction between them.

Suppose that in our imagined study of the variation in relative clause POSITION, there are a series of other variables that have been hypothesized to affect POSITION other than INFORMATION STATUS OF THE HEAD NP such as the RESTRICTIVITY OF THE RC (whether the RC is restrictive or non-restrictive), FUNCTION OF THE HEAD NP IN THE RC and FUNCTION OF THE HEAD NP IN THE MATRIX CLAUSE, and so on. We could look at the relationship of POSITION with each of these independent variables individually by running tests of independence for each association. If the results of these individual tests suggest that POSITION is significantly associated with every dependent variable, a multivariate analysis may be warranted which will allow for a comprehensive examination of the joint influence of all independent variables on POSITION, providing a more nuanced understanding of the complex relationships within the dataset. Random forest and conditional inference trees are a form of such multivariate analysis used alternatively to *multivariate linear regression*. While multivariate linear regression evaluates how each predictor (i.e., independent variable) affects the outcome based

on a mathematical equation, random forests, in contrast, work through the entire dataset and establish, by trial and error, if the variable is a useful predictor or not (Tagliamonte & Baayen 2012: 159). Random forests do so by creating multiple conditional inference trees using the data. The creation of these trees involves an algorithmic method called *binary recursive partitioning* which involves several steps (Levshina 2015: 291):

- I. The algorithm tests the associations of every independent variable with the dependent variable and chooses the one that shows the strongest association with the dependent variable.
- II. Next, the algorithm makes a binary split in this variable, creating two subsets which contain the values of the dependent variable.
- III. The first step is recursively reiterated for every subset until no variables display any association with the outcome at the significance level of 0.05.
- IV. The result of this process is a tree-like diagram where each binary split resembles the branches and the leaves of the tree.

Random forests and conditional inference trees are claimed to be particularly useful in situations when the sample size is small and the number of independent variables is large (Levshina 2015: 291), which is true of my dataset: I have as many as 214 observations across 9 variables.

3. Results & Discussion

3.1. Frequency Distributions of FUNCTION OF HEAD NP IN MATRIX CLAUSE & FUNCTION OF HEAD NP IN RC

Table 2 presents a frequency distribution of FUNCTION OF HEAD NP IN MATRIX CLAUSE.

FUNCTION OF HEAD NP IN MATRIX CLAUSE	Frequency
subject	64 (29.91%)
object	24 (11.21%)
existential theme	37 (17.29%)
adjunct	42 (19.63%)

other	47 (21.96%)
Total	214

Table 2. A frequency distribution of FUNCTION OF HEAD NP IN MATRIX CLAUSE.

RCs modifying subject heads are more prevalent in the Kazakh data than those modifying object heads, a pattern reverse to that of English RCs that tend to occur with object heads (Fox & Thompson 1990). Meanwhile, Korean, Japanese, and Chinese exhibit a balanced distribution concerning the relative frequency of subject-head RCs and object-head RCs (Kim and Shin 1994; Collier-Sanuki 1990; Pu 2007). Before offering my own explanation for the preponderance of subject relatives in Kazakh, let me summarize the arguments that Collier-Sanuki (1990) and Hwang (1994) make regarding this difference between Korean & Japanese and English. In order to account for such a difference, both Collier-Sanuki (1990) and Hwang (1994) draw on word order constraints on Information Flow. First, let me visualize and compare the basic structures of Korean, Japanese, and English sentences that contain RCs (see Figures 1 and 2 below).

subject-head RCs:	[RC]S	O	V
object-head RCs:	(S)	[RC]O	V

Figure 1. A schematic representation of the basic structure of Korean and Japanese sentences with RCs

subject-head RCs:	S[RC]	V	O
object-head RCs:	S	V	O[RC]

Figure 2. A schematic representation of the basic structure of English sentences with RCs

Collier-Sanuki (1990) argues that in English, subject heads, whose relevance to current discourse is not clear from prior discourse, need to be grounded by their post-nominal RCs because there is no other element preceding them capable of providing this grounding. When this happens, the English post-nominal subject-head RC “interrupts the flow of information in the main clause, while the one modifying [an object head] does not” (Hwang

1994: 679). This is the reason why English exhibits a preponderance of object-head RCs: they are cognitively more effective — they do not interrupt the flow of information in the main clause because they are preceded by S and V and will have already become grounded by these elements. This renders their post-nominal RCs to serve as ‘characterizing clauses’ rather than ‘grounding clauses’ (Fox and Thompson 1990). In *we get reports [that go to every department]* (Fox and Thompson 1990: 305), the grammatical object *reports* is grounded using the main clause elements *we* and *get*, rendering the clause-final object-head RC to function not as a grounding device but as a ‘characterizing’ clause. By contrast, in Japanese and Korean, neither subject nor object heads have the privilege of main-clause grounding since one of the main-clause elements, namely the verb, comes clause-finally, following subject and object heads. This suggests that these languages should not have any particular preference for either subject-head RCs or object-head RCs which is borne out by the data collected from these languages (Collier-Sanuki 1990; Kim and Shin 1994). The basic structure of Kazakh sentences containing RCs is the same as in Japanese and Korean. Nevertheless, both subject and object heads tend to be modified by pre-nominal RCs, suggesting that the pattern in Figure 1 above should be reflective of Kazakh and that, therefore, there should be no particular preference for either subject or object head RCs in Kazakh as well. However, the percentage of subject heads (29.91%) is two times higher than that of object heads (12.62%) in the data. I propose that the reason for this difference lies in the properties of grammatical subjects and objects that have been demonstrated in previous studies (DuBois 1987; Givón 1983) and their INFORMATION STATUS. In the Kazakh data, object heads tend to be non-human and New⁷, while subject heads tend to encode human

⁷ For object heads, there were 2 Given and 22 New referents, 5 human and 19 non-human referents.

referents⁸, both New (n = 34) and Given (n = 30). Since grammatical subjects have been shown to tend to be human, more topical, discourse-prominent, and agentive, with grammatical objects tending not to exhibit these properties (DuBois 1987; Givón 1983), the preponderance of subject heads is in line with this discourse pattern. Newly introduced subject heads may be recurrently referred to throughout discourse in the form of Given information due to their topicality. In this definition, topicality is best understood in Lambrecht's (1994) terms:

“[t]he topic of a sentence is the thing which the proposition expressed by the sentence is ABOUT. [...] Even though this topic definition is derived from the traditional definition of ‘subject,’ the two notions ‘topic’ and ‘subject’ cannot be conflated. Topics are not necessarily grammatical subjects, and grammatical subjects are not necessarily topics...” (p. 118).

Object heads, however, are less topical which explains a low number of Given object heads in the data; instead, the object slot has been shown to code New referents (DuBois 1987), which is borne out by my data as well. Thus, the non-humanness and low topicality of object heads contributes to the low frequency of object heads in the data, and hence the low frequency of object-head RCs.

Table 3 presents a frequency distribution of FUNCTION OF HEAD NP IN RC.

FUNCTION OF HEAD NP IN RC	Frequency
subject	130 (60.74%)
object	49 (22.90%)
adjunct	35 (16.36%)
Total	214

Table 3. A frequency distribution of FUNCTION OF HEAD NP IN RC

⁸ The distribution of all human referents in the data (N = 70) were skewed towards the subject position, a pattern significant at $\chi^2(2, N = 70) = 30.714, p < .05$.

There is a statistically significant preponderance of subject relatives in the data, at $\chi^2(2, N = 214) = 73.748, p < .05$. In fact, subject relatives ($n = 130/214$) outnumber object relatives ($n = 49/214$) by a ratio of about 3:1 and adjunct relatives ($n = 35/214$) by a ratio of around 4:1. In other words, utterances like (10) are more common than utterances like (11) and (12), repeated below as (27), (28), and (29), respectively.

(27) [subject RC] subject head

[Kel-gen] qonaq-tar, gostinica-ğa, jat-tı, besplatno.
 come-PTCP guest-PL hotel-DAT lay-PST for.free
 ‘The guests [who came] stayed at the hotel for free.’

(28) [object RC] subject head

[kimder, ayt-qan] přikoldar-ı, ne et-pe-ytin bol-dı, öt-pe-ytin bol-dı
 who.PL say-PTCP jokes-3.POSS what do-NEG-PTCP be-PST pass-NEG-PTCP be-PST
 ‘The jokes [that those people said] will no longer work.’

(29) [adjunct RC] subject head

prosto, [bar-atın,] nemene-miz=de normal’ny, bol-uw kerek
 just go-PTCP thing-1PL.POSS=also normal be-INF need
 ‘It is just that the thing [to which we will go] needs be good as well.’

In order to explain the preponderance of subject RCs in Kazakh over object RCs, I involve the cognitive factor of MARKEDNESS: unmarked forms are those that occur relatively more frequent in discourse and are formally less complex, and thus are easier to process, while marked forms are structurally more complex and less frequent in discourse, and thus are harder to process (Givón 1991). For example, in Japanese and Chinese, Pu (2007) and Prideaux (1982) have invoked the concept of MARKEDNESS as an explanatory framework for understanding why subject RCs, wherein the Head NP functions as the subject in the RC, exhibit a higher frequency compared to object RCs, wherein the Head NP functions as an object in the RC. According to Pu (2007) and Prideaux (1982), this is because subject RCs are considered the unmarked structures, while object RCs are viewed as marked structures, which explains the uneven distribution between the two. I argue that the preponderance of subject RCs in Kazakh over object RCs arises from constraints involved in the processing of relative

clauses due to MARKEDNESS, with the internal structure of objects RCs being more marked than that of subjects RCs, rendering their occurrence less preferred.

Numerous studies that focus on the processing of relative clauses find that subject relatives are easier to process than object relatives, a phenomenon termed as ‘subject-object asymmetry’ or ‘subject advantage’ (Turkish: Slobin 1986; Hungarian: MacWhinney and Pléh 1988; Korean: Kwon, Polinsky, and Kluender 2006; Japanese: Miyamoto and Nakamura 2003; Chinese: Lin 2006; Vasishth et al. 2013; Dutch: Frazier 1987; German: Mecklinger et al. 1995, among other many similar studies). Prideaux (1982) offers an explanation behind this subject advantage in Japanese based on MARKEDNESS, which I argue can also be taken as a valid explanation for subject advantage observed in Kazakh. From Figure 4 below⁹, which illustrates the internal structures of subject RCs and object RCs in Japanese as they occur in a main clause, we can see that subject relatives contain zero subjects, while object relatives contain zero objects (RCs are in brackets and zero arguments are indicated by the symbol ‘Ø’).

SRC-S: [Ø(O)V]S	O	V
SRC-O: (S)	[Ø(O)V]O	V
ORC-S: [(S)ØV]S	O	V
ORC-O: (S)	[(S)ØV]O	V

Figure 4. The internal structure of Japanese relative clauses as they occur in a main clause

Prideaux (1982) argues that the subject advantage in Japanese arises due to object relatives being more marked than subject relatives in their internal structure: “OV structures [in which a subject is zero] are more normal and natural [i.e., unmarked] than SV structures [in which the object is zero] (Prideaux 1982: 26).” His participants judged SRC-S and SRC-O structures more natural and comprehensible than ORC-S and ORC-O structures. Pu (2007)

⁹ SRC and ORC stand for subject and object RCs, respectively. S and O stand for subject and object, respectively.

argues that the same constraint of MARKEDNESS on subject advantage is borne out by his data from Chinese. He argues that this unmarked nature of subject relatives in Chinese and Japanese, i.e., of $[\emptyset(O)V]$ structures with zero subject, is due to the subject position being the position where zeroes occur more frequently when compared with the object position (Pu 2007: 40). This is because “zero subjects primarily function as topics in topic chains” and “subject NPs usually have the features of humanness, agentivity, and definiteness, all of which are high in topicality” (Pu 2007: 40).

Like Japanese RCs, Kazakh RCs typically precede their heads (Muhamedowa 2016). Therefore, the structures of sentences with subject and object RCs in Kazakh look like the one illustrated in Figures 4 for Japanese¹⁰. Like Chinese, Japanese, and many Turkic languages, Kazakh allows zero anaphora, or ‘pro-drop’ strategy (Johanson 2021; Muhamedowa 2016). Whether zero anaphora in Kazakh is more common in subject positions than in object positions has not yet been investigated for Kazakh. However, for the related language Turkish, a number of studies on zero anaphora focused on zero subjects (Çynar 2021; Özsoy 1987; Kerslake 1987; Taylan 1986; Enç 1986, among other studies), which gives the impression that zero subjects are perhaps a more pervasive phenomenon than zero objects in Turkic languages. Furthermore, in standard spoken Turkish, topic continuation, or topic chain, to use the term used in Pu (2007), is achieved by zero anaphora (Schroeder 1999). Since in Turkic languages, topics tend to be subjects (Johanson 2021), zero anaphora in Turkish discourse should be expected to affect subjects more frequently than other grammatical roles, rendering zero subjects a more common phenomenon than zero objects. Thus, zero subjects would be expected to be more frequent in Kazakh discourse as well. If in Turkic languages, topics occur sentence-initially and exhibit a tendency to be clausal subjects

¹⁰ Since my data shows that Kazakh, unlike Japanese, allows post-nominal as well as headless relative clauses, pre-nominal relative clauses still tend to be the most preponderant, and this observation is statistically significant both for subject and object relatives at $\chi^2(2, N = 64) = 8.8438, p = 0.01$ and $\chi^2(2, N = 24) = 9, p = 0.01$.

(Johanson 2021), then the use of topical subjects (T) should result in an unmarked word order, i.e., T/SOV, while the use of other grammatical roles as topics should result in a marked word order such as T/OSV if, say, object is topicalized. This difference in MARKEDNESS should be true of Kazakh as well. Thus, since subject RCs contain zero subjects and the internal word order of $\emptyset(O)V$, they are unmarked, and since object RCs contain zero objects and the internal word order of (S) $\emptyset V$, they are marked. In cognitive terms, marked forms in a language are more difficult to process than unmarked forms (Givón 1991). This cognitive constraint brought by MARKEDNESS, thus, interacts with the relative clause constructions available in Kazakh, giving rise to the preponderance of subject RCs in Kazakh.

3.2. A multivariate analysis of HEAD-RC COMBINATIONS and POSITION

In this and the following sections, I will present the results of an analysis of the formal distribution of relative clause constructions, as influenced jointly by the variables previously delineated in Section 2.1. These formal characteristics under investigation include various configurations of Head NP and relative clause represented by the variable HEAD-RC COMBINATIONS as well as the POSITION of the relative clause.

3.2.1. The analysis of HEAD-RC COMBINATIONS as the dependent variable

Figure 5¹¹ on the next page depicts a conditional inference tree run with the variable HEAD-RC COMBINATION as the dependent variable and all other variables as independent ones.

¹¹ Please refer to Table 1 to review the abbreviations used on the plot.

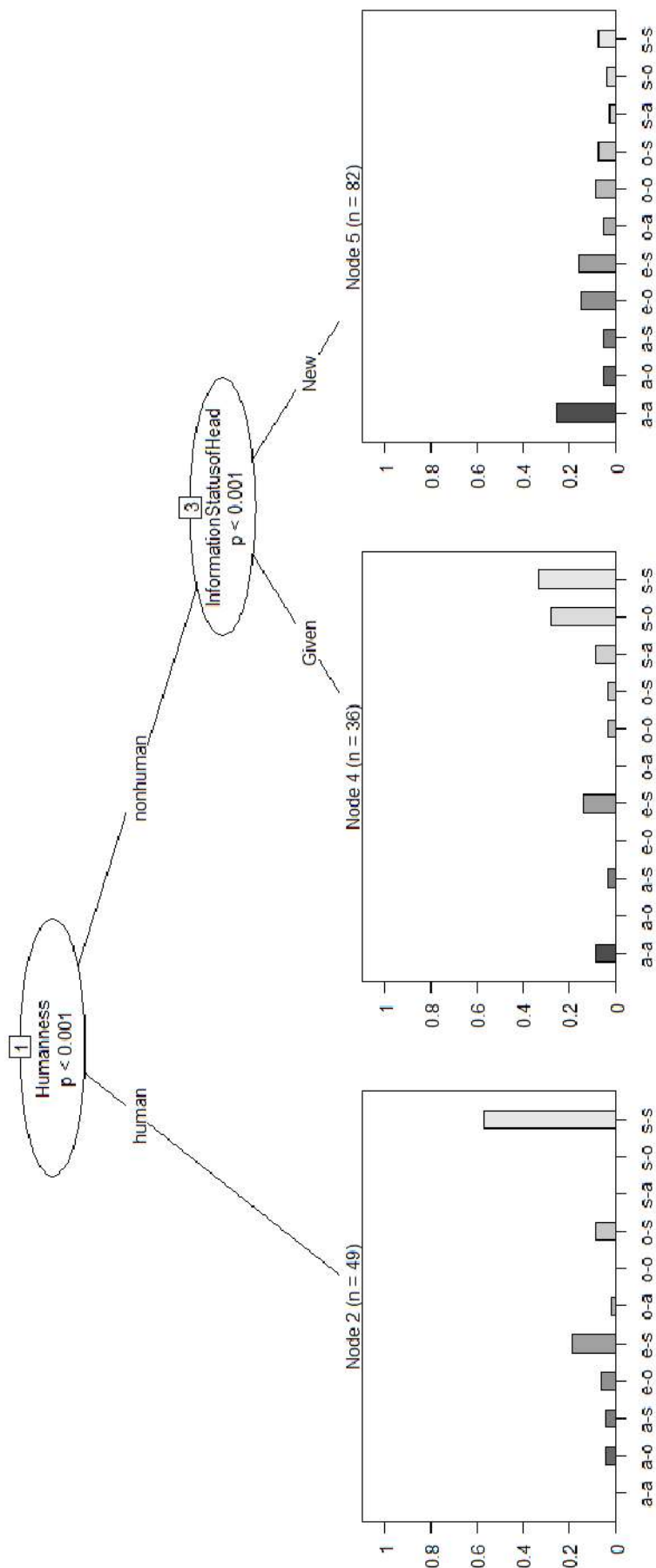


Figure 5. A conditional inference tree for of HEAD-RC COMBINATION as the dependent variables and all other variable as independent ones.

Using the HEAD-RC COMBINATION as the dependent variable and all other variables as independent ones, a random forest analysis was run. Figure 6 is a variable importance plot that was obtained as a result of the random forest analysis.

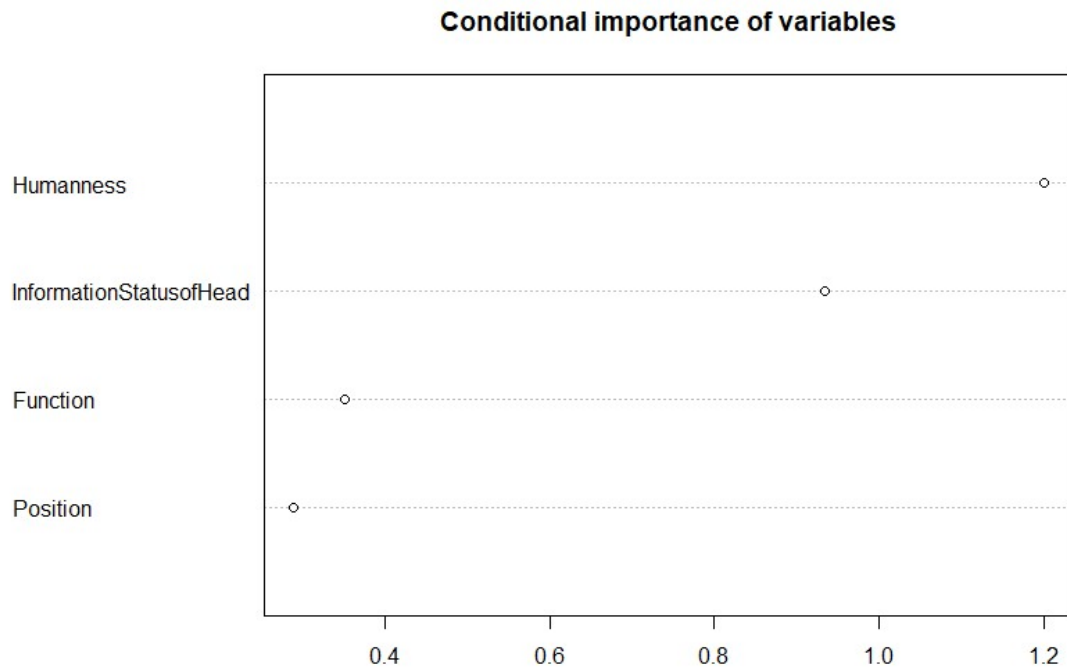


Figure 6. Variable importance scores from the random forest analysis of HEAD-RC COMBINATION as the dependent variables and all other variable as independent ones.

According to Figure 6, HUMANNES is the most important predictor of HEAD-RC COMBINATION, followed by INFORMATION STATUS OF HEAD NP, while FUNCTION OF RC and POSITION are the least important predictors. The conditional inference tree in Figure 5 confirms this because the first split at the top divides the combinations according to the Humanness of the Head NPs (node 1) into human and non-human, with the following split occurring at the left branch with non-human referents (node 3), dividing this subset further into Given non-human referents and New non-human referents according to their INFORMATION STATUS.

Table 4 below shows the confusion plot for the random forest model, which was used to evaluate the accuracy of the model. Each cell in the plot contains the count of instances

where the predicted class, displayed in columns, aligns with the actual class, displayed in rows. For example, the cell at row ‘a-a’ and column ‘a-a’ contains the count of instances the model predicted ‘a-a’ where the actual class was ‘a-a.’ Similarly, the cell at row ‘e-o’ and column ‘a-s’ contains the count of instances where the actual class is ‘e-o’ but the model predicted ‘a-s’.

predicted \ actual	a-a	a-o	a-s	e-o	e-s	o-a	o-o	o-s	s-a	s-o	s-s
a-a	20	4	3	5	8	2	4	5	2	3	6
a-o	0	0	0	0	0	0	0	0	0	0	0
a-s	0	0	0	0	0	0	0	0	0	0	0
e-o	1	0	1	6	3	1	3	1	0	0	0
e-s	0	0	0	1	2	1	0	0	0	0	0
o-a	0	0	0	0	0	0	0	0	0	0	0
o-o	0	0	0	0	0	0	0	0	0	0	0
o-s	0	0	0	0	0	0	0	0	0	0	0
s-a	0	0	0	0	0	0	0	0	0	0	0
s-o	0	0	0	0	0	0	0	0	0	0	0
s-s	3	2	3	3	14	1	1	5	3	10	40

Table 4. The confusion plot for the random forest analysis of HEAD-RC COMBINATION as the dependent variables and all other variable as independent ones.

The *baseline* of the model, i.e., the accuracy achieved by always predicting the majority class, is 0.2754, calculated by dividing the frequency of the most frequent class by the total number of instances. In other words, if the model were to predict the most frequent class for every instance, it would be correct approximately 27% of the time. The *accuracy* of the model measures the proportion of correctly predicted instances out of the total instances in the dataset, and for this model, it is 0.4701, calculated by adding up the counts of true positives and true negatives and dividing by the total number of instances. In this case, the model correctly predicts the class of approximately 40.7% of instances, which is higher than the baseline, suggesting that the model has a moderate performance.

On the leftmost side of the tree (node 2) in Figure 5 which represents human referents ($n = 49$), the combination of subject heads and subject RCs ($n = 28$) is the most preponderant;

no other combinations involving subject heads, i.e., subject & object RC and subject & adjunct RC combinations were contained in this node. Example (1) repeated below as (30) is an example of such a HEAD-RC COMBINATION.

- (30) [subject RC] subject head
 [Kel-gen] qonaq-tar, gostinica-ğa, jat-tı, besplatno.
 come-PTCP guest-PL hotel-DAT lay-PST for.free
 ‘The guests [who came] stayed at the hotel for free.’

The other combinations are used when the Head NP is non-human: 10 out of 13 subject head & object RC combinations are used with Given non-human referents, while 3 – with New non-human referents; similarly, 3 out of 5 subject head & adjunct RC combinations are used with Given non-human referents, while the remaining 2 are used with New non-human referents. The tendency seems to be that subject head & object RC and subject head & adjunct RC configurations tend to be used when the referent is non-human and Given.

As for object heads, there are only 5 humans but 17 non-humans. 4 of the human object heads are used in object & subject RC and 1 is used in an object & adjunct RC combination. The latter combination is used with 2 New non-human referents as well, while the former combination is also used with 6 New non-human referents and 1 Given non-human referent, similar to the object & object RC combination which is used with 7 New non-human Head NPs and only 1 Given non-human referent. The overall tendency is for object heads to be New non-human referents.

Regarding adjuncts, their combination with adjunct RCs is the most prevalent in Node 5, where 21 out of the total 24 such combinations are used with New adjunct heads, with the remaining 3 being Given adjunct heads. Adjunct heads tend not to be human (n = 2) but non-human (n = 33) and New, with adjunct head & subject RC and adjunct head & object RC combinations tending to code New adjunct heads (n = 4 and n = 4, respectively) over the Given ones (n = 1 and n = 0, respectively).

The proportion of RC constructions with existential themes is greater in Node 5 than in Nodes 2 and 4, with existential theme & subject RC and existential theme & object RC combinations tending to code New non-human referents over Given ones (13 New over 5 Given and 12 New over 0 Given, respectively). Overall, existential themes tend to be non-human ($n = 30$) than human (12).

Figure 7 below is a correlation matrix conducted on the variables presented earlier in Section 2.1.

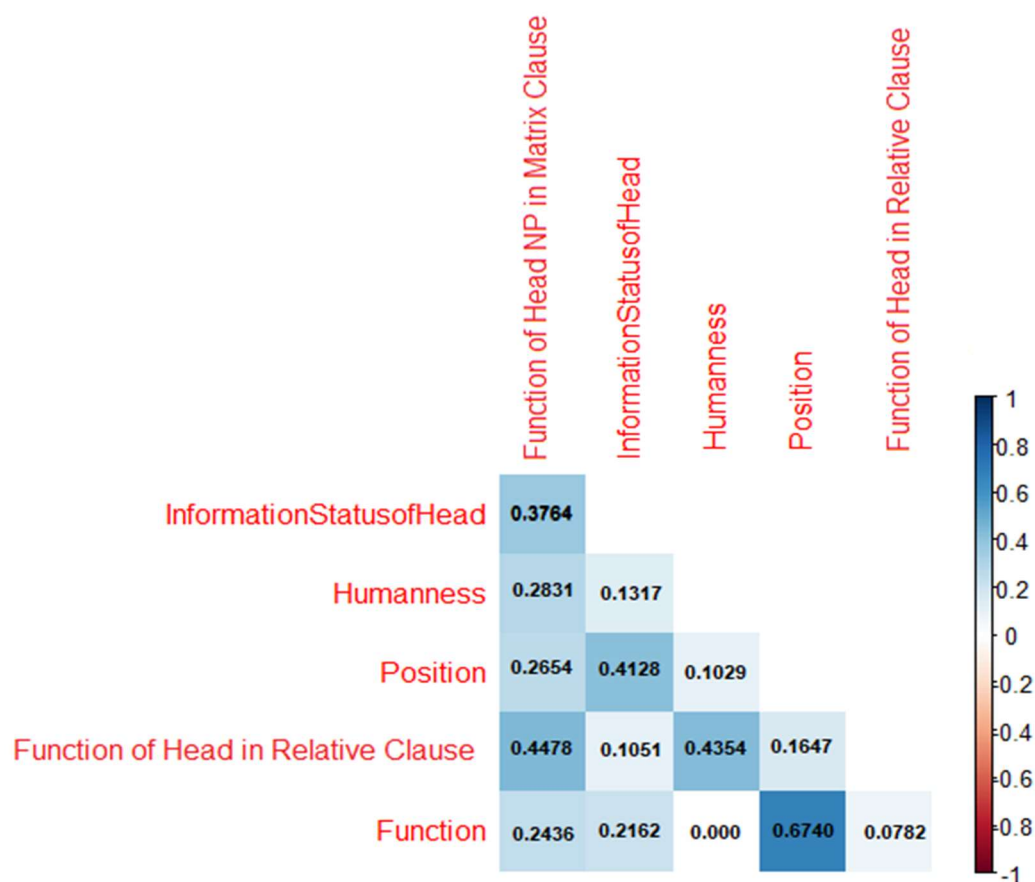


Figure 7. A correlation matrix for the variables in the data

According to Figure 7, POSITION and FUNCTION OF RC are highly correlated, scoring at around 0.6740. The analyses presented in Figures 5 and 6 are done without taking into account the influence of these highly correlated variables in the dataset. Collapsing highly correlated variables in a random forest model improves interpretability while reducing

redundancy, leading to simpler and more efficient predictions and enhancing the model's ability to uncover meaningful patterns and relationships in the data more reliably. Removing these two variables from the previous analyses did not alter the outcome of the conditional inference tree shown and the variable importance scores — HUMANNESS remained the most important predictor of HEAD-RC COMBINATIONS.

From the analyses above, recall that one of the most salient observations made is the predominance of subject head & subject RC combinations. In Pu's (2007) data on Chinese relative clauses, the same pattern is true, and she invoked HUMANNESS as the driving factor: human referents tend to be agentive, topical, and discourse prominent, and hence tend to be subjects, both in the matrix and relative clauses. In her data, there is a statistically significant difference in the humanness of subjects and objects, with subjects tending to encode human referents and objects tending to encode non-human referents, a finding in line with previous research (Givón 1983; DuBois 1987). Recall that the distribution of human Head NPs in my data is also skewed, with the grammatical role of subject being the most frequent position occupied by human referents, while the distribution of non-human Head NPs is not skewed towards any specific grammatical role. Furthermore, recall that human subject heads only occur with subject RCs. When we consider all the variables in the data set to explain this tendency, we find that it is largely predicted by the interaction of HUMANNESS and INFORMATION STATUS OF HEAD NP, according to the conditional inference tree analysis in Figure 5. On the leftmost side of the tree (node 2) which represents human referents ($n = 49$), the combination of subject heads and subject RCs ($n = 28$) is the most preponderant; no other combinations involving subject heads, i.e., subject & object RC and subject & adjunct RC combinations, were contained in this node – their distribution was largely affected by the INFORMATION STATUS OF HEAD NP variable. In other words, if the subject head is human, whether or not it is Given or New is not important for predicting what role it will have in the

relative clause; it is highly likely that it is going to be a subject in the RC as well. When the subject head is non-human, it may still be modified by a subject RC, in addition to being able to be modified by object and adjunct RCs. However, INFORMATION STATUS OF HEAD NP also comes into play, with Given non-human subjects heads being more frequent than New non-human subject heads, which leads to an uneven distribution of all three RC types across these two groups. The tendency seems to be that subject head & object RC and subject head & adjunct RCs tend to be used with Given non-human Head NPs. This can be explained by the tendency in conversations for non-human referents to be linked to human referents who possess, manipulate, or exert some action on them (Du Bois 1980: 269-270).

Recall that there is a predominance of adjunct heads to occur with adjunct RCs. Looking closely into the individual examples in the data, this seems to be due to the collocational preference of adjunct temporal nouns such as *kez* ‘time, period,’ and *kün* ‘day,’ to occur with ‘temporal RCs’ (Tao 2002). In fact, out of 24 adjunct & adjunct RC pairs, 21 adjunct heads were temporal nouns, with *kez* ‘time, period,’ occurring 18 times, *kün* ‘day,’ occurring once, and the remaining two being implied temporal nouns in headless adjunct RCs. In Kazakh, the temporal head *kez* ‘time, period’ is marked for the locative case in a collocational construction expressing time (Muhamedowa 2016: 51), illustrated in (31).

- (31) [adjunct RC (temporal-locative)] adjunct temporal head
 [Üy-de adam joq] kez-de, öziñ=de saw-a al-a-sıñ.
 home-LOC person NEG.COP.EXST time-LOC yourself=also milk-CVB AUX-PRS-2SG
 ‘In the time [when there is no one home], you yourself can milk (cows).’

The remaining three adjunct heads that also function as adjuncts in their relative clause were nouns designating places, two of them being a lexical noun *jer* ‘place’ and the other one being an implied spatial head noun, and these are illustrated in examples (32) and (33).

- (32) [adjunct RC (dative)] adjunct spatial head
 [biz-ge az qal-ğan] jerde, sayaxat bar
 1PL-DAT few remain-PTCP place-LOC sayahat COP.EXST

‘At the place [getting to which takes only a few (miles) for us], there is sayahat.’

- (33) [adjunct RC (locative)] (adjunct spatial head).
 [Mına svetofor-dan arı öt-ken]-de,...¹² bir neme-ge bar-dı-m.
 this traffic.light-ABL father pass-PTCP-LOC one what-DAT go-PST-1SG
 ‘At the (place) [where you pass these traffic lights] ... I went to one place.’

As for adjunct head & subject RC and adjunct head & object RC combinations, the data suggests that these combinations tend to be used with human and non-human heads, including spatial lexical heads. Out of 13 such combinations, 4 were human referents, 3 adjunct heads were non-human entities, 5 were non-human spatial nouns, and only one was a temporal noun. Examples (34) and (35) illustrate the use of these adjunct heads.

- (34) [object RC] human adjunct head
 komnat-ta [tan-ıtın] adam-men bol-ğan=da jaqsı=ğoy.
 room-LOC know-PTCP person-COM be-PTCP=also good=EMPH
 ‘You know, it is good to live in a room with a person [whom you know].’
- (35) [object RC] non-human adjunct head
 [sen kī-ip jür-gen] jaman Bişkek-tiñ älgı büytken-i-nen bes ese.
 2SG wear-CVB AUX-PTCP bad Bişkek-GEN that like.this-3.POSS-ABL five time
 ‘(It) is five times (better) than that bad thing like this [that you are wearing].’
- (36) [subject RC] non-human spatial adjunct head
 Men [ber jaq-ta-ğı,] Zerde-den tüs-ip qal-ayın=ba?
 1SG nearby side-LOC-ADJZ Zerde-ABL get.off-CVB AUX-1SG.HORT=Q
 ‘Shall I get off at Zerde [which is on the nearby side]?’

This finding can be linked to another tendency described earlier, namely that of adjunct human heads disfavoring adjunct relatives in adjunct & adjunct RC combinations, and this happens precisely because human heads are mostly preferred in adjunct head & subject RC and adjunct head & object RC combinations.

Even though adjunct head & subject RC and adjunct head & object RC combinations tend to occur with human and non-human referents, including non-human spatial lexical

¹² The three dots mean that I have omitted some portion from this utterance.

heads, it is, however, equally possible to have these heads to function as adjuncts in their relative clauses (i.e., to occur in adjunct head & adjunct RC combinations), as in English examples from (37) to (39) derived through introspection.

- (37) non-human spatial adjunct head [adjunct RC (locative)]
I went to the new restaurant [where they make modern Kazakh cuisine].
- (38) non-human adjunct head [adjunct RC (instrumental)]
The musician dared to perform with an instrument [with which he had issues].
- (39) human adjunct head [adjunct RC (dative/benefactive)]
 I took a walk with my colleague [for whom you sang a song].

In fact, in my data, spatial adjunct heads did occur as adjuncts in their relative clauses as I have shown earlier in examples (32) and (33). Non-human spatial adjunct heads are the only category that can function as adjuncts within their RCs in addition to being able to function as subjects and objects in their RCs. (40) is another example presented earlier for such an adjunct & adjunct RC combination in which the adjunct head is a spatial noun.

- (40) [adjunct RC (locative)] non-human spatial adjunct head
 [banki tur-ğan] jer-ge qoy-a sal?
 jar stand-PTCP place-DAT put-CVB AUX
 ‘Just put (it) in the place [where the jars are].’

However, there are no examples in my data of human and non-spatial non-human adjunct heads being used as adjuncts in the relative clause. This opens up a question regarding why non-human spatial adjunct heads are more suited to function as adjuncts in their RCs while human and non-spatial non-human adjunct heads are not when, theoretically, they are.

I explain this by considering the semantic properties of these lexical nouns and their role in Information Flow. Adjunct heads are generally associated with New Information, i.e., they tend to be Non-Identifiable and Non-Given (Thompson 1997: 70). As we know, New Referents need to be made relevant for interlocutors at the point of their introduction by means of grounding. Accordingly, it has also been noted that New human referents are usually

grounded by being linked to their own activities, while New non-human object referents are usually grounded by being linked to Given human referents who possess, manipulate, or exert some action on them (Fox and Thompson 1990). From this it follows that if the New adjunct head functions as an object (usually non-human) in the RC, then the RC should be expected to include a Given human subject head to whom it can be related. Similarly, if the adjunct head functions as a subject (usually human) in the RC, the RC may be expected to include a non-human object. In example (41), a human adjunct is anchored in the RC to an object *ne* ‘what’ by being coded as the subject of the verb, while in example (42), the non-human adjunct head *jaman Biškektiñ älgı büytkeni* ‘that Bishkek’s bad thing like this’ is anchored by being linked to a second person pronoun of a transitive verb in the RC.

- (41) [subject RC] human adjunct head
 [ne et-ken] adamdar-ğa, jumıs köp qoy.
 what do-PTCP people-DAT work a.lot EMPH
 ‘For people [who do such things], jobs are abundant.’
- (42) [object RC] non-spatial non-human adjunct
 [sen kï-ip jür-gen] jaman Bişkek-tiñ älgı büytken-i-nen bes ese.
 2SG wear-CVB AUX-PTCP bad Bıshek-GEN that like.this-3.POSS-ABL five time
 ‘(It) is five times (better) than that Bıshek’s bad thing like this [that you are wearing].’

For non-human spatial adjuncts, however, their core semantic meaning is that of locations and physical entities in which something may be located, stored, or kept, which is why the retention of their adjunct function in RCs is not surprising. Human subject referents and non-spatial non-human referents, however, are typically associated with core argument functions rather than non-core (i.e., adjunct or oblique) functions. Thus, the tendency of non-spatial non-human and human adjunct referents not to occur with adjunct relatives is constrained by their semantic features as well as their INFORMATION STATUS.

In sum, we have seen that the tendency of adjunct heads to favor adjunct relatives can be explained by the collocational preference of temporal adjunct heads to occur with temporal relative clauses. We have also seen that human and non-spatial non-human adjunct heads tend to be used by speakers in adjunct & subject RC and adjunct & object RC combinations. This is because such nouns exhibit properties of prototypical agent-like subjects and patient-like objects, which is why their information status as New adjunct referents can be taken care of object and subject relatives that ground them accordingly. Finally, non-human spatial adjunct heads, by virtue of their semantics, are able to function as adjuncts in their relative clauses like their temporal counterparts in addition to being able function as objects and subjects like their human and non-spatial non-human counterparts, which allows speakers to ground New non-human spatial heads using either of the three RC combinations.

3.2.2. The analysis of POSITION as a dependent variable

A conditional inference tree run on the variable POSITION as a dependent variable and other variables as independent variables, excluding HEAD-RC COMBINATIONS. Figure 8 on the next page depicts this tree.

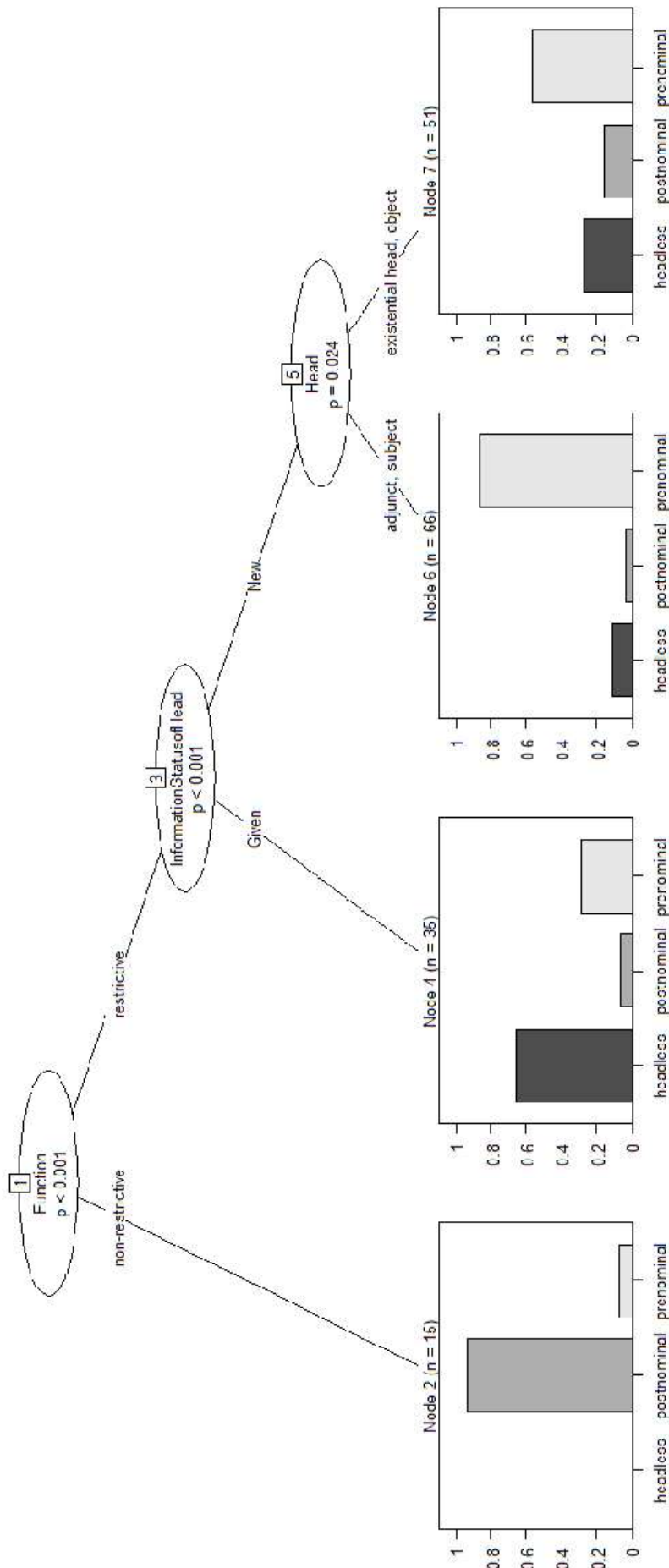


Figure 8. A conditional inference tree for of POSITION as the dependent variables and all other variable as independent ones.

Using POSITION as the dependent variable and all other variables as independent ones, a random forest analysis was run. Figure 9 is a variable importance plot that was obtained as a result of this random forest analysis.

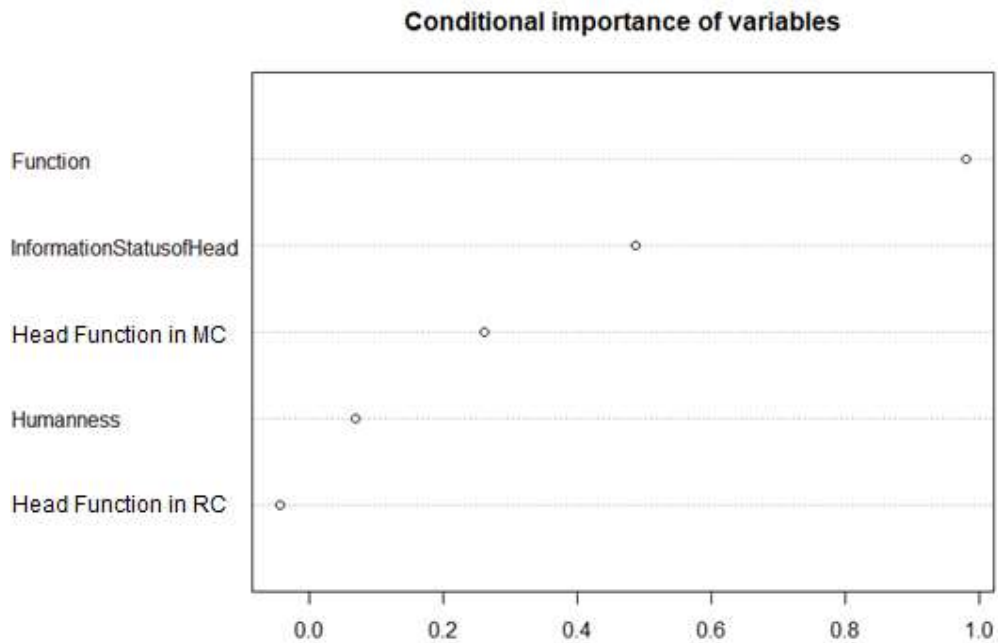


Figure 9. Variable importance scores from the random forest analysis of POSITION as the dependent variables and all other variable as independent ones.

According to Figure 9, FUNCTION OF RC is the most important predictor of POSITION , followed by INFORMATION STATUS OF HEAD NP, while HEAD FUNCTION IN MC have very little predictive power, with HUMANNES and HEAD FUNCTION IN RC being virtually negligible. The outcome of the conditional inference tree analysis suggests that post-nominal RCs tend to be non-restrictive, while restrictive RCs tend to be headless when the referent is Given or pre-nominal when the referent is New. While for New adjunct and subject heads tend to occur with restrictive pre-nominal RCs more often than with restrictive headless and post-nominal RCs, New object and existential themes occur with the latter two a little more often, at the same time preferring to occur with pre-nominal RCs.

Table 5 below shows the confusion plot for the random forest model, which was used to evaluate the accuracy of the model.

predicted actual	headless	post-nominal	pre-nominal
headless	21	2	8
post-nominal	0	14	1
pre-nominal	23	10	88

Table 5. A confusion plot for the random forest model

The baseline accuracy of the model is 0.5808, while the true accuracy of the model is 0.7365, which signals a moderate performance of the model.

Since FUNCTION and POSITION are highly correlated according to the correlation matrix in Figure 7, these analyses were repeated again without the variable FUNCTION. The results of the conditional tree analysis and the variable importance scores for the second round of analyses are presented in Figures 10 and 11 on the next pages, respectively.

Table 6 below shows the confusion plot for the collapsed random forest model, which was used to evaluate the accuracy of the second model.

predicted actual	headless	post-nominal	pre-nominal
headless	21	10	9
post-nominal	0	0	0
pre-nominal	23	16	88

Table 6. A confusion plot for the collapsed random forest model

This time, the baseline accuracy of the model is 0.5808, while the true accuracy of the model is 0.6527, which also signals a moderate performance of the model.

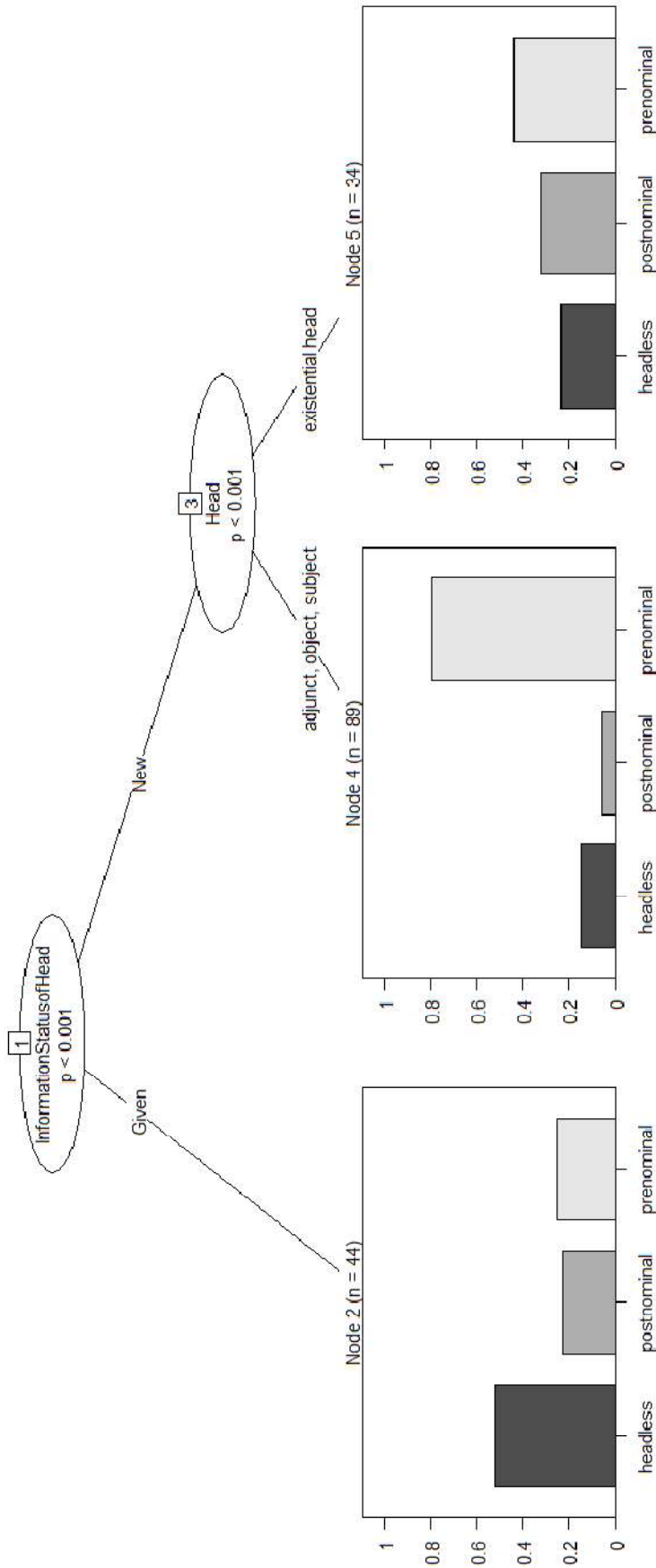


Figure 10. A conditional inference tree for of POSITION as the dependent variables and all other variable as independent ones after collapsing the highly correlated variables.

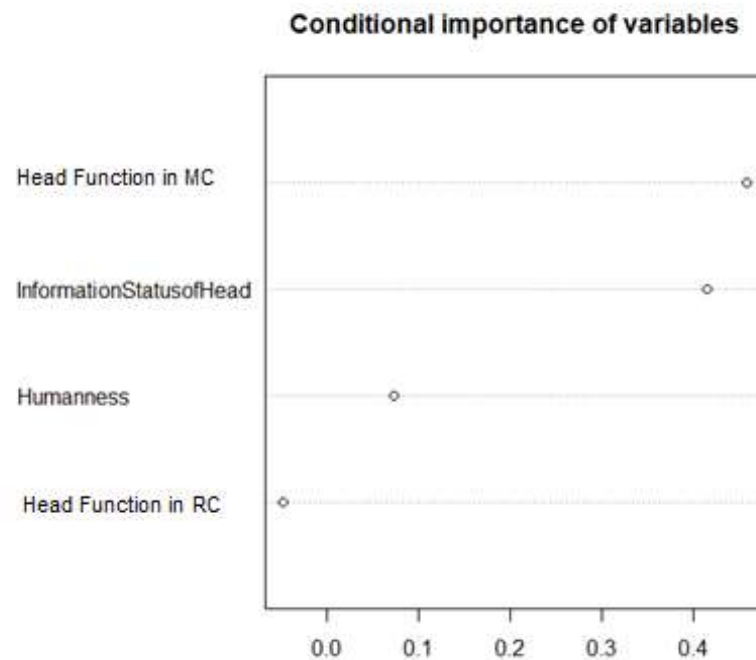


Figure 11. Variable importance scores for the collapsed random forest analysis of POSITION as the dependent variables and all other variable as independent ones.

A collapsed random forest analysis shows that FUNCTION OF HEADN NP MC is now higher in predictive power than INFORMATION STATUS OF HEAD HP. This is also reflected in the second conditional inference tree analysis. At node 3, there is a two-way split for all the New Head NPs according to their FUNCTION IN MC: adjunct, object, and subject heads versus existential themes. While the former three tend to be modified by pre-nominal RCs, existential themes do not have a sharp skew across the three values of POSITION. Given head NPs, represented by the leftmost split from node 1 at INFORMATION STATUS OF HEAD NP, are modified by post-nominal and pre-nominal RCs equally while giving more preference to headless RCs.

As the analyses show, in general, POSITION is primarily affected by the interaction of FUNCTION OF RC and INFORMATION STATUS OF HEAD NP. Recall the outcome of the conditional inference tree analysis suggesting that post-nominal RCs tend to be non-restrictive, while restrictive RCs tend to be headless when the referent is Given or pre-nominal when the referent is New. Let me now attempt to explain why this has to be the case. Examples 43-46

illustrate a non-restrictive post-nominal RC, restrictive headless and restrictive pre-nominal RC, respectively.

- (43) existential head [subject RC]
mamandıq-tar bar ğoy, [bıznes-qa jaqın]
 major-PL EXST EMPH business-DAT close
 ‘You know there are majors [which are close (in focus) to business.]’
- (44) [object RC (Given non-human subject head)]
osılar ğoy, [ötkende, qara-ğan-ımız.]
 these EMPH a.while.ago see-PTCP-1PL.POSS
 ‘(Our clothes), [which we saw a while ago] are these.’
- (45) [subject RC] New human subject head
 [biz-diñ jumıs-ta-ǵı] bir qız aytpaqşı, florist
 1PL-GEN work-LOC-ADJZ one girl by.the.way florist
- bol-ıp jumıs ist-eyin de-gen
 be-CVB work do-HORT say-PST
- ‘By the way, one girl [from our work] was planning to work as a florist.’

Non-restrictive RCs in the data strongly tend to be post-nominal. Comrie (1989) writes that “the non-restrictive relative is a way of presenting new information on the basis of the assumption that the referent can already be identified” (p. 139) and Ariel (1990) writes that “[non-restrictive relative clauses] can and often do introduce New information [...] Intonationally (and in English punctuationally as well), [non-restrictive relative clauses] differ from [restrictive relative clauses] in that they call for a break before the relative clause is uttered, again pointing to the relative separation” (p. 152). From this it follows that non-restrictive RCs, by virtue of functioning as non-essential for the identification of a referent but as a New piece of information, are likely to be separated from an utterance that properly identifies the referent. In the case of Kazakh, a post-nominal RC accomplishes this separation from the antecedent head NP.

Second, RCs in the data tend to be headless when modifying Given referents and pre-nominal when modifying New referents. Given referents are typically more accessible or

already introduced in the discourse, whereas New referents require more attention and contextualization. Since in headless relative clauses, the head noun is already known or easily retrievable from the context, we can assume that headless RCs are then used to provide additional information about Given referents. Conversely, pre-nominal relative clauses are preferred for New referents to help establish their identity and prominence within the discourse. This explains the tendency, mentioned above, of adjunct heads preferring to occur with pre-nominal RCs over headless and post-nominal RCs. This is because adjunct heads usually present New information (Thompson 1997: 70).

3.3. Pairwise Comparisons between Variables

Below I present the results of the pairwise association analyses conducted among the six variables (excluding the HEAD-RC COMBINATIONS variable), comprising a total of 15 pairs. I will present the pairs in order of their strength as indicated in the correlation matrix presented earlier in Figure 7.

3.2.1. POSITION & FUNCTION OF RC

Figure 12 below illustrates the relationship between POSITION & FUNCTION OF RC.

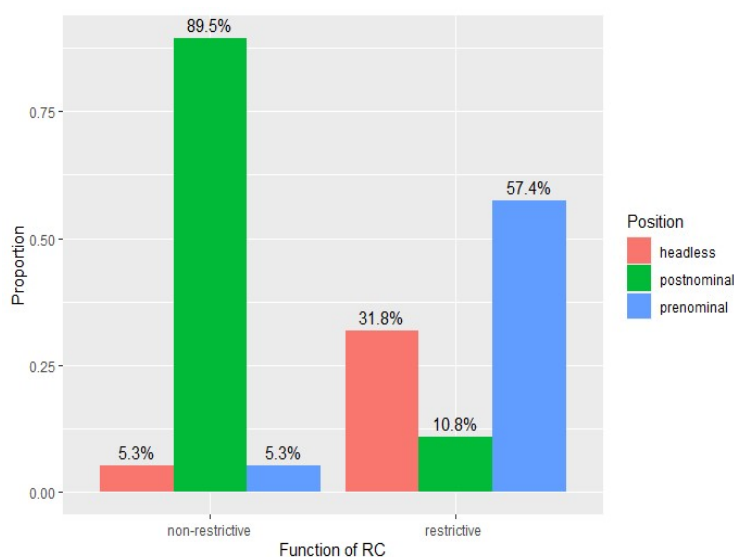


Figure 12. The distribution of POSITION across FUNCTION OF RC.

This distribution of the observed values of POSITION across FUNCTION OF RC is significantly different from the expected values under a chi-squared test at $\chi^2(2, N = 214) = 73.46, p < .05$. The strength of this relationship is moderate (Cramér's $V = 0.586$), with non-restrictive RCs tending to be post-nominal and restrictive RCs tending to be pre-nominal. I accounted for this pattern in 3.2.2.

3.2.2. FUNCTION OF HEAD NP IN MATRIX CLAUSE & FUNCTION OF HEAD NP IN RC

Figure 13 below shows the distribution of FUNCTIONS OF HEAD NP IN MATRIX CLAUSE across FUNCTIONS OF HEAD NP IN RC.

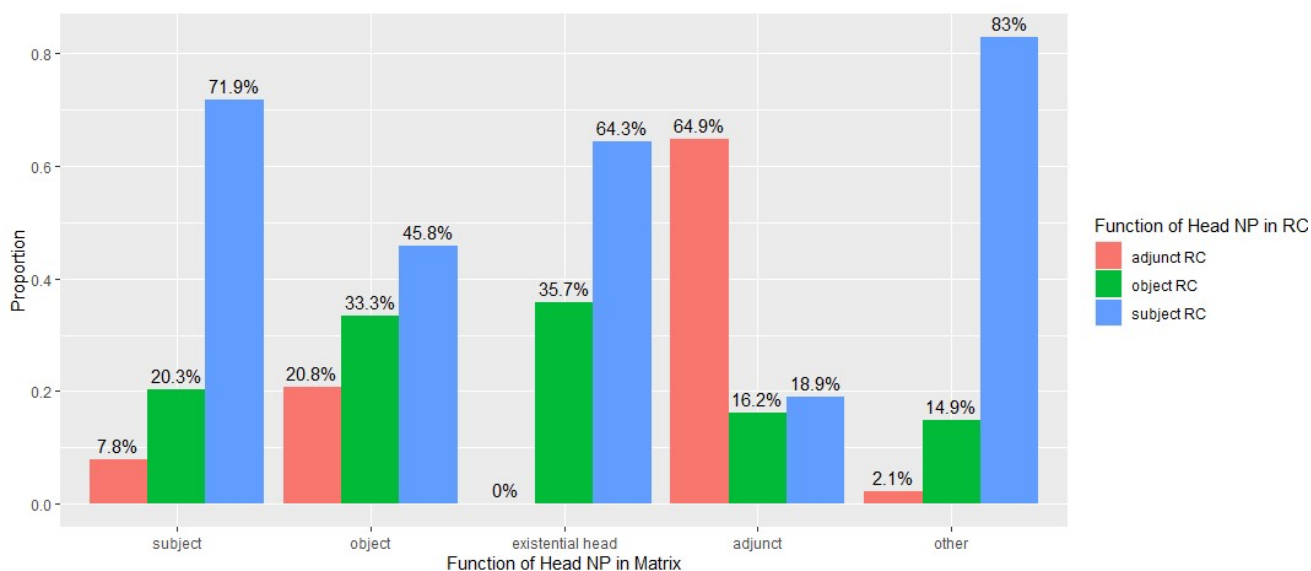


Figure 13. The distribution of FUNCTIONS OF HEAD NP IN MATRIX CLAUSE across FUNCTIONS OF HEAD NP IN RC

This distribution of the observed values of FUNCTIONS OF HEAD NP IN MATRIX CLAUSE across FUNCTIONS OF HEAD NP IN RC is significantly different from the expected values under a chi-squared test at $\chi^2(8, N = 214) = 92.2, p < .05$. The strength of this association is moderate, at Cramér's $V = 0.464$. In other words, overall, speakers' choice of Head NP functions in matrix clauses is moderately sensitive to the grammatical function that this Head NP will have in the RC.

When this association is assessed overall, the distribution of adjunct heads and adjunct RCs are the most skewed. A mosaic plot of χ^2 residual values for each category in Figure 14

shows that there are large statistically significant discrepancies between the expected and observed values for the adjunct head & adjunct RC, existential theme & adjunct RC, and adjunct head & subject RC combinations.

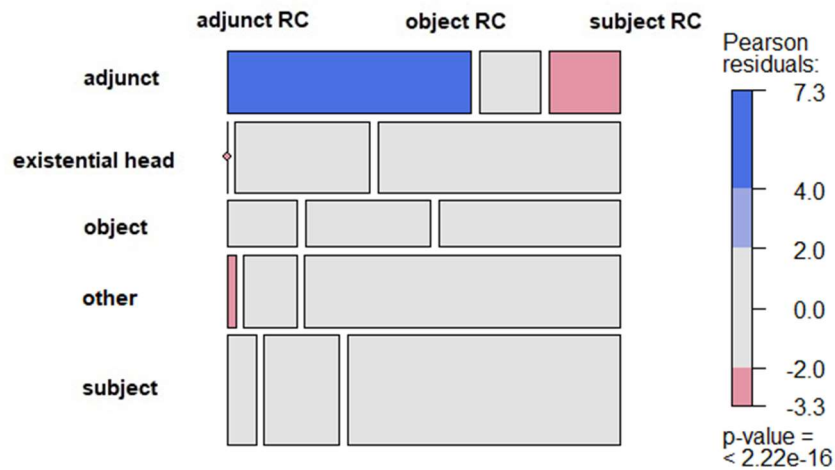


Figure 14. A mosaic plot for the distribution of FUNCTION OF HEAD NP IN MATRIX CLAUSE across FUNCTION OF HEAD NP IN RC

Adjunct heads tend to reserve a strong preference to occur with adjunct RCs while strongly dispreferring subject RCs, as was also confirmed by the conditional inference tree analysis. Existential themes do not, at all, tend to occur with adjunct RCs. In other words, adjunct heads tend to occur in utterances like (21) rather than in utterances like (19) repeated below as (46) and (47), respectively, while existential themes do not occur in utterances like a constructed example in (48).

- (46) [adjunct RC] adjunct head
 [banki tur-ğan] jer-ge qoy-a sal?
 jar stand-PTCP place-DAT put-CVB AUX
 ‘Just put (it) in the place [where the jars are].’
- (47) [subject RC] adjunct head
 eşçe [qasımda otır-ğan] adam-dar-ğa sovet et-ip
 also near.me sit-PTCP person-PL-DAT advice do-CVB
 ‘I was also giving advice to people [who were sitting near me].’
- (48) [adjunct RC] existential theme (constructed example)
 Biz-diñ üy-de [et saqta-ytın] jer bar.
 1PL-GEN home-LOC meat store-PTCP place COP.EXST
 ‘There is a place at our home [in which to store meat].’

Unfortunately, the studies on the distribution of relative clauses in other languages (Kim and Shin 1994; Collier-Sanuki 1990; Fox and Thompson 1990; Pu 2007) focus mostly on objects and subjects, ignoring other grammatical roles, which makes it hard to make any comparison regarding the distribution of adjuncts between Kazakh and these languages.

In addition, the distribution of observed values for all matrix functions of Head NPs other than the object role across the three functions in RCs is significantly different from the expected values under a chi-squared test. For instance, the observed distribution of subject heads across the three functions in RCs is statistically different from its expected distribution at $\chi^2(2, N = 64) = 44.281, p < .05$, with subject heads predominantly occurring with subject RCs, almost five times and ten times more frequent than with object and adjunct RCs, respectively. In other words, an utterance like (10) is much more statistically preferred than utterances in (11) and (12), repeated below as (49), (50), and (51), respectively.

- (49) [subject RC] subject head
 [Kel-gen] qonaq-tar, gostinica-ğa, jat-tı, besplatno.
 come-PTCP guest-PL hotel-DAT lay-PST for.free
 ‘The guests [who came] stayed at the hotel for free.’
- (50) [object RC] subject head
 [Kim-der, ayt-qan] prikol-dar-ı, ne et-pe-ytin bol-dı,
 who-PL say-PTCP joke-PL-3.POSS what do-NEG-PTCP be-PST
 öt-pe-ytin bol-dı
 pass-NEG-PTCP be-PST
 ‘The jokes [that those people said] will no longer be relevant.’
- (51) [adjunct RC] subject head
 Prosto, [bar-atın,] nemene-miz=de normal’niy, bol-uw kerek
 just go-PTCP thing-1PL.POSS=also normal be-INF need
 ‘It is just that the thing [to which we will go] needs be good as well.’

Similarly, in Chinese, subject heads tend to be modified by subject RCs, especially when the referent is human (Pu 2007). In Korean, the same is true also for non-human referents (Kim and Shin 1994). The tendency of subject RCs to modify subject heads is also found in written Japanese (Collier-Sanuki 1990). Fox and Thompson, in contrast, have found

that the reverse pattern is true for English non-human referents: subject heads do not tend to occur with subject heads but tend to occur with object relatives (1990: 302)

Existential themes did not occur with adjunct relatives, and this pattern is significant.¹³ In English, existential themes favor subject RCs over object RCs (Fox and Thompson 1990), while existential themes in Korean were indifferent to the choice of function of the Head NP in the RC (Kim and Shin 1994). For Chinese and Japanese, there are no discussions of existential themes in relative clauses (Pu 2007; Collier-Sanuki 1990).

3.2.3. HUMANNES & FUNCTION OF HEAD NP IN RC

This distribution of the observed values of FUNCTION OF HEAD NP IN RC across HUMANNES is significantly different from the expected values under a chi-squared test at $\chi^2(2, N = 214) = 41.567, p < .05$. The proportion of grammatical functions within RC for both human and non-human referents is illustrated in Figure 15.

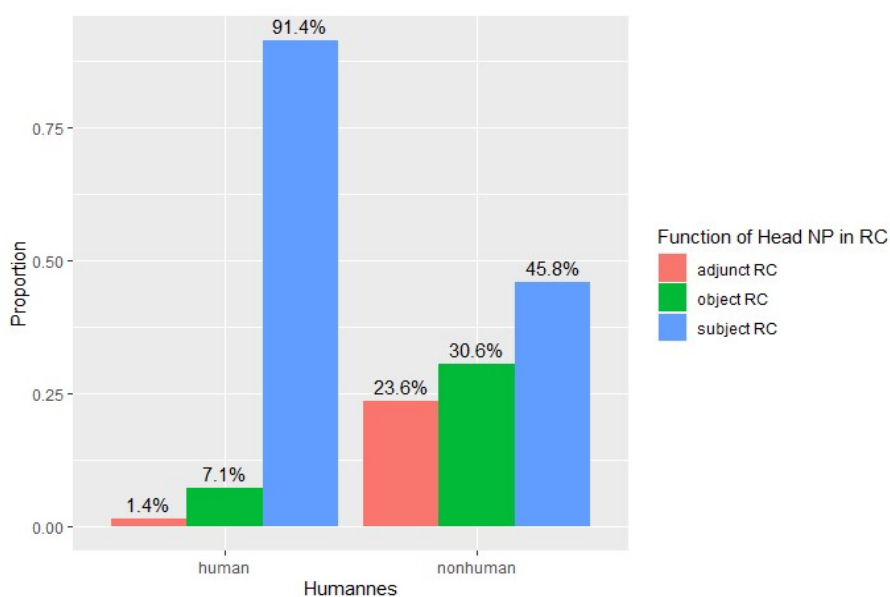


Figure 15. The distribution of HUMANNES across FUNCTION OF HEAD NP IN RC

¹³ The results of a chi-square test for existential themes are significant at $\chi^2(2, N = 42) = 26.143, p < .05$

The skews in this distribution suggest that subject relatives tend to modify human referents rather than non-human referents, which is consistent with the findings for Chinese (Pu 2007). Adjunct RCs tend to modify non-human referents rather than human referents, with object RCs tending to occur very rarely with human referents as well.

If we take a closer look at the distribution of both FUNCTION OF HEAD NP IN MATRIX CLAUSE and FUNCTION OF HEAD NP IN RC considering the humanness of each Head NP, we will get the following distribution illustrated in Table 7.

Function in RC	Subject		Object		Adjunct		Total	
	Human	Non-human	Human	Non-human	Human	Non-human	Human	Non-human
Subject	28 (100%)	18 (50.00%)	0	13 (34.11%)	0	5 (13.89%)	28	36
Object	4 (80.00%)	7 (36.84%)	0	8 (42.11%)	1 (20.0%)	4 (21.05%)	5	19
Adjunct	2 (50%)	5 (15.15%)	2 (50%)	4 (12.12%)	0	24 (72.73%)	4	33
Existential Theme	9 (75%)	18 (60%)	3 (25%)	12 (41.38%)	0	0	12	30
Other	21 (100%)	18 (69.23%)	0	7 (26.92%)	0	1 (3.85%)	21	26
Total	130 (60.74%)		49 (22.90%)		35 (16.36%)		214	

Table 7. The distribution of FUNCTION OF HEAD NP IN MATRIX CLAUSE across FUNCTION OF HEAD NP IN RC according to the HUMANNESSE of Head NPs

According to Table 7, there were no human subject heads occurring with object and adjunct RCs in the data. A closer look at the distribution of human and non-human subject heads gives the following picture in the mosaic plot in Figure 16.

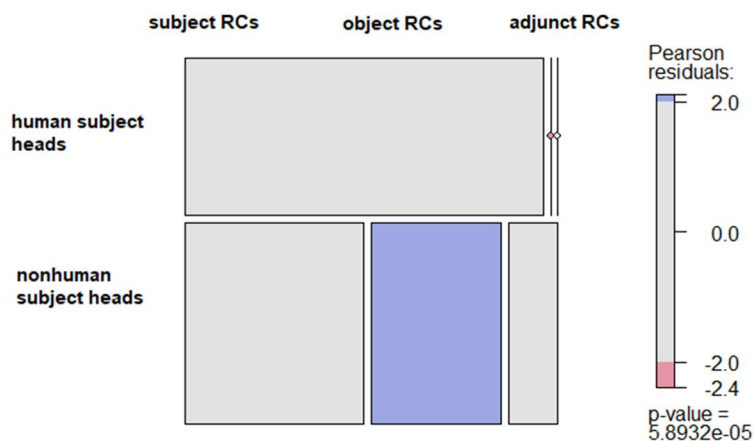


Figure 16. The mosaic plot for the distribution of HUMANNESSE of subject heads across FUNCTION OF HEAD NP IN RC

We find that while non-human subject heads favor object relatives¹⁴, human subject heads tend not to occur with object relatives. That is to say, utterances like in (52) are more common than utterances like in (53) which is possible to elicit but not attested in practice.

(52) [object RC] non-human subject
 [Iste-ytin] zattar köp de-y-di=dä.
 do-PTCP things a.lot say-PRS-3=EMPH
 ‘Things [to be done] are abundant, (they) say, you know.’

(53) [object RC] human subject
 [jaqsı kör-gen] qızdar-ı köp deydi.
 good see-PTCP girls-3.POSS many say-PRS-3.
 ‘The girls [whom he likes] are abundant, (they) say.’ (constructed example)

There were no human object heads modified by object RCs and no human adjunct heads modified by adjunct RCs in the data either.

3.2.4. POSITION & INFORMATION STATUS OF HEAD NP

This distribution of the observed values of POSITION across INFORMATION STATUS OF HEAD NP is significantly different from the expected values under a chi-squared test at $\chi^2(2, N$

¹⁴ This is a statistically significant observation at $\chi^2(2, N = 64) = 19.478, p < .05$. The discrepancies observed between observed and expected frequencies for the human subject & object RC and non-human subject & object RC combinations are significant, with their standardized Pearson’s residuals equaling -3.562086 and 3.562086, respectively.

= 214) = 34.68, $p < .05$ and is of moderate strength at Cramér's $V = 0.403$. Figure 17 is a mosaic plot for this association. The mosaic plot suggests that headless RCs tend to occur with Given referents, while pre-nominal RCs tend to occur with New referents. Post-nominal RCs do not exhibit any significant skews in their distribution.



Figure 17. A mosaic plot for distribution of INFORMATION STATUS OF HEAD NP across POSITION

3.2.5. FUNCTION OF HEAD NP IN THE MATRIX CLAUSE & INFORMATION STATUS OF HEAD NP

This distribution of the observed values of FUNCTION OF HEAD NP IN MATRIX CLAUSE across INFORMATION STATUS OF HEAD NP is significantly different from the expected values under a chi-squared test at $\chi^2(4, N = 214) = 27.546, p < 0.05$. Figure 18 presents the proportions of Given and New Head NPs distributed across the matrix roles. While New Head NPs tend to be adjuncts, existential themes, and objects New¹⁵, subjects do not exhibit a statistically significant preference for being either New or Given.

¹⁵ These figures are significant at $\chi^2(1, N = 37) = 22.73, p < 0.05$, $\chi^2(1, N = 41) = 16.095, p < 0.05$, and $\chi^2(1, N = 24) = 16.667, p < 0.05$, respectively.

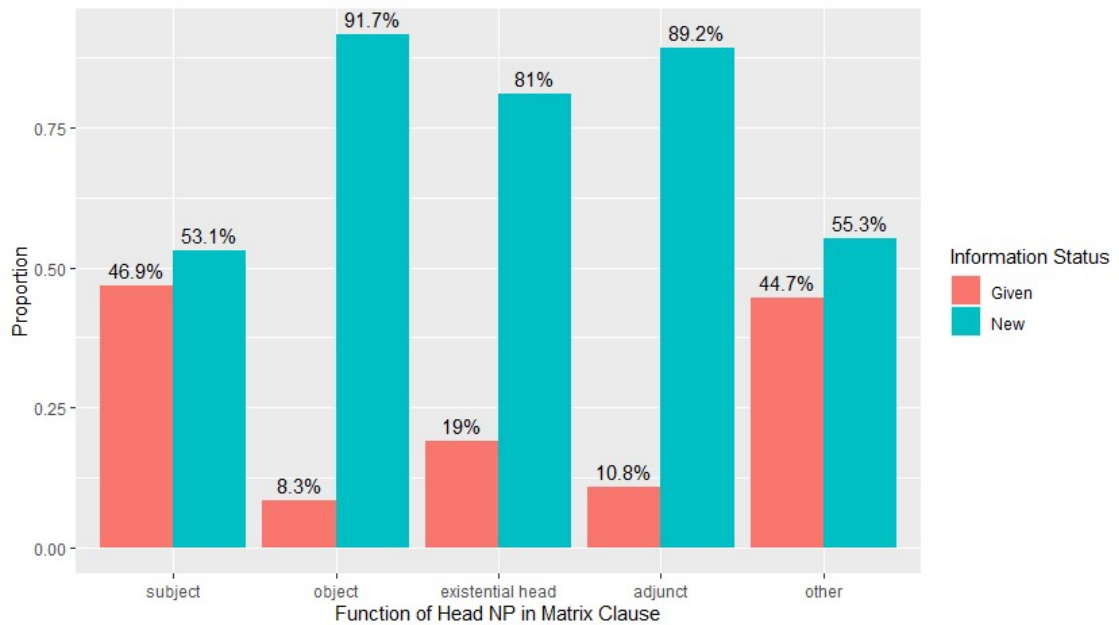


Figure 18. The distribution of INFORMATION STATUS OF HEAD NP across FUNCTION OF HEAD NP IN MATRIX CLAUSE

3.2.6. HUMANNESS & FUNCTION OF HEAD NP IN MATRIX CLAUSE

Figure 19 below visualizes the relationship between HUMANNESS and FUNCTION OF HEAD NP IN MATRIX CLAUSE. This distribution of the observed values of HUMANNESS across FUNCTION OF HEAD NP IN MATRIX CLAUSE is significantly different from the expected values under a chi-squared test at $\chi^2(4, N = 214) = 16.53, p = 0.002$. While adjunct, existential themes, and objects tend to be non-human¹⁶, there is no statistically significant tendency for subjects to be either human or non-human.

¹⁶ These figures are significant at $\chi^2(1, N = 37) = 22.73, p < 0.05$, $\chi^2(1, N = 41) = 7.7143, p = 0.005$, and $\chi^2(1, N = 24) = 8.1667, p = 0.004$, respectively.

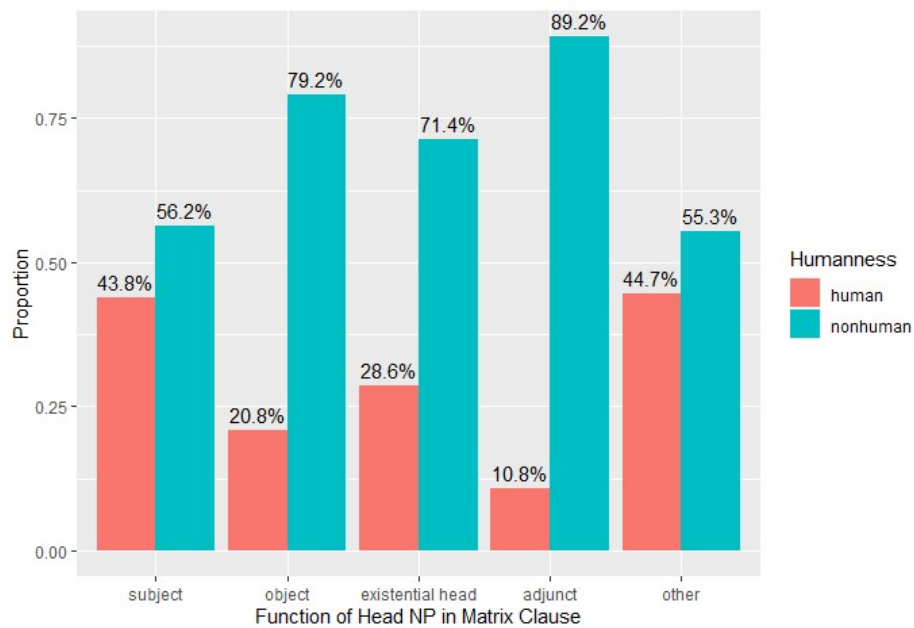


Figure 19. The distribution of HUMANNES across FUNCTION OF HEAD NP IN MATRIX CLAUSE

3.2.7. FUNCTION OF HEAD NP IN MATRIX CLAUSE & POSITION

This distribution of the observed values of POSITION across FUNCTION OF HEAD NP IN MATRIX CLAUSE is significantly different from the expected values under a chi-squared test at $\chi^2(8, N = 214) = 31.165, p < .05$. The Cramér's V score for the relationship between FUNCTION OF HEAD NP IN MATRIX CLAUSE and POSITION is 0.27, which signifies a weak association. The bar plot below in Figure 20 illustrates the proportions of headless, post-nominal, and pre-nominal RCs across the matrix roles. It suggests that, among all roles, adjunct heads behave most differently, tending not to occur with headless RCs; in addition, there were no data points for adjunct heads occurring with post-nominal RCs. In other words, an utterance like (54) is preferred over (55); constructed utterances like in (56) were not found in the data.

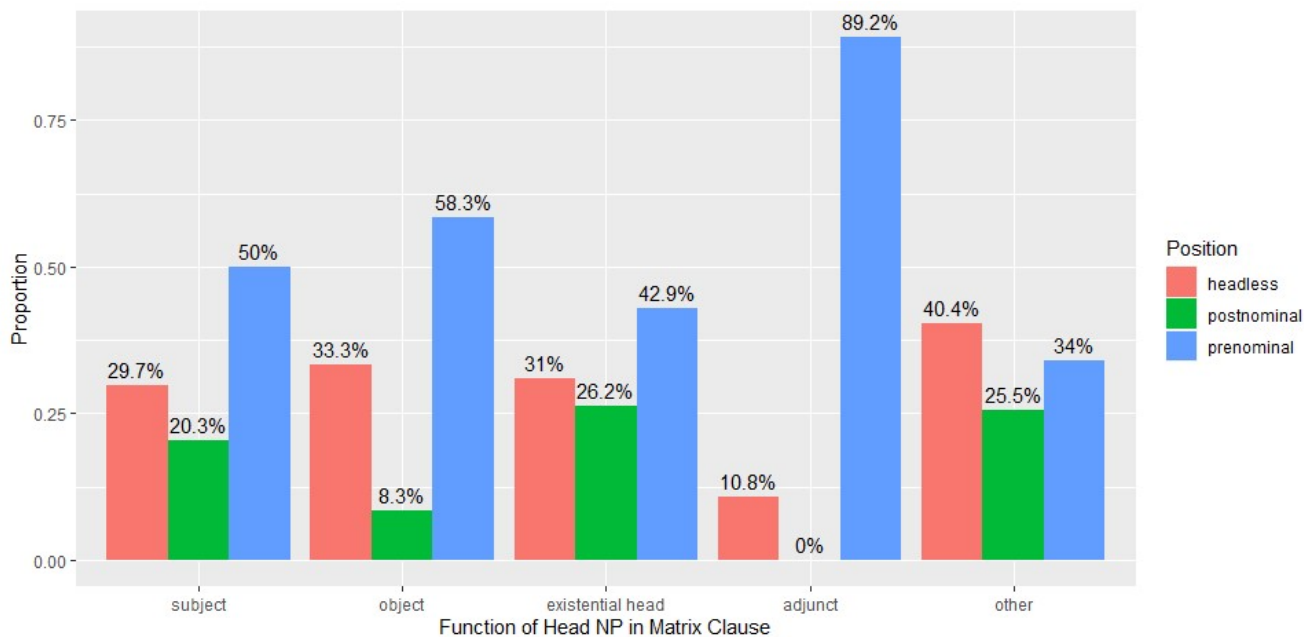


Figure 20. The distribution of FUNCTIONS OF HEAD NP IN MATRIX CLAUSE across POSITION

- (54) [adjunct RC] adjunct head
 [Aldında, Dīdar apay-men kel-gen] kez-de al-ğan-bız.
 a.while.ago PN sister-COM come-PST time-LOC buy-PTCP-1PL
 ‘We bought (it) at the time [when we came (here) with aunt Didar a while ago].’
- (55) [adjunct RC] (adjunct head)
 [o-dan keyin-gi]-si-nde, ... ber-di-k=aw de-y-min
 3SG-ABL after-ADJZ-3.POSS-LOC give-PST-1PL=maybe say-PRS-1SG
 ‘I think we ordered in the (time) [which was after that time].’
- (56) adjunct head [adjunct RC] (constructed example)
Taw-ğa bar-ayın dep jatır. [Qaraqat ös-etin]
 mountain-DAT go-1SG.HORT saying AUX currant grow-PTCP
 ‘(He/she) is planning to go to the mountains [where currants grow].’

When looked at individually, subject and object heads showed significant associations with POSITION at $\chi^2(2, N = 64) = 8.8438, p = 0.01$ and $\chi^2(2, N = 24) = 9, p = 0.01$, both of them tending to occur with pre-nominal RCs more frequently. In other words, examples (2) and (4), repeated below as (57) and (58), occur more frequently than (59) and (60).

- (57) [object RC] subject head
 [Kim-der, ayt-qan] prikol-dar-ı, ne et-pe-ytin bol-dı,
 who-PL say-PTCP joke-PL-3.POSS what do-NEG-PTCP be-PST
 öt-pe-ytin bol-dı
 pass-NEG-PTCP be-PST
 ‘The jokes [that those people said] will no longer be relevant.’
- (58) [subject RC] object head
 [elıw mıñ dannie bar,] kod-tı jiber-e-di,
 fifty thousand data EXST code-ACC send-PRS-3
 ‘(They) send code [that has fifty thousand data].’
- (59) subject head [object RC]
 endi ne zat anaw. [Qara-p otır-ğan,]
 then what thing that watch-CVB sit-PTCP
 ‘Then what that is that thing [that you are watching]?’
- (60) object head [subject RC]
köp adam-dar-dı, vrode kör-di-m, [kotoryie real’no,
 many person-PL-ACC I.guess see-PST-1SG REL.NOM.PL really
 mamandıq-tar-ı-n, ot i do awıstır-ıp jat-qan]=şe
 profession-PL-3.POSS-ACC from and to change-CVB AUX-PTCP=EMPH
 ‘I, like, saw a lot of people [who are really changing their professions entirely], you know.’

3.2.8. FUNCTION OF HEAD NP IN THE MATRIX CLAUSE & FUNCTION OF RC

Figure 21 depicts the relationship between these two variables. Overall, it is clear that all matrix roles tend to be modified by restrictive RCs. In fact, in the data, restrictive RCs (n = 195/214) outnumber non-restrictive RCs by ten times (n = 19/214), indicating that the use of restrictive RCs is overwhelmingly more common than the use of non-restrictive RCs.

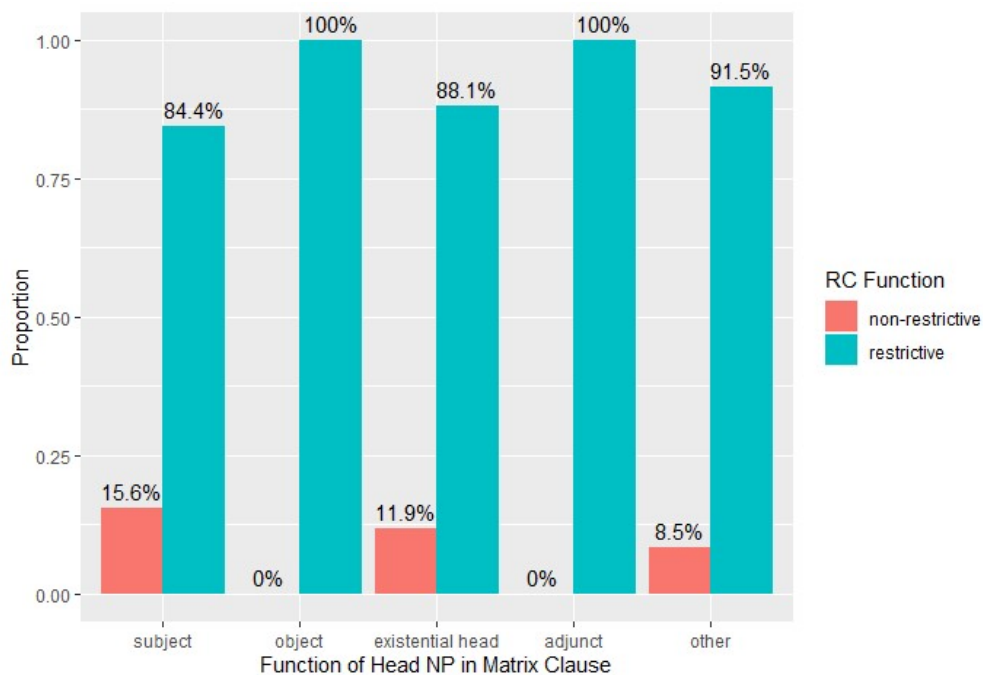


Figure 21. The distribution of FUNCTION OF RC across FUNCTION OF HEAD NP IN MATRIX CLAUSE

3.2.9. INFORMATION STATUS OF HEAD NP & FUNCTION OF RC

This relationship is visualized in Figure 22 below.

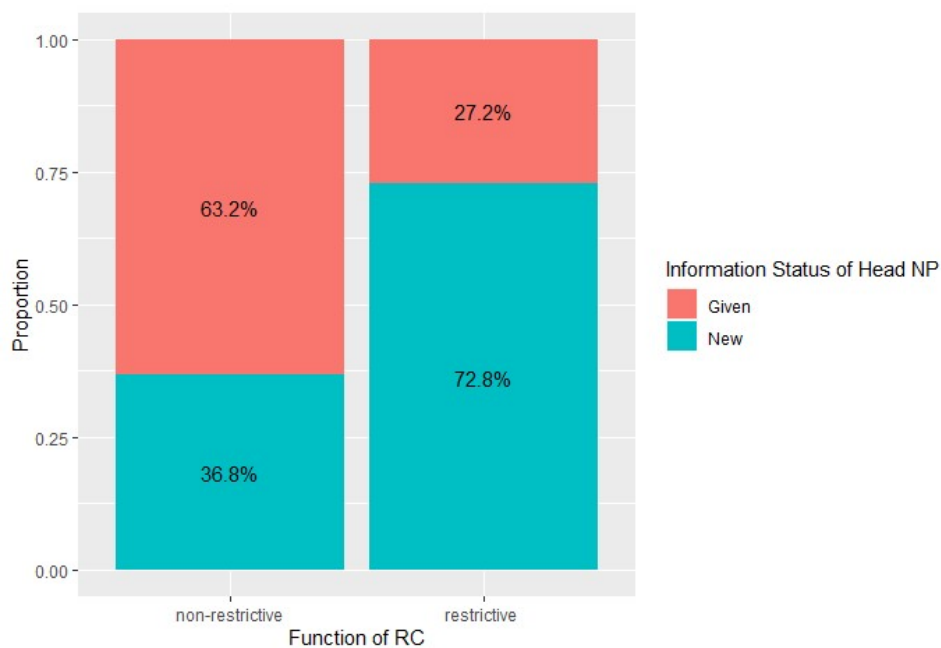


Figure 22. The distribution of INFORMATION STATUS OF HEAD NP across HUMANNES

This distribution of the observed values of INFORMATION STATUS OF HEAD NP across FUNCTION OF RC is significantly different from the expected values under a chi-squared test at $\chi^2(1, N = 214) = 8.9641, p = .002753$. It suggests that Given referents tend to be modified by RCs that are non-restrictive, while New referents tend to be modified by restrictive RCs.

3.2.10. POSITION & HUMANNES

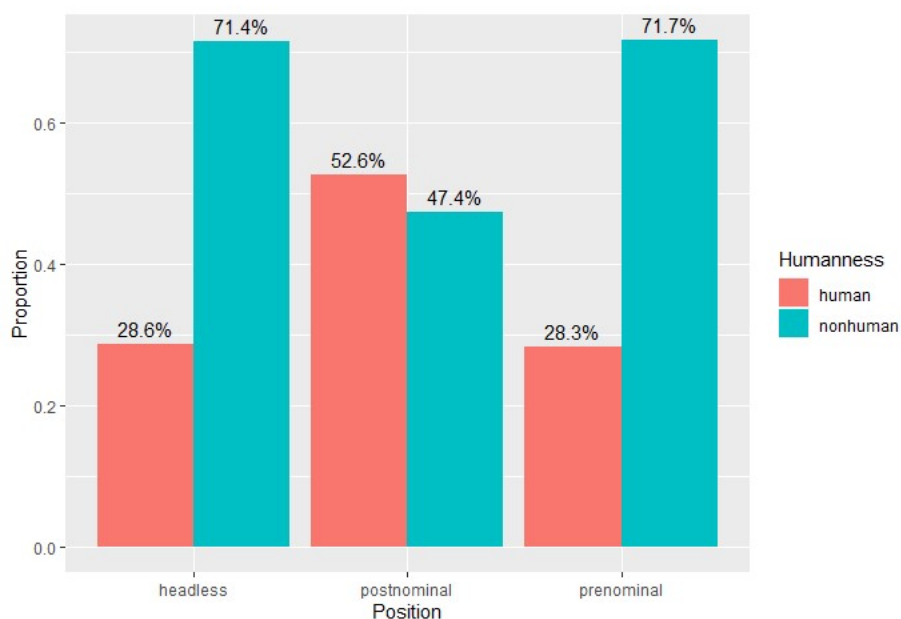


Figure 23. The distribution of HUMANNES across POSITION

This distribution of the observed values of POSITION across HUMANNES is significantly different from the expected values under a chi-squared test at $\chi^2(2, N = 214) = 8.332, p = .01551$. While headless and pre-nominal RCs tend to modify non-human referents, post-nominal RCs seem not to exhibit any strong preference for either human or non-human referents.

3.2.11. FUNCTION OF HEAD NP IN RC & POSITION

The relationship between these two variables is visualized in Figure 24 below.

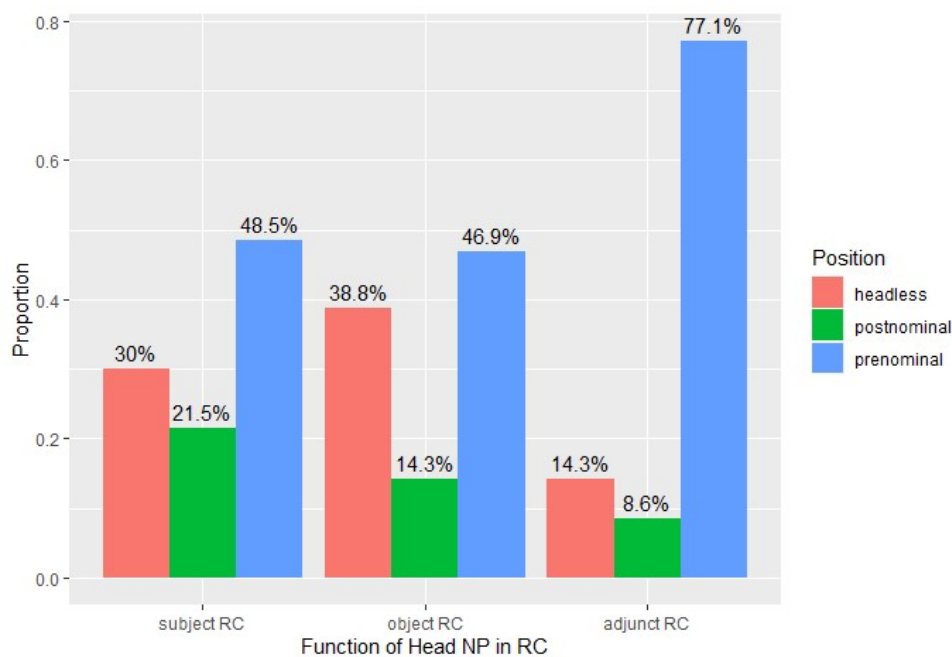


Figure 24. The distribution of FUNCTION OF HEAD IN RC across POSITION

This distribution of the observed values of POSITION across FUNCTION OF HEAD IN RC is significantly different from the expected values under a chi-squared test at $\chi^2(4, N = 214) = 11.947, p = .01775$. The association, however, is very weak, at Cramér's $V = 0.167$. It is clear, however, that every RC type tends to be pre-nominal.

3.2.12. INFORMATION STATUS OF HEAD NP & HUMANNES

The relationship between these variables is visualized in Figure 25. Overall, although the distribution of the observed values of HUMANNES across INFORMATION STATUS OF HEAD NP is significantly different from the expected values under a chi-squared test at $\chi^2(4, N = 214) = 4.5898, p = .03216$, it is very weak, at Cramér's $V = 0.157$. Both human and non-human referents tend to be New in the data; but the number of New human referents are 4 times greater than Given ones, while New non-human referents only 2 times greater than their Given counterparts.

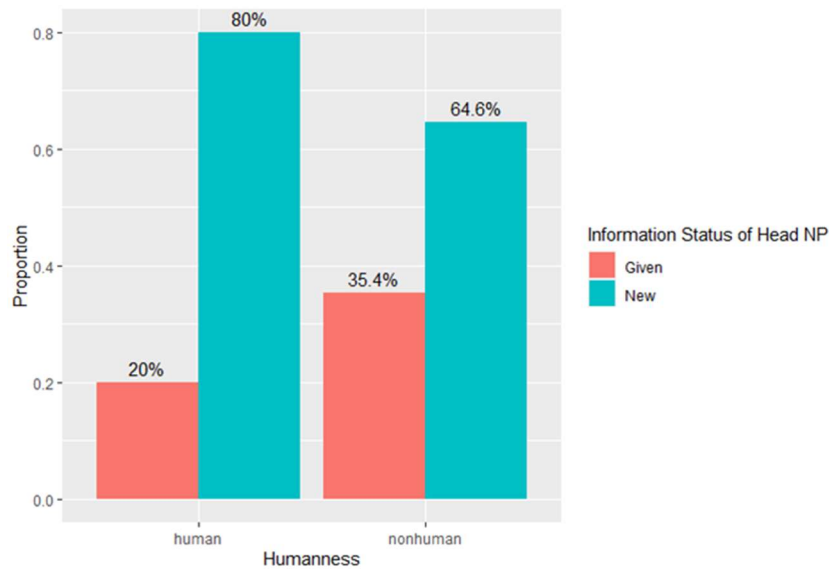


Figure 25. The distribution of INFORMATION STATUS OF HEAD NP across HUMANNES

3.2.13. FUNCTION OF HEAD NP IN RC & INFORMATION STATUS OF HEAD NP

The relationship between these variables is visualized in Figure 25.

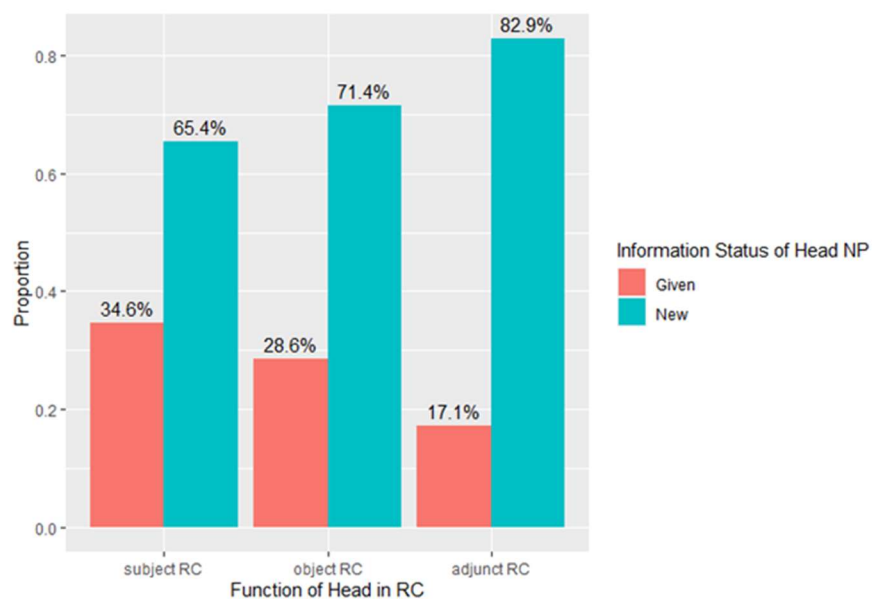


Figure 26. The distribution of INFORMATION STATUS OF HEAD NP across FUNCTION OF HEAD IN RC

This distribution of the observed values of INFORMATION STATUS OF HEAD NP across FUNCTION OF HEAD IN RC is not significantly different from the expected values under a chi-squared test. Hence, there is no meaningful relationship between these variables.

3.2.14. FUNCTION OF HEAD NP IN RC & FUNCTION OF RC

The relationship between these variables is visualized in Figure 26.

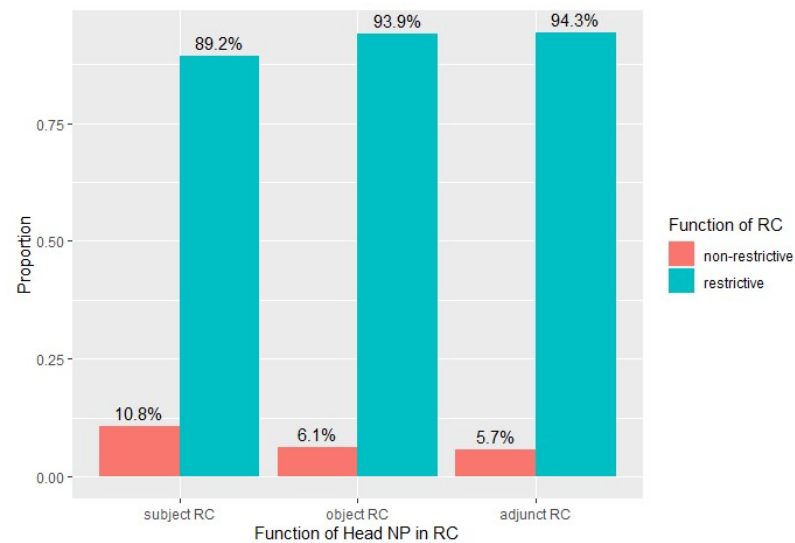


Figure 27. The distribution of FUNCTION OF HEAD NP IN RC across FUNCTION OF RC

This distribution of the observed values of FUNCTION OF HEAD NP IN RC across FUNCTION OF RC is not significantly different from the expected values under a chi-squared test. Hence, there is no meaningful relationship between these variables.

3.2.15. FUNCTION OF RC & HUMANNESS

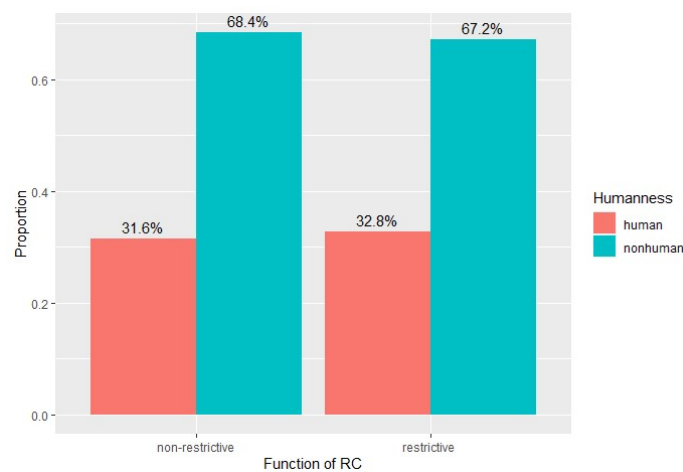


Figure 28. The distribution of HUMANNESS across FUNCTION OF RC

This distribution of the observed values of HUMANNES across FUNCTION OF RC is not significantly different from the expected values under a chi-squared test. Hence, there is no meaningful relationship between the two variables.

3.4. A note on the use of the Russian relative pronoun *kotor-* in Kazakh RCs

Out of 214 relative clauses in the dataset, only 8 contained a Russian relative pronoun *kotor-*. 5 out of 8 RCs contained the singular nominative form *kotoryi*, while the remaining three contained the plural nominal form *kotoryie*. Example (60) presented earlier in Section 3.2.7 is a relative clause that contains the plural nominative form *kotoryie*. It is repeated below as (61) where the relative pronoun is shown in bold.

- (61) köp adam-dar-dı, vrode kör-di-m, [**kotoryie** real'no,
many person-PL-ACC I.guess see-PST-1SG REL.NOM.PL really
mamandıq-tar-ı-n, ot i do awıstır-ıp jat-qan]=şe
profession-PL-3.POSS-ACC from and to change-CVB AUX-PTCP=EMPH

‘I, like, saw a lot of people [who are really changing their professions entirely], you know.’

Since the dataset contains a very low number of this relative pronoun, a statistical analysis would not be able to explain its distribution adequately; therefore, I did not treat the presence/absence of this pronoun as a separate variable in the statistical analyses. Future research with data containing more instances of this relative pronoun can shed light on its functions in Kazakh relative clauses.

3.5. Summary of results

Overall, there are clearly cognitive and discourse pressures at play in naturally occurring Kazakh conversations that yield meaningful skews in the distribution of relative clauses. First, the tendency of subjects to be human and topical (DuBois 1987; Givón 1983) yields a considerable number of relative clauses that modify subject Head NPs. The same

discourse tendency interacts with the cognitive constraint of MARKEDNESS (Givón 1991) to yield a high proportion of subject RCs. Thus, a human referent is highly likely to be used in subject head & subject RC combinations. Second, the joint influence of the semantic factor of HUMANNESSESS and the cognitive factor of INFORMATION STATUS leads to frequent HEAD-RC configurations: non-human referents are highly likely to be used in subject head & object RC and subject head & adjunct RC combinations. This is explained by the tendency of non-human referents in conversations to be linked to human referents who possess, manipulate, or exert some action on them (Du Bois 1980: 269-270). Third, most adjunct heads are temporal nouns, and this semantic property naturally yields the predominance of adjunct head & adjunct RC combinations. Fourth, POSITION of the relative clause is largely determined by the joint influence of FUNCTION and INFORMATION STATUS OF HEAD NP. If the RC is restrictive, it is likely to be formulated as either headless or pre-nominal, with headless RCs tending to modify Given referents and pre-nominal RCs – New referents. If the RC is non-restrictive, it is highly likely to be formulated as post-nominal due to the cognitive separation that is demanded by the juxtaposition of the New information in the non-restrictive RC and the modified referent (Ariel 1991: 152). Finally, a small portion of the dataset shows that the Russian relative pronoun *kotor-* is an additional means available to Kazakh speakers for forming relative clause constructions, and its low frequency in the data highlights the necessity for further comprehensive research into its distribution with more data.

4. Conclusion

The goal of this thesis was twofold: first, to corroborate the importance of studying grammar in use within naturally occurring conversational interactions, and second, to contribute to the grammatical description of Kazakh-in-use.

Serving as a universal social canvas, conversations embody the very essence of everyday social interactions (Schegloff 2015). They are a medium for language acquisition

(Clark & Casillas 2015) and serve as a crucible where linguistic evolution takes shape (Du Bois 2003; Chafe 1994). As such, I subscribe to the theoretical position that grammar is a crystallization of recurrent organic behavior as much as it is a resource for the conduct of such behavior (Ochs, Schegloff & Thompson 1996: 38). Taking relative clauses as the primary object of my research, I attempted to link their syntactic behavior to the frequent choices speakers make when engaged in organic face-to-face conversational exchanges. The findings of my research indicate that the organic use of relative clauses by Kazakh speakers is constrained by a number of linguistic, cognitive, and discourse-driven factors that naturally arise within conversational exchanges, as was expected. As such, the paper has outlined the specific configurations of relative clause constructions observed in naturally occurring Kazakh data that were obtained by running statistical models on a set of variables that had previously been shown to affect the distribution of relative clauses. Namely, the factors involving HUMANNES, INFORMATION STATUS OF HEAD NP, and FUNCTION OF THE RELATIVE CLAUSE were shown to be the most important predictors of speakers' choice of specific relative clause constructions.

The grammar of Kazakh has been studied by linguists both in its spoken and written varieties (Balakaev 1959; 1962; Amanzholov 1994; Zhanpeisov 2002; Zholshayeva 2016; Muhamedowa 2016). However, grammatical descriptions of spoken Kazakh have been either based on introspection or non-naturally occurring data such as fieldwork interviews (Muhamedowa 2009; 2005). As such, this thesis aimed to fill this gap by utilizing the existing corpus of naturally occurring Kazakh discourse called Multimedia Corpus of Modern Spoken Kazakh (Filchenko, Troiani, Du Bois & Sarseke et al. 2023). Relative clauses have been claimed to exhibit exactly the same syntactic behavior in both spoken and written Kazakh (Muhamedowa 2016: 39). The findings of my work challenge this claim in light of conversational evidence. Particularly, the use of non-finite post-nominal relative clauses, a

feature not typical of Turkic languages according to grammars, has been shown to fulfil an important interactional function of non-restriction whereby it gives the hearer “an added piece of information about an already identified entity, but [does not] identify that entity” (Comrie 1989: 138). Pre-nominal and headless relative clauses, which have been attested in the grammars of Kazakh, are used to help single out or identify a referent from potential members of a class (Comrie 1989:138). It was also shown that, in Kazakh conversations, pre-nominal relative clauses tend to modify New referents, while headless relative clauses – Given referents, thus showing that the cognitive factor of INFORMATION STATUS also plays a role in further skewing these distributions. Additionally, a small fraction of the data suggests that the Russian relative pronoun *kotor-* also serves as a means to form relative clauses in Kazakh conversations, a topic that requires further research. Thus, the ‘grammar’ of relative clauses, as used by everyday conversationalists, emerges as a dynamic phenomenon, an observation that could not have been made with non-conversational data.

The issue of so-called ‘spoken grammars’ and their important role in foreign language instruction, particularly in the context of English, has been the subject of much discussion in recent years (Carter & McCarthy 2017). If it is held that grammars of Kazakh as a foreign language needs to be “paradigmatically adapted for learners” given the current expansion of its scope of use (Naraliyeva et al. 2015: 347), then linguists need to complement modern grammars of Kazakh with empirically derived grammatical descriptions. As such, research informed by naturally occurring language use will not only contribute to a holistic representation of Kazakh but also to an effective teaching and learning of the language. The work of ‘spoken grammar’ researchers focus on three main aspects of language (Carter & McCarthy 2017: 5): (1) phenomena that are more frequent or differently distributed in speech compared to writing; (2) aspects of language use often overlooked due to an emphasis on written language; (3) the conditions under which these phenomena occur and the mechanisms

by which they illuminate fact-to-face interactions. Thus, future research in this direction for Kazakh should start from the analysis of phenomena in spoken Kazakh that the current grammars fail to account for by utilizing all the existing resources such as the Multimedia Corpus of Modern Spoken Kazakh (Filchenko, Troiani, Du Bois & Sarseke et al. 2023) and the works of conversation-oriented researchers.

5. References

- Amanzholov, Sarsen. 1950. *Qazaq ädebî tili sintaksisiniñ qısqaşa kwrsı* [A Short Course on the Syntax of Literary Kazakh]. Almaty.
- Andrews, Avery D. 2007. Relative clauses. In: Timothy Shopen (ed.), *Language Typology and Syntactic Description: Complex Constructions*. 206–236. Cambridge: Cambridge University Press.
- Ariel, Mira. 1990. *Accessing Noun-Phrase Antecedents*. London: Routledge
- Ariel, Mira. 2009. Discourse, grammar, discourse. *Discourse Studies* 11 (1): 5-36.
- Balakaev, Maulen. 1959. *Sovremennyy kazahskij jazyk. Sitaksi* [Modern Kazakh Language: Syntax]. Alma-Ata: Izd-vo Akad. Nauk Kaz. SSR.
- Balakaev, Maulen. 1962. *Sovremennyy kazahskij jazyk. Fonetika i morfologija* [Modern Kazakh Language: Phonetics and Morphology]. Alma-Ata: Izd-vo Akad. Nauk Kaz. SSR.
- Bazarbayeva, Zeinep, Sholpan Zharkynbekova, Aisaule Amanbayeva, Zhanar Zhumabayeva, & Ainur Karshygayeva. 2023. The National Corpus of Kazakh Language: Development of Phonetic and Prosodic Markers. *Zhurnal Sibirskogo federal'nogo universiteta. Gumanitarnye nauki* 16 (8): 1256-1270.
- Bloomfield, Leonard. 1935. *Language*. London: Allen & Unwin
- Bureau of National Statistics of the Agency for Strategic Planning and Reforms of the Republic of Kazakhstan. 2021. *2021 population census* [Data file]. Retrieved from <https://stat.gov.kz/upload/medialibrary/e62/b1e0sokkht34a1iyu2qdmu30dayt6sz1/Краткие%20итоги%20Переписи%20населения.pdf>
- Carter, Ronald & Michael McCarthy. Spoken Grammar: Where Are We and Where Are We Going?, *Applied Linguistics* 38(1): 1–20.
- Chafe, Wallace. 1987. Cognitive constraints on Information Flow. In: Russel Tomlin (ed.), *Coherence and Grounding in Discourse*. 21-51. Amsterdam: John Benjamins Publishing Company.
- Chafe, Wallace. 1994. *Discourse, consciousness, and time: The flow and displacement of conscious experience in speaking and writing*. Chicago: University of Chicago Press.
- Comrie, Bernard. 1989. *Language Universals and Liguistic Typology: Second Edition*. Chicago: The University of Chicago Press.

- Chomsky, Noam. 1965. *Aspects of the Theory of Syntax*. Massachusetts: MIT Press
- Clift, Rebecca. 2007. Grammar in Time: The Non-Restrictive Which-Clause as an Interactional Resource. *Essex Research Reports in Linguistics* 55: 51–82. Essex: University of Essex.
- Collier-Sanuki, Yoko. 1990. Relative clauses and discourse strategies. In: Soonjia Choi (ed.), *Japanese/Korean Linguistics*. 3:54–66. San Diego: San Diego State University.
- Couper-Kuhlen, Elizabeth, & Margret Selting. 2017. *Interactional linguistics: Studying language in social interaction*. Cambridge: Cambridge University Press.
- Çynar, Oktay. 2021. On the typology of the null subject parameter: A proposal on Turkish. In: Alper Kumcu & Ayşe Selmin Söylemez (eds.), *Synergy II: Linguistics: Contemporary Studies on Turkish Linguistics*. 59-76. Berlin: Peter Lang Publishing
- Dahl, Osten, and Kari Fraurud. 1996. Animacy in grammar and discourse. *Pragmatics and Beyond New Series*: 47-64.
- de Saussure, Ferdinand. 1916. *Course in General Linguistics*. London: Duckworth
- Du Bois, John. 1980. Beyond definiteness: The trace of identity in discourse. In: Wallace Chafe (ed.), *The Pear Stories: Cognitive, Cultural, and Linguistic Aspects of Narrative Production*. 203-274. Westport: Praeger.
- Du Bois, John. 1983. Outline of Discourse Transcription. In Jane A. Edwards & Martin D. Lampert (eds), *Talking Data: Transcription and Coding in Discourse Research*: 45-89. London: Routledge
- Du Bois, John. 1987. The discourse basis of ergativity. *Language* 63 (4): 805-855.
- Du Bois, John. 2003. Discourse and Grammar. In: Micheal Tomasello (ed.), *The New Psychology of Language: Cognitive and Functional Approaches to Language Structure, Volume 2*. 47-88. New Jersey: Lawrence Erlbaum Associates.
- Du Bois, John W., Wallace L. Chafe, Charles Meyer, Sandra A. Thompson, Robert Englebretson, and Nii Martey. 2000-2005. *Santa Barbara corpus of spoken American English, Parts 1-4*. Philadelphia: Linguistic Data Consortium.
- Enç, Mürvet. 1986. Topic switching and pronominal subjects in Turkish. In: Dan I. Slobin & Karl Zimmer (eds.), *Studies in Turkish Linguistics*. 195-209. Amsterdam: John Benjamins Publishing Company.
- Fairclough, Norman. 1993. Critical discourse analysis and the marketization of public discourse: The universities. *Discourse & Society* 4 (2): 133-168.
- Filchenko, Andrey, Giorgia Troiani, John W. Du Bois, Gulnar Sarseke, Akyl Akanov, Moldir Bizhanova, Nikolay Mikhailov, Tansulu Temirbekova, Bibarys Seitak, and Zhansaya Turaliyeva. 2023. *Multimedia Corpus of Spoken Kazakh Language*. Astana: Nazarbayev University.
- Ford, Cecilia E., & Sandra A. Thompson. 1996. Interactional units in conversation: Syntactic, intonational, and pragmatic resources for the management of turns. In: Elinor Ochs, Emanuel A. Schegloff & Sandra A. Thomopson (eds.), *Interaction and grammar*. 134-184. Cambridge: Cambridge University Press

- Fox, Barbara, & Sandra A. Thompson. 1990. A discourse explanation of the grammar of relative clauses in English conversation. *Language* 66 (2): 297-316.
- Fox, Barbara, & Sandra A. Thompson. 2007. Relative clauses in English conversation: relativizers, frequency, and the notion of construction. *Studies in Language* 31 (2): 293-326.
- Frazier, Lyn. 1987. Syntactic processing: Evidence from Dutch. *Natural Language and Linguistic Theory* 5 (4): 519-559.
- Givón, Thomas. 1979. *On Understanding Grammar*. New York: Academic Press
- Givón, Thomas. 1983. Topic continuity and word-order pragmatics in Ute. In: Thomas Givón (ed.), *Topic Continuity in Discourse: A Quantitative Cross-Language Study*. 141-214. Amsterdam: John Benjamins Publishing Company.
- Givón, Thomas. 1991. Markedness in Grammar: Distributional, communicative and cognitive correlates of syntactic Structure. *Studies in Language* 15 (2): 335-375.
- Givón, Thomas. 1993. *English grammar: A function-based introduction*. Amsterdam: John Benjamins Publishing Company.
- Günthner, Susanne. 2011a. Between Emergence and Sedimentation: Projecting Constructions in German Interactions. In: Peter Auer & Stefan Pfänder (eds.), *Constructions: Emerging and Emergent*. 156-185. Berlin: De Gruyter.
- Günthner, Susanne. 2011b. N Be That-Constructions in Everyday German Conversation. In: Ritva Laury & Ryoko Suzuki (eds.), *Subordination in Conversation: A Cross-Linguistic Perspective*. 11-36. Amsterdam: John Benjamins Publishing Company.
- Halliday, M. A. K. & Christian M. I. M. Matthiessen. 2013. *Halliday's Introduction to Functional Grammar: Fourth Edition*. London: Routledge.
- Higashiizumi, Yuko. 2011. Are Kara 'Because'-Clauses Causal Subordinate Clauses in Present-Day Japanese? In: Ritva Laury & Ryoko Suzuki (eds.), *Subordination in Conversation: A Cross-Linguistic Perspective*. 191-207. Amsterdam: John Benjamins Publishing Company.
- Hopper, Paul, and Sandra Thompson. 2008. Projectability and Clause Combining in Interaction. In: Riva Laury (ed.), *Crosslinguistic Studies of Clause Combining: The Multifunctionality of Conjunctions*, 99-123. Amsterdam: John Benjamins Publishing Company.
- Hwang, Shin Ja J. 1994. Relative clauses, adverbial Clauses, and Information Flow in discourse. *Language Research* 30 (4): 673-705.
- Imo, Wolfgang. 2011. Clines of Subordination – Constructions with the German 'Complement-Taking Predicate' *Glauben*. In: Ritva Laury & Ryoko Suzuki (eds.), *Subordination in Conversation: A Cross-Linguistic Perspective*. 165-190. Amsterdam: John Benjamins Publishing Company.
- Johanson, Lars. 2021. *Turkic*. Cambridge: Cambridge University Press.
- Keenan, Edward L. 1985. Relative clauses. In: Timothy Shopen (ed.), *Language Typology and Syntactic Description: Complex Constructions*. 141–170. Cambridge: Cambridge University Press.

- Keevallik, Leelo. 2011. Interrogative ‘Complements’ and Question Design in Estonian. In: Ritva Laury & Ryoko Suzuki (eds.), *Subordination in Conversation: A Cross-Linguistic Perspective*. 37-68. Amsterdam: John Benjamins Publishing Company.
- Kerslake, Celtek. 1987. Noun phrase deletion and pronominalization in Turkish. In: Ludo Verhoeven & Erik Boeschoten (eds.), *Studies on Modern Turkish*, 91–104. Tilburg: Tilburg University Press.
- Kim, Alan Hyun-Oak & Hyon-Sook Shin. 1994. Information Flow and relative-clause constructions in Korean discourse.” In: Young-Key Kim-Renaud (ed.), *Theoretical Issues in Korean Linguistics*. 463–494. Chicago: Center for the Study of Language.
- Koivisto, Aino, Ritva Laury, and Eeva-Leena Seppänen. 2011. Syntactic and Actional Characteristics of Finnish Että-Clauses. In: Ritva Laury & Ryoko Suzuki (eds.), *Subordination in Conversation: A Cross-Linguistic Perspective*. 69-102. Amsterdam: John Benjamins Publishing Company.
- Kwon, Na-Young, Maria Polinsky, & Robert Kluender. 2006. Subject preference in Korean. In: Donald Baumer (ed.), *Proceedings of the West Coast Conference on Formal Linguistics 25*: 1–14.
- Lambrecht, Knud. 1994. *Information Structure and Sentence Form: Topic, focus, and the mental representations of discourse referents*. Cambridge: Cambridge University Press.
- Laury, Ritva, and Ryoko Suzuki. 2011. *Subordination in Conversation: A Cross-Linguistic Perspective*. Amsterdam: John Benjamins Publishing Company.
- Laury, Ritva, and Shigeko Okamoto. 2011. Teyuuka and I Mean as Pragmatic Parentheticals in Japanese and English. In: Ritva Laury & Ryoko Suzuki (eds.), *Subordination in Conversation: A Cross-Linguistic Perspective*. 209-238. Amsterdam: John Benjamins Publishing Company.
- Levinson, Stephen C. 2012. Action formation and ascription. In: Jack Sidnell & Tanya Stivers (eds.), *The handbook of conversation analysis*. 101-130. New Jersey: John Wiley & Sons
- Levshina, Natalia. 2015. *How to do Linguistics with R*. Amsterdam: John Benjamins Publishing Company.
- Lin, Chien-er Charles. 2006. *Grammar and Parsing: A Typological Investigation of Relative-Clause Processing*. Arizona: University of Arizona. Doctoral Dissertation.
- Linell, Per. 2004. *The Written Language Bias in Linguistics: Its Nature, Origins and Transformations*. London: Routledge.
- MacWhinney, Brian & Csaba Pléh. 1988. The processing of restrictive relative clauses in Hungarian. *Cognition* 29 (2): 95-141.
- Makhambetov, Olzhas, Aibek Makazhanov, Zhandos Yessenbayev, Bakhyt Matkarimov, Islam Sabyrgaliyev, & Anuar Sharafudinov. 2013. Assembling the kazakh language corpus. In David Yarowsky, Timothy Baldwin, Anna Korhonen, Karen Livescu & Steven Bethard (eds.), *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*: 1022-1031.
- Manz, Beatrice. 2018 *Central Asia in historical perspective*. London: Routledge

- Mecklinger, A., H. Schriefers, K. Steinhauer, & Angela D. Friederici. 1995. Processing relative clauses varying on syntactic and semantic dimensions: An analysis with event-related potentials. *Memory & Cognition* 23 (4): 477-494.
- Mithun, Marianne. 2015. Discourse and Grammar. In: Deborah Tannen, Deborah Schiffrin, & Heidi Hamilton (eds.), *The Handbook of Discourse Analysis*. 11-41. New Jersey: Wiley-Blackwell
- Miyamoto, Edson, & Michiko Nakamura. 2003. Subject/Object asymmetries in the processing of relative clauses in Japanese." In: Gina Garding & Mimu Tsujimura (eds.), *Proceedings of the West Coast Conference on Formal Linguistics*. 342–355. Burnaby: Simon Fraser University.
- Muhamedowa, Raihan. 2005. Kasachisch-Russisches Code-Mixing: Ein Fall von Morphologischer Vereinfachung [Kazakh-Russian Code Mixing: A Case of Morphological Simplification]. *Zeitschrift Fur Sprachwissenschaft* 24 (2): 263-309.
- Muhamedowa, Raihan. 2009. The use of Russian conjunctions in the speech of bilingual Kazakhs. *International Journal of Bilingualism* 13 (3): 331-356.
- Muhamedowa, Raihan. 2016. *Kazakh: A Comprehensive Grammar*. London: Routledge
- Mussakhojayeva, Saida, Yerbolat Khassanov, & Huseyin Atakan Varol. 2022. KSC2: An Industrial-Scale Open-Source Kazakh Speech Corpus. *INTERSPEECH*: 1367-1371.
- Naraliyeva, Rakhila, Laura Mukhanbekkyzy, Maira Toiganvekova, Balgabay Dosanov, & Bybygul Sultanova. 2015. Modern methods of teaching Kazakh as a foreign language: Search, Innovation, Quality, Result. *Review of European Studies* 7 (7): 347.
- Ochs, Elinor, Emanuel A. Schegloff, & Sandra A. 1996. Introduction. In: Elinor Ochs, Emanuel A. Schegloff & Sandra A. Thomopson (eds.), *Interaction and grammar*. 1-52. Cambridge: Cambridge University Press.
- Ótót-Kovács, Eszter. 2015. The syntax of non-finite clauses in Kazakh. Szeged: University of Szeged. Doctoral dissertation.
- Özsoy, Sumru. 1987. Null subject parameter in Turkish. In: Ludo Verhoeven & Boeschoten Erik (eds.), *Studies on Modern Turkish*. 82–90. Tilburg: Tilburg University Press.
- Pekarek Doehler, Simona. 2011. Clause-Combining and the Sequencing of Actions: Projector Constructions in French Talk-in-Interaction. In: Ritva Laury & Ryoko Suzuki (eds.), *Subordination in Conversation: A Cross-Linguistic Perspective*. 103-148. Amsterdam: John Benjamins Publishing Company.
- Posit team. 2024. RStudio: Integrated Development Environment for R. Posit Software, PBC. <http://www.posit.co/>. MA: Boston
- Prideaux, Gary D. 1982. The processing of Japanese relative clauses. *Canadienne de Linguistique* 27 (1): 23–30.
- Pu, Ming Ming. 2007. The distribution of relative clauses in Chinese discourse. *Discourse Processes* 43 (1): 25–53.
- Schroeder, Christoph. 1999. *The Turkish Nominal Phrase in Spoken Discourse*. Wiesbaden: Harrassowitz Verlag.

- Schegloff, Emanuel A. 2015. Conversational interaction: The embodiment of human sociality. In: Deborah Tannen, Heidi E. Hamilton, & Deborah Schiffrin (eds), *The handbook of discourse analysis*. 346-366. New Jersey: Wiley-Blackwell.
- Sinor, Denis. 1995. Languages and cultural interchange along the Silk Roads. *Diogenes* 43 (171): 1-13.
- Slobin, Dan. 1986. The acquisition and use of relative clauses in Turkish and Indo-European languages." In: Dan Slobin & Karl Zimmer (eds.), *Studies in Turkic Linguistics*. 273-297. Amsterdam: John Benjamins Publishing Company.
- Smagulova, Juldyz. 2006. Kazakhstan: Language, identity and conflict. *Innovation: The European Journal of Social Science Research* 19 (3-4): 303-320.
- Suzuki, Ryoko. 2011. A Note on the Emergence of Quotative Constructions in Japanese Conversation. In: Ritva Laury & Ryoko Suzuki (eds.), *Subordination in Conversation: A Cross-Linguistic Perspective*. 149-165. Amsterdam: John Benjamins Publishing Company.
- Tagliamonte, Sali A., & Harald Baayen. 2012. Models, forests, and trees of York English: Was/were variation as a case study for statistical practice. *Language variation and change* 24, (2): 135-178.
- Taleghani-Nikazm, Carmen. 2006. *Request sequences: The intersection of grammar, interaction and social context*. Amsterdam: John Benjamins Publishing.
- Tao, Hongyin. 2002. Hànyǔ kǒuyǔ xùshì tǐ guānxì cóngjù jiégòu de yǔyì hé piānzhāng shǔxìng [Semantic and discourse properties of relative clauses in Chinese oral narratives]. *Contemporary Research in Modern Chinese* 4: 47-57.
- Tao, Hongyin, and Michael J McCarthy. 2001. Understanding non-restrictive which-clauses in spoken English, which is not an easy thing. *Language Sciences* 23 (6): 651-77.
- Taylan, Eser. 1986. Pronominal vs. zero representation of anaphora in Turkish. In: Dan Slobin & Karl Zimmer, *Studies in Turkish Linguistics*. 209-232. Amsterdam: John Benjamins Publishing Company.
- Thompson, Sandra A. 1997. Discourse motivations for the core-oblique distinction as a language universal. In: Akio Kamio (ed.), *Directions in Functional Linguistics*. 59-82. Amsterdam: John Benjamins Publishing Company.
- Thompson, Sandra A. 2002. Object Complements and Conversation towards a Realistic Account. *Studies in Language* 26 (1): 125-164.
- Troiani, Giorgia. 2023. Representing a language in use: corpus construction, prosody, and grammar in Kazakh. Santa-Barbara: University of California. Doctoral Dissertation.
- Troiani, Giorgia, John Du Bois, & Andrey Filchenko. (in press). Corpus as a slice of life: Representing naturally occurring language and its speakers. In: Robbie Love (ed.), *Research in Corpus Linguistics: Special Issue "Innovations in the Compilation and Analysis of Spoken Corpora"*.
- Vasishth, Shravan, Zhong Chen, Qiang Li, & Gueilan Guo. 2013. Processing Chinese relative clauses: Evidence for the subject-relative advantage. *PLoS ONE* 8 (10): e

- Wang, Fang, and Fuyun Wu. 2020. Post-nominal relative clauses in Chinese. *Linguistics* 58(6): 1501–1542.
- Zhanpeisov, Yerbol, Kobey Husain, Nurzhamal Oralbayeva, Seilbek Isaev, Aitbay Aigabyluly, Myrzatay Sergaliev, Rakysh Amir, and Nurgeldi Ualiev, eds. 2002. *Qazaq grammatikası: Fonetika. Sözjasam. Morfologiya. Sintaksis*. [The Grammar of Kazakh: Phonetics, Word Formation, Morphology, and Syntax]. Astana.
- Zholshayeva, Mayra. 2016. *Qazaq tili: funkcionaldy grammatika* [The Kazakh language: functional grammar]. Almaty: Print World.
- Zhubanov, Askar. 2009. Korpwstıq lingvıstika – qazaq tilbiliminin jaña bağıtı [Corpus Linguistics is the new direction in Kazakh linguistics]. *Tiltanym* 2 (34): 3-11. Almaty: QR BjĜM A.Baytursınılı atındağı Til bilimi instıtıwtı

6. Appendices

6.1. List of glossing abbreviations

ABL	ablative
ACC	accusative
ADJ	adjectival
AUX	auxiliary
CAUS	causative
COP	copula
COM	comitative
COMP	complementizer
CVB	converb
DAT	dative
EXST	existential
EMPH	emphatic
EVD	evidential
ENG	engagement marker
F	feminine
GEN	genitive
HORT	hortative
INF	infinitive
INST	instrumental
LOC	locative
M	masculine
NEG	negation marker
NOM	nominative
OPT	optative
POSS	possessive clitic
POL	polite
PN	proper name
PL	plural
PTCP	participle
PST	past
PROG	progressive

Q	interrogative
REFL	reflexive
REL	relativizer
SG	singular
1	first person
2	second person
3	third person

6.2. Discourse-Functional Transcription Conventions

Symbol	Name	Meaning
.	Final contour	intonation marks current action as complete
,	Continuing contour	intonation marks current action as incomplete
?	Rising contour	intonation marks a rising contour
—	Truncated contour	current intonation unit is interrupted
wor-	Truncated contour	word not completed as projected
@	Laughter	one symbol per pulse of laughter
#	Unintelligible	one symbol per unintelligible syllable
:	Prosodic lengthening	follows a lag in speech rate determined by prosody and not phonology
(H)	In-breath	audible inhalation
(Hx)	Out-breath	audible exhalation
¬	Pseudograph	name change for anonymity
(EVENT)	Non vocal event	sounds not produced by the speakers or not produced in the vocal tract

6.3. R code

```
#INTRO
#####
## Author: Akyl Akanov
## Last Update: 15.04.2024

## This code is written for analyzing RCs in Spoken Kazakh
## This code assumes the reader is familiar with the the linguistic terms used in the thesis

#load packages
library(ggplot2); library(rcompanion); library(vcd); library(party);
library(partykit); library(caret); library(randomForest); library(corrplot)

#1. DATA
# load the data
relclauses <- read.csv("relclauses.csv")

#turn categorical variables into factors
relclauses$Type <- as.factor(relclauses$Type) #FUNCTION OF HEAD NP IN RC
relclauses$Head <- as.factor(relclauses$Head) #FUNCTION OF HEAD NP IN MATRIX
relclauses$NPrel <- as.factor(relclauses$NPrel) #same as the variable FUNCTION OF HEAD NP IN RC
relclauses$Position <- as.factor(relclauses$Position)
relclauses$Humanness <- as.factor(relclauses$Humanness)
```

```

relclauses$Form <- as.factor(relclauses$Form) #the morphological form of the participle used
relclauses$InformationStatusofHead <- as.factor(relclauses$InformationStatusofHead)
relclauses$Function <- as.factor(relclauses$Function) #FUNCTION OF RC

#2. Frequency Distributions

#2.1. Frequency Distribution of FUNCTION OF HEAD NP IN MATRIX CLAUSE

tableHeads <- table(relclauses$Head)
propHeads <- prop.table(tableHeads)

#2.2. Frequency Distribution of FUNCTION OF HEAD NP IN RC
tableTypes <- table(relclauses$Type)
propTypes <- prop.table(tableTypes)

#a chi-squared test
chisq.test(tableTypes)

#3 A multivariate analysis of HEAD-RC COMBINATIONS and POSITION

#3.1. The analysis of HEAD-RC COMBINATIONS as the dependent variable

#create a data frame without the Head function of "Other" since I am focusing only on
#subject, object, existential theme, and adjunct

filtered_relclauses_withoutOther <- subset(relclauses, Head != "other")

#add a new variable HEAD-RC COMBINATIONS into the dataframe
filtered_relclauses_withoutOther$Head_Type <- paste(substr(filtered_relclauses_withoutOther$Head, 1, 1),
substr(filtered_relclauses_withoutOther$Type, 1, 1), sep = "-")

#turn categorical variables into factors
filtered_relclauses_withoutOther$Head_Type <- as.factor(filtered_relclauses_withoutOther$Head_Type)
filtered_relclauses_withoutOther$Type <- as.factor(filtered_relclauses_withoutOther$Type) #explained below
filtered_relclauses_withoutOther$Head <- as.factor(filtered_relclauses_withoutOther$Head) #grammatical function of
modified NP in main clause
filtered_relclauses_withoutOther$NPrel <- as.factor(filtered_relclauses_withoutOther$NPrel) #grammatical function of
modified NP in RC
filtered_relclauses_withoutOther$Position <- as.factor(filtered_relclauses_withoutOther$Position) #headless, pre-nominal,
post-nominal
filtered_relclauses_withoutOther$Humanness <- as.factor(filtered_relclauses_withoutOther$Humanness) #whether NP is
human/non-human
filtered_relclauses_withoutOther$Form <- as.factor(filtered_relclauses_withoutOther$Form) #the morphological form of the
participle used
filtered_relclauses_withoutOther$InformationStatusofHead <-
as.factor(filtered_relclauses_withoutOther$InformationStatusofHead) #whether NP is New or Given
filtered_relclauses_withoutOther$Function <- as.factor(filtered_relclauses_withoutOther$Function) #whether the RC is
restrictive or non-restrictive

#Run and plot a condition inference tree with HEAD-RC COMBINATIONS as the dependent variable and others as
independent variables

set.seed(124)
HeadRcCombo.ct <- ctree(Head_Type ~ InformationStatusofHead + Humanness + Position + Function, data =
filtered_relclauses_withoutOther)
plot(HeadRcCombo.ct, gp = gpar(fontsize = 11))

#Run a random forest with HEAD-RC COMBINATIONS as the dependent variable and others as independent variables

```



```

set.seed(365)
HeadRcCombo.cf <- cforest(Head_Type ~ InformationStatusofHead + Humanness + Position + Function,
  data = filtered_relclauses_withoutOther)

#Get the scoring for variable importance
HeadRcCombo.varimp <- varimp(HeadRcCombo.cf, conditional = TRUE)
round(HeadRcCombo.varimp, 3)
dotchart(sort(HeadRcCombo.varimp), main = "Conditional importance of variables")

#Report confusion matrix for the random forest model
predictions_headRCcombo <- predict(HeadRcCombo.cf, newdata = filtered_relclauses_withoutOther)
conf_matrix_headRCcombo <- table(predictions_headRCcombo, filtered_relclauses_withoutOther$Head_Type)

#Calculate accuracy
baseline <- max(table(filtered_relclauses_withoutOther$Head_Type)) /
sum(table(filtered_relclauses_withoutOther$Head_Type))
accuracy <- sum(diag(conf_matrix_headRCcombo)) / sum(conf_matrix_headRCcombo)
confusionMatrix(conf_matrix_headRCcombo)

#export the matrix as a table
write.table(conf_matrix_headRCcombo, file = "confusion_matrix_HeadRCCombo.txt", sep = ",", quote = FALSE,
row.names = F)

#3.2. Correlation Matrix
#load a filtered dataframe showing the variables only without metadata
relclausesVariablesOnly <- read.csv("relclausesSub.csv")

#remove all rows containing "Other"
relclausesVariablesOnly <- subset(relclausesVariablesOnly, !grepl("Other", Head, ignore.case = TRUE))

#turn categorical variables into factors
relclausesVariablesOnly$Type <- as.factor(relclausesVariablesOnly$Type) #explained below
relclausesVariablesOnly$Head <- as.factor(relclausesVariablesOnly$Head) #grammatical function of modified NP in main
clause
relclausesVariablesOnly$Position <- as.factor(relclausesVariablesOnly$Position) #headless, pre-nominal, post-nominal
relclausesVariablesOnly$Humanness <- as.factor(relclausesVariablesOnly$Humanness) #whether NP is human/non-human
relclausesVariablesOnly$InformationStatusofHead <- as.factor(relclausesVariablesOnly$InformationStatusofHead) #whether
NP is New or Given
relclausesVariablesOnly$Function <- as.factor(relclausesVariablesOnly$Function) #whether the RC is restrictive or non-
restrictive

#Calculate Cramer's V
cramers_v <- function(x, y) {
  confusion_matrix <- table(x, y)
  chi_square <- chisq.test(confusion_matrix)$statistic
  n <- sum(confusion_matrix)
  phi_square <- chi_square / n
  min_dim <- min(dim(confusion_matrix)) - 1
  crammers_v <- sqrt(phi_square / min_dim)
  return(cramers_v)
}

#Calculate correlation matrix
correlation_matrix <- function(relclausesVariablesOnly) {
  cols <- names(relclausesVariablesOnly)
  n <- length(cols)
  result_matrix <- matrix(NA, nrow = n, ncol = n, dimnames = list(cols, cols))

```

```

for (i in 1:n) {
  for (j in 1:n) {
    if (i == j) {
      result_matrix[i, j] <- 1
    } else {
      result_matrix[i, j] <- crammers_v(relclausesVariablesOnly[[i]], relclausesVariablesOnly[[j]])
    }
  }
}

return(result_matrix)
}

correlation_matrix_final <- correlation_matrix(relclausesVariablesOnly)

#visualize the correlation matrix
print(correlation_matrix_final)
corrplot(correlation_matrix_final, method = "color")

#3.3. The analysis of POSITION as a dependent variable dependent variable

#Run and plot a conditional inference tree
set.seed(1200)
position.ct <- ctree(Position ~ InformationStatusofHead + Humanness + Function + Head + Type, data =
filtered_relclauses_withoutOther)
plot(position.ct, gp = gpar(fontsize = 10))

#Run a random forest
set.seed(315)
position.cf <- cforest(Position ~ InformationStatusofHead + Humanness + Function + Head + Type,
data = filtered_relclauses_withoutOther)

#Get the scoring for variable importance
position.varimp <- varimp(position.cf, conditional = TRUE)
round(position.varimp, 3)
dotchart(sort(position.varimp), main = "Conditional importance of variables")

#Report confusion matrix for the model
predictions_position <- predict(position.cf, newdata = filtered_relclauses_withoutOther)
conf_matrix_position <- table(predictions_position, filtered_relclauses_withoutOther$Position)
confusionMatrix(conf_matrix_position)

#Calculate accuracy
baseline <- max(table(filtered_relclauses_withoutOther$Position)) / sum(table(filtered_relclauses_withoutOther$Position))
accuracy <- sum(diag(conf_matrix_position)) / sum(conf_matrix_position)

#export the matrix as a table
write.table(conf_matrix_position, file = "position_confusion_matrix.txt", sep = ",", quote = FALSE, row.names = F)

#Repeat all of the above for POSITION after collapsing the highly correlated variables (without FUNCTION)

#Run and plot a conditional inference tree
set.seed(1200)
position.ct.collapsed <- ctree(Position ~ InformationStatusofHead + Humanness + Head + Type, data =
filtered_relclauses_withoutOther)
plot(position.ct.collapsed, gp = gpar(fontsize = 10))

```

```

#Run a random forest
set.seed(315)
position.cf.collapsed <- cforest(Position ~ InformationStatusofHead + Humanness + Head + Type,
                                data = filtered_relclauses_withoutOther)

#Get the scoring for variable importance
position.varimp.collapsed <- varimp(position.cf.collapsed, conditional = TRUE)
round(position.varimp.collapsed, 3)
dotchart(sort(position.varimp.collapsed), main = "Conditional importance of variables")

#Report confusion matrix for the model
predictions_position_collapsed <- predict(position.cf.collapsed, newdata = filtered_relclauses_withoutOther)
conf_matrix_position_collapsed <- table(predictions_position_collapsed, filtered_relclauses_withoutOther$Position)
confusionMatrix(conf_matrix_position_collapsed)

#Calculate accuracy
baseline <- max(table(filtered_relclauses_withoutOther$Position)) / sum(table(filtered_relclauses_withoutOther$Position))
accuracy <- sum(diag(conf_matrix_position_collapsed)) / sum(conf_matrix_position_collapsed)

#export the matrix as a table
write.table(conf_matrix_position_collapsed, file = "position_confusion_matrix_collapsed.txt", sep = ",", quote = FALSE,
            row.names = F)

#4. Pairwise comparisons

#4.1. POSITION & FUNCTION OF RC
tableFunctionPosition <- table(relclauses$Function, relclauses$Position)
propFunctionPosition <- prop.table(tableFunctionPosition, 1)
propFunctionPositiondf <- as.data.frame(propFunctionPosition)

#visualize the association
ggplot(propFunctionPositiondf, aes(x = Var1, y = Freq, fill = Var2)) +
  geom_col(position = "dodge") +
  xlab("Function of RC") +
  ylab("Proportion") +
  labs(fill = "Position") +
  geom_text(aes(label = paste0(round(Freq * 100, 1), "%")), position = position_dodge(width = 0.9), vjust = -0.5)

#a chi-squared test
chisq.test(tableFunctionPosition)

#4.2. FUNCTION OF HEAD NP IN MATRIX CLAUSE & FUNCTION OF HEAD NP IN RC

tableTypeHead <- table(relclauses$Head, relclauses$Type)
propTypeHead <- prop.table(tableTypeHead, 1)
propTypeHeaddf <- as.data.frame(propTypeHead)
propTypeHeaddf$Var1 <- factor(propTypeHeaddf$Var1, levels = c("subject", "object", "existential head", "adjunct", "other"))

#visualize
ggplot(propTypeHeaddf, aes(x = Var1, y = Freq, fill = Var2)) +
  geom_col(position = "dodge") +
  xlab("Function of Head NP in Matrix") +
  ylab("Proportion") +
  labs(fill = "Function of Head NP in RC") +
  geom_text(aes(label = paste0(round(Freq * 100, 1), "%")),
            position = position_dodge(width = 0.9), vjust = -0.5)

```

```

#a chi-squared test
chisq.test(tableTypeHead)

#mosaic plot
mosaic(tableTypeHead, shade = TRUE, varnames = FALSE)

#strength of the association
assocstats(tableTypeHead)

#test the distribution of subject heads across RC Types for significance
TypeSubject <- cbind(c(46), c(13), c(5)) #in the following order: Subject RCs, Object RCs, and
                                     #Adjunct RCs
chisq.test(TypeSubject)

#test the distribution of object heads across RC Types for significance
TypeObject <- cbind(c(11), c(8), c(5))
TypeObject
chisq.test(TypeObject)
chisq.test(TypeObject)$expected

#test the distribution of adjunct heads across RC Types for significance
TypeAdjunct <- cbind(c(7), c(6), c(24))
TypeAdjunct
chisq.test(TypeAdjunct)
chisq.test(TypeAdjunct)$expected

#test the distribution of existential heads across RC Types for significance
TypeExH <- cbind(c(27), c(15), c(0))
TypeExH
chisq.test(TypeExH)
chisq.test(TypeExH)$expected

```

#4.3. HUMANNESS & FUNCTION OF HEAD NP IN RC

```

tableTypeHumanness <- table(relclauses$Humanness, relclauses$Type)
propTypeHumanness <- prop.table(tableTypeHumanness, 1)
propTypeHumannessdf <- as.data.frame(propTypeHumanness)

#visualize
ggplot(propTypeHumannessdf, aes(x = Var1, y = Freq, fill = Var2)) +
  geom_col(position = "dodge") +
  xlab("Humannes") +
  ylab("Proportion") +
  labs(fill = "Function of Head NP in RC") +
  geom_text(aes(label = paste0(round(Freq * 100, 1), "%")), position = position_dodge(width = 0.9), vjust = -0.5)

#a chi-squared test
chisq.test(tableTypeHumanness)

#a chi-squared test for the distribution of humans and non-humans across FUNCTION OF HEAD NP IN RC
TypeHuman <- cbind((1), (5), (64))
TypeNon-human <- cbind((34), (44), (66))
chisq.test(TypeHuman)
chisq.test(TypeNon-human)

```

```

#testing the Humanness of subject heads across FUNCTION OF HEAD NP IN RC
HumanHeads <- cbind(c(28), c(5), c(4), c(12), c(21)) #the order from left to right:
#subjects, objects, existential heads, and adjuncts
Non-humanHeads <- cbind(c(36), c(19), c(33), c(30), c(26))
chisq.test(HumanHeads)
chisq.test(HumanHeads)$expected
chisq.test(Non-humanHeads)
chisq.test(Non-humanHeads)$expected

```

#4.4 POSITION & INFORMATION STATUS OF HEAD NP

```

tablePositionInfoStatus<- table(relclauses$Position, relclauses$InformationStatusofHead)
propPositionInfoStatus <- prop.table(tablePositionInfoStatus, 1)
propPositionInfoStatusdf <- as.data.frame(propPositionInfoStatus)

```

```

#mosaic plot
mosaic(tablePositionInfoStatus, shade = TRUE, varnames = FALSE)

```

```

#strength of the association
assocstats(tablePositionInfoStatus)

```

```

#a chi-squared test
chisq.test(tablePositionInfoStatus)

```

#4.5 FUNCTION OF HEAD NP IN THE MATRIX CLAUSE & INFORMATION STATUS OF HEAD NP

```

tableInfoStatusHeadNP <- table(relclauses$Head, relclauses$InformationStatusofHead)
propInfoStatusHeadNP <- prop.table(tableInfoStatusHeadNP, 1)
propInfoStatusHeadNPdf <- as.data.frame(propInfoStatusHeadNP)
propInfoStatusHeadNPdf$Var1 <- factor(propInfoStatusHeadNPdf$Var1, levels = c("subject", "object", "existential head",
"adjunct", "other"))

```

```

#visualize
ggplot(propInfoStatusHeadNPdf, aes(x = Var1, y = Freq, fill = Var2)) +
  geom_col(position = 'dodge') +
  xlab("Function of Head NP in Matrix Clause") +
  ylab("Proportion") +
  labs(fill = "Information Status")+
  geom_text(aes(label = paste0(round(Freq * 100, 1), "%")), position = position_dodge(width = 0.9), vjust = -0.5)

```

```

#a chi-squared test
chisq.test(tableInfoStatusHeadNP)

```

```

#strength of the association
assocstats(tableInfoStatusHeadNP)

```

```

#test each head for skews regarding Information Status
adjunctGivenNew <- cbind((4), 33)
chisq.test(adjunctGivenNew)

```

```

ExGivenNew <- cbind((8), 34)
chisq.test(ExGivenNew)
chisq.test(ExGivenNew)$expected

```

```

objectGivenNew <- cbind((2), (22))
chisq.test(objectGivenNew)

```

```
chisq.test(objectGivenNew)$expected

subjectGivenNew <- cbind((30),(34))
chisq.test(subjectGivenNew)
chisq.test(subjectGivenNew)$expected
```

#4.6 HUMANNESSESS & FUNCTION OF HEAD NP IN MATRIX CLAUSE

```
tableHeadHumanness <- table(relclauses$Head, relclauses$Humanness)
propHeadHumanness <- prop.table(tableHeadHumanness, 1)
propHeadHumannessdf <- as.data.frame(propHeadHumanness)
propHeadHumannessdf$Var1 <- factor(propHeadHumannessdf$Var1, levels = c("subject", "object", "existential head",
"adjunct", "other"))

#visualize
ggplot(propHeadHumannessdf, aes(x = Var1, y = Freq, fill = Var2)) +
  geom_col(position = "dodge") +
  xlab("Function of Head NP in Matrix Clause") +
  ylab("Proportion") + labs(fill = "Humanness") +
  geom_text(aes(label = paste0(round(Freq * 100, 1), "%")), position = position_dodge(width = 0.9), vjust = -0.5)

#a chi-squared test
chisq.test(tableHeadHumanness)

#testing each head
adjunctHumanness<- cbind((4), 33)
chisq.test(adjunctHumanness)
chisq.test(adjunctHumanness)$expected

ExHHumanness <- cbind((12), 30)
chisq.test(ExHHumanness)
chisq.test(ExHHumanness)$expected

objectHumanness <- cbind((5), (19))
chisq.test(objectHumanness)
chisq.test(objectHumanness)$expected

subjectHumanness <- cbind((28),(36))
chisq.test(subjectHumanness)
chisq.test(subjectHumanness)$expected
```

#4.7 FUNCTION OF HEAD NP IN MATRIX CLAUSE & POSITION

```
tableHeadPosition <- table(relclauses$Head, relclauses$Position)
propHeadPosition <- prop.table(tableHeadPosition, 1)
propHeadPositiondf <- as.data.frame(propHeadPosition)
propHeadPositiondf$Var1 <- factor(propHeadPositiondf$Var1, levels = c("subject", "object", "existential head", "adjunct",
"other"))

#visualize
ggplot(propHeadPositiondf, aes(x = Var1, y = Freq, fill = Var2)) +
  geom_col(position = "dodge") +
  xlab("Function of Head NP in Matrix Clause") +
  ylab("Proportion") +
  labs(fill = "Position") +
  geom_text(aes(label = paste0(round(Freq * 100, 1), "%")), position = position_dodge(width = 0.9), vjust = -0.5)
```

```
#test the association for significance
chisq.test(tableHeadPosition)
chisq.test(tableHeadPosition)$expected
```

```
#create a mosaic plot
mosaic(tableHeadPosition, shade = TRUE, varnames = FALSE)
```

```
#test the strength of the association
assocstats(tableHeadPosition)
```

```
#test the associations of Heads across Positions for significance
#order: prenom, postnom, headless
tableSubjPosition <- cbind(c(32), c(13), c(19))
chisq.test(tableSubjPosition)
```

```
tableOtherPosition <- cbind(c(16), c(12), c(19))
chisq.test(tableOtherPosition) #not significant
```

```
tableObjPosition <- cbind(c(14), c(2), c(8))
chisq.test(tableObjPosition)
```

```
tableExHPosition <- cbind(c(18), c(11), c(13))
chisq.test(tableExHPosition) #not significant
```

#4.8. FUNCTION OF HEAD NP IN THE MATRIX CLAUSE & FUNCTION OF RC

```
tableFunctionHeadNP <- table(relclauses$Head, relclauses$Function)
propFunctionHeadNP <- prop.table(tableFunctionHeadNP, 1)
propFunctionHeadNPdf <- as.data.frame(propFunctionHeadNP)
propFunctionHeadNPdf$Var1 <- factor(propFunctionHeadNPdf$Var1, levels = c("subject", "object", "existential head",
"adjunct", "other"))
```

```
#visualize
ggplot(propFunctionHeadNPdf, aes(x = Var1, y = Freq, fill = Var2)) +
  geom_col(position = "dodge") +
  xlab("Function of Head NP in Matrix Clause") +
  ylab("Proportion") +
  labs(fill = "RC Function") +
  geom_text(aes(label = paste0(round(Freq * 100, 1), "%")), position = position_dodge(width = 0.9), vjust = -0.5)
```

```
#a chi-squared test
chisq.test(tableFunctionHeadNP)
chisq.test(tableFunctionHeadNP)$expected
```

```
#strength of the relationship
assocstats(tableInfoStatusHeadNP)
```

#4.9 INFORMATION STATUS OF HEAD NP & FUNCTION OF RC

```
tableFunctionInfoStatus <- table(relclauses$Function, relclauses$InformationStatusofHead)
propFunctionInfoStatus <- prop.table(tableFunctionInfoStatus, 1)
propFunctionInfoStatusdf <- as.data.frame(propFunctionInfoStatus)
```

```
#visualize
ggplot(propFunctionInfoStatusdf, aes(x = Var1, y = Freq, fill = Var2)) +
```

```
geom_col() +
xlab("Function of RC") +
ylab("Proportion") +
labs(fill = "Information Status of Head NP") +
geom_text(aes(label = paste0(round(Freq * 100, 1), "%")), position = position_stack(vjust = 0.5))
```

```
#a chi squared test
chisq.test(tableFunctionInfoStatus)
chisq.test(tableFunctionInfoStatus)$expected
```

```
#strength of the relationship
assocstats(tableFunctionInfoStatus)
```

#4.10 POSITION & HUMANNES

```
tablePositionHumanness<- table(relclauses$Position, relclauses$Humanness)
propPositionHumanness <- prop.table(tablePositionHumanness, 1)
propPositionHumannessdf <- as.data.frame(propPositionHumanness)
```

```
#visualize
ggplot(propPositionHumannessdf, aes(x = Var1, y = Freq, fill = Var2)) +
  geom_col(position = "dodge") +
  xlab("Position") +
  ylab("Proportion") +
  labs(fill = "Humanness") +
  geom_text(aes(label = paste0(round(Freq * 100, 1), "%")), position = position_dodge(width = 0.9), vjust = -0.5)
```

```
#a chi-squared test
chisq.test(tablePositionHumanness)
chisq.test(tablePositionHumanness)$expected
```

```
#strength of the relationship
assocstats(tablePositionHumanness)
```

#4.11 FUNCTION OF HEAD NP IN RC & POSITION

```
tableTypePosition <- table(relclauses$Type, relclauses$Position)
propTypePosition <- prop.table(tableTypePosition, 1)
propTypePositiondf <- as.data.frame(propTypePosition)
propTypePositiondf$Var1 <- factor(propTypePositiondf$Var1, levels = c("subject RC", "object RC", "adjunct RC"))
```

```
#visualize
ggplot(propTypePositiondf, aes(x = Var1, y = Freq, fill = Var2)) +
  geom_col(position = "dodge") +
  xlab("Function of Head NP in RC") +
  ylab("Proportion") +
  labs(fill = "Position") +
  geom_text(aes(label = paste0(round(Freq * 100, 1), "%")), position = position_dodge(width = 0.9), vjust = -0.5)
```

```
#test the association for significance
chisq.test(tableTypePosition)
chisq.test(tableTypePosition)$expected
```

```
#strength of the relationship
assocstats(tableTypePosition)
```

```
#test the individual associations for significance
```



```
tableSubjRCPosition <- cbind(c(63), c(28), c(39)) #order: prenom, postnom, hdless
chisq.test(tableSubjRCPosition)
chisq.test(tableSubjRCPosition)$expected
```

```
tableObjRCPosition <- cbind(c(23), c(7), c(19))
chisq.test(tableObjRCPosition)
chisq.test(tableObjRCPosition)$expected
```

```
tableAdjRCPosition <- cbind(c(27), c(3), c(5))
chisq.test(tableAdjRCPosition)
```

#4.12 INFORMATION STATUS OF HEAD NP & HUMANNES

```
tableHumannessInfoStatus<- table(relclauses$Humanness, relclauses$InformationStatusofHead)
propHumannessInfoStatus<- prop.table(tableHumannessInfoStatus, 1)
propHumannessInfoStatusdf <- as.data.frame(propHumannessInfoStatus)
```

```
#visualize
ggplot(propHumannessInfoStatusdf, aes(x = Var1, y = Freq, fill = Var2)) +
  geom_col(position = "dodge") +
  xlab("Humanness") +
  ylab("Proportion") +
  labs(fill = "Information Status of Head NP") +
  geom_text(aes(label = paste0(round(Freq * 100, 1), "%")), position = position_dodge(width = 0.9), vjust = -0.5)
```

```
#a chi squared test
chisq.test(tableHumannessInfoStatus)
chisq.test(tableHumannessInfoStatus)$expected
```

```
#strength of the relationship
assocstats(tableHumannessInfoStatus)
```

#4.13 FUNCTION OF HEAD NP IN RC & INFORMATION STATUS OF HEAD NP

```
tableTypeInfoStatus <- table(relclauses$Type, relclauses$InformationStatusofHead)
propTypeInfoStatus <- prop.table(tableTypeInfoStatus, 1)
propTypeInfoStatusdf <- as.data.frame(propTypeInfoStatus)
propTypeInfoStatusdf$Var1 <- factor(propTypeInfoStatusdf$Var1, levels = c("subject RC", "object RC", "adjunct"))
```

```
#visualize
ggplot(propTypeInfoStatusdf, aes(x = Var1, y = Freq, fill = Var2)) +
  geom_col(position = "dodge") +
  xlab("Function of Head NP in RC") +
  ylab("Proportion") +
  labs(fill = "Information Status of Head NP") +
  geom_text(aes(label = paste0(round(Freq * 100, 1), "%")), position = position_dodge(width = 0.9), vjust = -0.5)
```

```
chisq.test(tableTypeInfoStatus)
```

```
#strength of the relationship
assocstats(tableTypeInfoStatus)
```

#4.14 FUNCTION OF HEAD NP IN RC & FUNCTION OF RC

```
tableTypeRCFunction <- table(relclauses$Type, relclauses$Function)
propTypeRCFunction <- prop.table(tableTypeRCFunction, 1)
```

```
propTypeRCFunctiondf <- as.data.frame(propTypeRCFunction)
propTypeRCFunctiondf$Var1 <- factor(propTypeRCFunctiondf$Var1, levels = c("subject RC", "object RC", "adjunct RC"))
```

```
#visualize
```

```
ggplot(propTypeRCFunctiondf, aes(x = Var1, y = Freq, fill = Var2)) +
  geom_col(position = "dodge") +
  xlab("Function of Head NP in RC") +
  ylab("Proportion") +
  labs(fill = "Function of RC") +
  geom_text(aes(label = paste0(round(Freq * 100, 1), "%")), position = position_dodge(width = 0.9), vjust = -0.5)
```

```
#a chi-squared test
```

```
chisq.test(tableTypeRCFunction)
chisq.test(tableTypeRCFunction)$expected
```

```
#strength of the relationship
```

```
assocstats(tableTypeRCFunction)
```

#4.15 FUNCTION OF RC & HUMANNESS

```
tableFunctionHumanness<- table(relclauses$Function, relclauses$Humanness)
propFunctionHumanness<- prop.table(tableFunctionHumanness, 1)
propFunctionHumannessdf <- as.data.frame(propFunctionHumanness)
```

```
#visualize
```

```
ggplot(propFunctionHumannessdf, aes(x = Var1, y = Freq, fill = Var2)) +
  geom_col(position = "dodge") +
  xlab("Function of RC") +
  ylab("Proportion") +
  labs(fill = "Humanness") +
  geom_text(aes(label = paste0(round(Freq * 100, 1), "%")), position = position_dodge(width = 0.9), vjust = -0.5)
```

```
#a chi-squared test
```

```
chisq.test(tableFunctionHumanness)
chisq.test(tableFunctionHumanness)$expected
```

```
#strength of the relationship
```

```
assocstats(tableFunctionHumanness)
```