UNIVERSITY OF CALIFORNIA
SANTA CRUZ

THE DISTRIBUTION AND PROCESSING OF REFERENTIAL EXPRESSIONS: EVIDENCE FROM ENGLISH AND CHAMORRO

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

MASTER OF ARTS

in

LINGUISTICS

by

Scarlett Clothier-Goldschmidt

June 2015

The thesis of Scarlett Clothier-Goldschmidt
is approved:

_____
Associate Professor Matthew Wagers, Chair

_____
Professor Sandra Chung

_____
Associate Professor Adrian Brasoveanu

_____
Tyrus Miller
Vice Provost and Dean of Graduate Studies

Contents

List of Figures

List of Tables

Abstract

THE DISTRIBUTION AND PROCESSING OF REFERENTIAL EXPRESSIONS: EVIDENCE FROM ENGLISH AND CHAMORRO

by

Scarlett Clothier-Goldschmidt

This research is a broad investigation of the distribution of nominal expressions in natural language. I explore the rigidity of the person-animacy hierarchy constraint of the Chamorro language (Chung 1998) using corpus data in Chapter 1. In Chapter 2, I present evidence from a self-paced reading study that the 2 >3 component of this constraint is not respected in relative clause processing in English. In Chapter 3, I present corpus data exploring the Accessibility Hierarchy (Ariel 1991) and Centering Theory (Grosz et. al 1995) in English weblogs. While the data presented comes from different languages and different types of measures, I conclude that the distribution and processing of nominals is sensitive to certain grammatical hierarchies and discourse structure.

Dedication

Dedicated to the late Annais Rittenberg, who inspires me every day to do my best and follow my dreams.

Acknowledgements

It all began in the Winter 2012 Syntax I class, co-taught by Sandy Chung and Matt Wagers, where I discovered I wanted to be a linguist and met the people who would make realizing this dream possible.

It's hard to express enough gratitude to one's advisor in a paragraph, but I am forever indebted to my thesis chair Matt Wagers for teaching me everything I know about research; for his wisdom, guidance, patience, and support through the laughter and the tears, through the good times and the janky ones; and for the hours spent discussing experiments and linguistic theory for the past three years.

I equally acknowledge my mentor and kindred spirit Sandy Chung for thought-provoking and insightful conversations; for always having her door open and talking to me whenever I walked by, which was often; for introducing me to the Chamorro language, feeding me amazing Chamorro food, and writing millions of immediate email replies about Chamorro morphology; for being one of my favorite humans; and for our mutual appreciation of Peeps and emojis for the past three years.

I must also thank Adrian Brasoveanu and Pranav Anand for constructive and insightful feedback in the development of this research.

I thank my cohort members Deniz Rudin, Jason Ostrove, Jenny Bellik (and Isaac and Ozan), Kelsey Kraus, and Maho Morimoto for providing friendship, support, shoulders to cry on, babies to play with, and generally good times. I also thank Karl DeVries and Nate Arnett for hours spent working together and being there for me when I needed it.

I must acknowledge my best friend Chelsea Miller, without whom this thesis would not exist because I would have died before writing it. This statement is not an exaggeration.

Lastly, I thank my siblings Pike and Devo, my parents Daniel and Anita, and my grandparents Birchard, Marlynne, Debby, Peter, Grace, and Martin for supporting me

and believing in me in everything I do. My family believes in me more than I do and I wouldn't have gotten here without them.

Introduction

This thesis is a broad investigation of the constraints on the distribution of referential expressions in natural language, in both individual clauses and larger discourses consisting of many sentences. From a functional perspective, language is used by humans to communicate with other humans about events and entities in the world. Some entities have inherent perceptible properties such as animacy, and other properties which shift in relation to a discourse, such as speaker and addressee roles, definiteness, and salience. The referential expressions we use to talk about entities in the world encode some of these properties, and there are often multiple expressions that could be used to refer to an entity. The research presented in this thesis—drawn from a variety of sources, including English and Chamorro translations of the New Testament of the Bible, data from English relative clause processing, and English weblogs—suggests that choices of referring expressions used in natural language are sensitive to person, animacy, and whether the DP is a pronoun. At the discourse level, referential expressions are chosen with respect to structural configurations of preceding sentences as well as the information structure of the discourse.

Some languages rank nominal expressions according to their person and animacy features. Chamorro, an Austronesian language with approximately 40,000 speakers, is an example of such a language. Transitive clauses in Chamorro are constrained by a person-animacy hierarchy. Nominals are ranked according to person (2 >3), animacy (animate >inanimate), and pronominal status (pronoun >non-pronoun), and no canonical transitive clause may have a direct object which outranks its subject according to the hierarchy (Chung 1998). In Chapter 1, I used corpus measures to determine that

the person-animacy hierarchy is never violated and is in fact a hard constraint of the Chamorro grammar. Given that this constraint is a hard constraint, I explored whether it is an idiosyncratic part of the Chamorro grammar or if it is rooted in universal parsing preferences by looking at English, which has no such hard constraint. Bresnan et. al (2001) argue that hard constraints in some languages are mirrored by soft constraints in other languages. In Chapter 2, I consider whether the person component of this hierarchy is mirrored in English.

In Chapter 2, I present data from a self-paced reading study focused on English relative clause processing. There is a well-established finding that subject relative clauses are harder to process than object relative clauses (Gordon et al. 2001). However, this difference in difficulty is eliminated when one of the DPs is a local person pronoun. I compared 2nd and 3rd person pronouns to see whether they act uniformly to alleviate this processing asymmetry. Given that 2nd and 3rd person pronouns both eliminate the asymmetry, I conclude that the 2 >3 ranking of the Chamorro person-animacy hierarchy is not respected in English. In addition to data from the reading study, I consider the distribution of argument pairings within transitive clauses in the English New Testament and find that these data support the experimental result that the person hierarchy is not respected in English.

In Chapter 3, I focus on the properties are referring expressions which shift during a discourse—namely, definiteness and salience. I turn to a corpus of English weblogs to look at the types of nominal expressions which are used to refer to entities throughout a larger personal narrative. I consider the predictions of Centering Theory (Grosz et al. 1995), a theory of parallelism between sentences in a discourse, and the Accessibility Hierarchy (Ariel 1991), a ranking of nominals based on their uniqueness and informational content. I find that these constraints are largely obeyed in this type of personal narrative, and that the choice of referential expressions used is sensitive to the number of entities under discussion and the shifting topic matter of the discourse.

The research presented here makes several contributions through the use of com-

putational and experimental methods. The person-animacy hierarchy is sometimes violated in elicitation, but the data presented here suggests that it is actually a hard constraint. Virtually all Chamorro speakers also speak English, and this finding highlights the importance of using multiple measures to evaluate the strength of a constraint when informants have knowledge of two distinct grammars with divergent constraints. These measures also help us to evaluate the the claim made by Bresnan et al. (2001), which is based on distributional patterns in the English Switchboard Corpus (Godfrey et al. 1992). Corpus measures give us an insight into aggregated patterns in natural language that may not be visible in individual sentences alone. If hard constraints are mirrored by soft constraints, this suggests that such constraints are rooted in innate preferences which surface cross-linguistically. We would expect that constraints which are based on innate cognitive preferences should surface not only in aggregated corpus statistics, but in the real-time processing of sentences. In the case of the person component of the person-animacy hierarchy, we do not find evidence that it is mirrored in English in either measure. This finding adds to our understanding of language-specific constraints and the extent to which they are the reflexes of universal parsing preferences.

I leave open further possibilities for connecting the research presented in this thesis. Since the person-animacy hierarchy is a constraint which operates over single clauses while the Accessibility Hierarchy and Centering Theory operate over larger discourses, we might expect interactions between these constraints. Thus, it would worthwhile to look at both animacy, person, and pronominal features in addition to definiteness and salience within the same corpus. We might expect that the ways these constraints interact would be different for English and Chamorro, since the person-animacy is a hard constraint of the Chamorro grammar which is not called off in favor of discourse structure. The corpora used in this thesis were the New Testament and English weblogs, which represent types of language both distinct from one another and distinct from other types of written language like short stories or newspaper articles. Further kinds of corpora could be used to see whether the trends reported here are representative of

the distributions found in other types of writing. While there are not many Chamorro weblogs, we have access to at least one where the data presented in Chapter 3 could be explored for Chamorro.

While the evidence presented here is diverse, an overall picture emerges. The research shows that referring expressions are not randomly distributed, but are chosen based on grammatical features and information structure of a discourse. While the restrictions on nominal distribution vary cross-linguistically, we find the same grammatical features that affect linguistic processes across languages. The work presented here adds to our understanding of hard constraints and their relationship to psycholinguistic processing.

Chapter 1

The Chamorro Person-Animacy Hierarchy

1.1 Background

Cultural Context

Chamorro is an Austronesian language spoken in the Mariana Islands by approximately 40,000 speakers. English is the language of public settings in the region where Chamorro is spoken. Because of this, the majority of Chamorro speakers also speak English and therefore have knowledge of two distinct grammars with different constraints. This has consequences for elicitation because speakers' judgements about Chamorro sentences may reflect their knowledge of English grammar.

A cultural fact about languages of the Pacific is that speakers do not want to say 'no' to outsiders who are learning the language (Chung, p.c.). This also has consequences for elicitation data, because linguists often want to discover which sentences are ill-formed through the elicitation of negative judgements. The aim of the current research is to supplement speakers' reported linguistic judgements about a particular constraint of the language in an interest to address these features of the cultural context of Chamorro.

The Person-Animacy Hierarchy

An example of a constraint of the Chamorro grammar is the person-animacy hierarchy:

(1) [1] 2nd person > 3rd person animate pronoun > animate > inanimate

This hierarchy restricts the co-occurrence of certain subjects and direct objects within a transitive clause, such that no clause may have a direct object which outranks the subject according to the hierarchy (Chung 1998). Grammatical transitive clauses in Chamorro can have a subject and a direct object which are equal on the hierarchy:

(2) Kao para u na'homlu' gui' gi Sabaf
Q FUT AGR:3.SG heal him P Sabbath
'if he [Jesus] would heal him on the Sabbath' 3 anim pro = 3 anim pro
Mark 3:2 (Camacho 2007)

Additionally, because 1st person is unranked, a transitive clause with a 1st person subject or object will be well-formed. But if the direct object outranks the subject, the clause is ill-formed. This fact is exhibited by the following example:

(3) * Kao **ha** kuentusi **hao** åntis di u hånao?
Q **agr:3.sg** speak.to **you** before AGR:3.SG go
'Did he speak to you before he left?' *3>2
(Chung 2012)

Here, the bolded 3rd person singular agreement marker which occurs with realis transitive verb forms is bolded. This morphology signals that the subject of the transitive clause is 3rd person. The object of the transitive clause is the 2nd person pronoun *hao*. Because the object is 2nd person, it outranks the subject according to the hierarchy, and the structure is ungrammatical.

To avoid a violation of the hierarchy, the sentence above could be expressed in the passive voice:

[1] Note that 1st person is not ranked according to this hierarchy. The fact that Chamorro's hierarchy does not include first person makes it typologically rare (Aissen 1999). Cross-linguistically, there seems to be an implicational relationship—if a hierarchy includes 2nd person, it will also include 1st person.

(4) Kao kuinentusi **hao** åntis di u hånao?
Q **pass**.speak.to **you** before AGR:3.SG go
'Were you spoken to by him before he left?'
(Chung 2012)

The infix *-in-* indicates that the verb is passivized and that its external argument is definite and singular. Because the verb form is realis intransitive, we do not see any agreement markers. The subject of the intransitivized verb is the 2nd person pronoun *hao*. Because there is only a subject and there is no direct object in this clause, the hierarchy constraint is evaded.

The alternation shown above could be accounted for syntactically. In the passive, the direct object of the transitive verb is promoted to subject position, and the transitive subject is expressed as an internalized argument as indicated by the features of the passive infix. From this, we might conclude that person-animacy hierarchy is a constraint which is satisfied by alignment of a syntactic hierarchy (subject >direct object) and a hierarchy of prominence with respect to other grammatical categories (animacy, person, pronoun). DPs are ranked on the person-animacy hierarchy based on their prominence, and the hierarchy is satisfied when the argument in the clause which is more grammatically prominent is placed in the more prominent syntactic position (Aissen 1999).

But passive clauses are not the only environment where the hierarchy is evaded. There are certain types of transitive clauses where the hierarchy is also evaded, despite having the same syntactic configuration as the violating clauses. The generalization is that only transitive clauses with canonical transitive morphology are subject to the person-animacy hierarchy. If this constraint is a constraint on the alignment of grammatical hierarchies, it requires such alignment only in canonical transitive clauses.

Morphologically Marked Verb Forms

Another passive example is given below:

passive

(5) sinedda' gui' as Jesus
find:PASS 3.pro by Jesus
'He was found by Jesus' 3 anim pro>animate

John 5:14 (Camacho 2007)

Here, the agent of *find* is *Jesus* and the patient is a 3rd person pronoun, which would violate the ranking of 3rd person animate pronoun > animate if expressed as a transitive clause. Instead, the verb bears the passive infix *-in-* and its subject is *gui'*, the 3rd person pronoun, while *Jesus* is expressed in a syntactic oblique.

Chamorro also has an antipassive construction[2] :

antipassive

(6) Manoppe si Jesus
answer:ANITPASS CASE Jesus
'Jesus answered him' 3 anim pro>animate

John 18:34 (Camacho 2007)

The antipassive either expresses the direct object of its corresponding transitive as a syntactic oblique, or the argument is implicit as in this example. The transitive verb *answer* bears the antipassive prefix *man-* which makes the clause intransitive. The subject of *manoppe* is the DP *Jesus*, and the direct object pronoun corresponding to *him* is omitted.

As stated in the previous section, these passive and antipassive examples are compatible with a syntactic analysis. Because *man-* and *-in-* render a transitive verb intransitive, it could be argued that the hierarchy is avoided because the clause is intransitive and only has a subject. However, there are verb forms which show that this is not the only way to avoid a violation of the hierarchy. Some transitive verbs are exceptionally inflected with possessor agreement:

[2]Examples are cited in their original orthography.

possessor

(7) Yan-ñiha hao.
Like-agr:3.pl you
'They like you.' 3 anim pro>2

In Chamorro, the verb *ya-* is an example of such a transitive. This verb is inflected with the possessor marker to indicate agreement with its subject. In this example, the subject is a 3rd person plural pronoun, and the verb bears 3rd person plural possessor agreement. The direct object is 2nd person, which outranks the subject according to the hierarchy, and yet the structure is fully grammatical. The fact that examples like this one are grammatical indicate that the hierarchy is not simply a syntactic constraint. Here, there is nothing special about the shape of the clause, it just does not bear canonical transitive morphology.

Finally, we have an example of a verb with wh-agreement morphology:

wh

(8) i Yi'os todu i grasia, ni umågang hamyo
the god all the grace, who called:agr.WH you.pl
'The god of all the grace, who called you' animate>2

1 Peter 5:10 (Camacho 2007)

In this example, we have a relative clause *ni umgang hamyo*, in which the transitive verb *ågang* (call) bears the wh-agreement morpheme *um-*. The subject of this verb is a relativized 3rd person DP while the direct object is a 2nd person pronoun. According to the hierarchy, this argument pairing in a transitive clause should be ungrammatical, but the violation of the hierarchy is avoided because the verb bears special wh-agreement.

We can conclude from the examples shown in this section that any verb which exhibits special morphological marking is exempt from the restrictions imposed by the person-animacy hierarchy. Since there are transitive clauses which avoid the hierarchy, we consider the person-animacy not to be a strictly syntactic constraint, but one which is sensitive to the morphology of the verb (Chung 1998). It can be stated as a syntactic

constraint which only targets transitive clauses with canonical morphology.

1.2 Method

Data from elicitation suggest that the person-animacy hierarchy is a constraint of the Chamorro grammar (Chung 1998). However, it is difficult to determine the strength of the person-animacy hierarchy constraint based on this data alone. Elicitation data report speakers' introspective judgements about the well-formedness of sentences, which may be difficult to interpret for reasons discussed in the introduction of this chapter—namely, that almost all Chamorro speakers also speak English, which has an influence on their grammar, and there is a hesitation to reject sentences produced by non-native speakers.

In the present study, we examine a corpus of written Chamorro to supplement speaker judgements. For this analysis, we used the Chamorro *Nuebu Testamento*, a translation of the English New Testament written by Bishop Tomas A. Camacho, who is a native Chamorro speaker, in collaboration with a small group of Chamorro native speakers (Camacho 2007). If sentences with hierarchy-violating argument pairings are ungrammatical in Chamorro because the person-animacy hierarchy is a hard constraint, we would not expect a native speaker to produce sentences which violate the hierarchy. By looking at a translation of the Bible, we can see what types of Chamorro structures a native speaker would actually produce given the English text.

We annotated transitive clauses found in the New American Standard version of English Bible (New American Standard Bible 1997), which is the version that was used to write the Chamorro translation. We identified the subject and object of each clause, and annotated their person, number, animacy, and form features. We then looked to see how these clauses were translated in the Chamorro text. Our clause categorizations are given in Table 1.1. For each type of clause, the English column gives an example from the English translation of the New Testament (New American Standard Bible 1997)

and the Chamorro column provides a schematic example of how the clause would be translated in the Chamorro translation (Camacho 2007).

Type	English	Chamorro
Passive	the glory of God has illumined it	because it is illuminated by the glory of God
Antipassive	whenever I go to Spain — for I hope to see you	I want to visit in my travels to Spain
Wh	an enemy has done this	it was my enemies who did this
Circumlocution	all who are with me greet you	greetings from all my friends here

Table 1.1: Clause Categorizations

1.3 Results

We collected 600 English tokens. Due to certain features of Chamorro syntax, we excluded some from this analysis. All clauses containing a 1st person argument were excluded, since the person-animacy does not rank 1st person. As discussed above, clauses with wh-agreement are exempt from the person-animacy hierarchy restriction. Thus, English clauses which are relative clauses are also excluded from this tabulation because they are not candidates for hierarchy-violating clauses. Lastly, in Chamorro there is an independent restriction on transitive clauses with 3rd person plural non-pronoun subjects (Chung 1998). Thus, in tabulating the number of clauses which violate the hierarchy, we exclude English clauses which have subjects of this type.

Our sample then consists of 356 tokens (300 violating, 56 non-violating), with the proportion of each translation type reported in Figure 1.1:

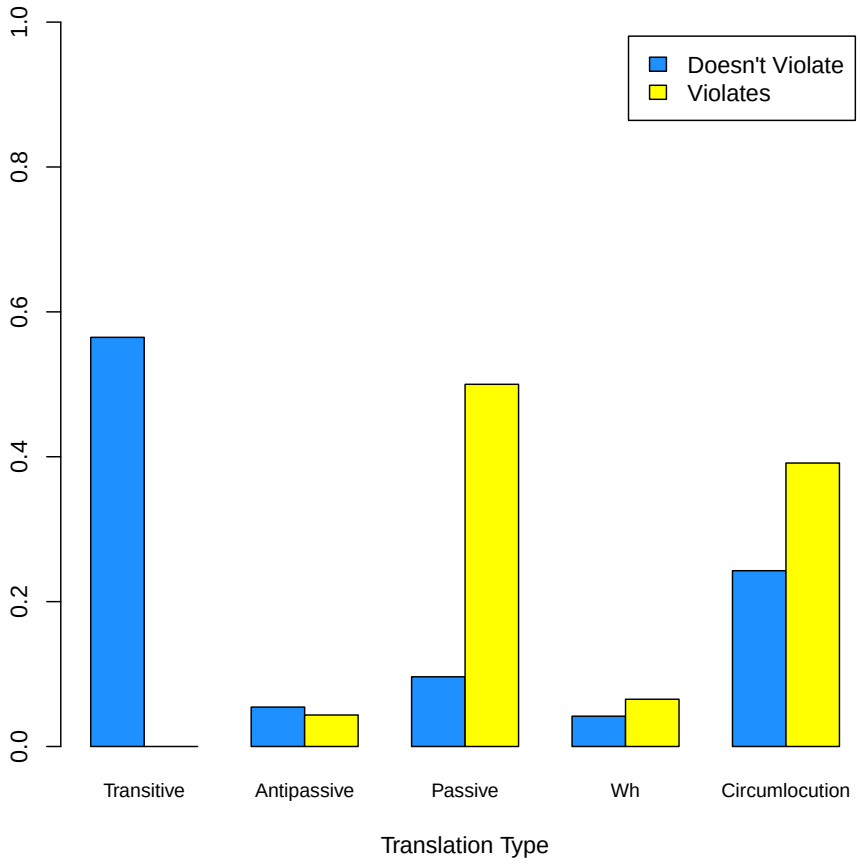


Figure 1.1: Probability of Translation Type Given English Argument Pairing

There are no cases which violate the hierarchy in English and are translated as transitives in Chamorro. Such clauses are usually translated as passive, followed by circumlocution. Wh-agreement is used in focus constructions, and these likely have semantic differences from canonical transitive clauses which may explain why this translation type is not very frequent. The elevated rate of passives compared to antipassives reflects the fact that passives are more frequent than antipassives in Chamorro. The elevated rate of circumlocutions is because this category is very broad. Any clause that was not transitive, passive, wh, or antipassive was treated as a circumlocution. For instance, if a verb is transitive but one of the DPs from the English is also changed to avoid a hier-

archy violation, these cases were counted as circumlocutions. Finer-grained annotation categories are needed to address these cases.

Violating Cases

In our corpus investigation, we only found two examples that violate the hierarchy. The first comes from Revelation 20:3:

(9) Ha yute' **gui'** hålom i anget gi lugåt anai
agr:3.sg throw him inside the angel to place where
manmapreresu i aniti siha
agr:3.pl.PASS.imprison.PROG the soul PL
'the angel threw him into the place where the souls were imprisoned'

Revelation 20:3 (Camacho 2007)

In this example, *the angel* is the subject of the transitive verb *throw*, and the direct object is the pronoun *gui'*. According to the 3rd person animate pronoun > animate ranking, this example should be ungrammatical. However, there are contextual factors surrounding this example which may account for its acceptability. In this context, *gui'* is used to refer to Satan in a serpent form. In general, *gui'* is used for human referents. Because the snake also refers to Satan, who could be construed as human, *gui'* is used, but this could be for effect. The pronoun does not have the same human status that it normally does, and thus this example could be construed as acceptable even though it would not be if *gui'* had a canonical human referent.

Another possible explanation is related to the subject matter of the discourse. Tomlin (1983) explores hierarchies which emerge as a property of material under discussion in the discourse. In an analysis of sentences from hockey game narration, he found that the following hierarchy accounted for the syntactic distribution of 98% of sentences in the corpus:

player with puck > puck > player without puck

If any of these DP types co-occurred within a transitive clause, the higher ranked DP would be the subject. If the higher ranked DP would be an object of a transitive, the clause was passivized. This hierarchy does not make reference to grammatical features such as person, animacy, and number, but rather it makes reference to entities referred to in the discourse and their relevance as participants in the discourse.

In the violating case shown above, the discourse is about angels and demons. It is possible that the referents of the DPs *i anghet* and *gui'* form a hierarchy in which an angel is above a demon, and thus a violation of the person-animacy hierarchy, a grammatical constraint, is permitted in this case due to the elevated discourse status of an angel compared to a demon in this context. While this possibility would be interesting to explore, we simply do not have enough violating examples to be able to evaluate whether such a discourse-level constraint is active here.

The other violating case is a relativized context without wh-agreement:

(10) ya gaigi gi hilo' kada ulu nå'an ni ha ensusutta si Yu'os.
and there.were on top each head name which agr.3.sg insulted CASE god
'and on his heads were blasphemous names'

Revelation 13:1 (Camacho 2007)

Here, the relativizer *ni* indicates a relative clause, but the verb bears canonical transitive morphology, and thus we expect this case to be a violation if names are inanimate and God is animate, since the names are the subject of *insult* while God is the object. Given that we have so many data points and this is one of two violating cases we have found, we can reasonably interpret this clause as noise in the data.

1.4 A Hard Constraint

This corpus study strongly suggests that the person-animacy hierarchy is in fact a hard constraint of the Chamorro grammar. Given this, how are we to understand this constraint? Is it a language-specific morphological constraint that is part of the

grammar, or can we derive this constraint from independent processing principles?

Minkoff (2010) argues for universal processing constraints whose interactions produce animacy hierarchy effects. He states that (11) has no independent status as a constraint of the grammar if a language is VSO and allows *pro*-drop of subjects:

(11) The subject of any transitive verb must be at least as animate as its object.

Rather, this constraint is the natural by-product of the interaction of two independently motivated principles:

Graded Animate Agent Preference (GAAP): The processor prefers a DP to be an agent in proportion to its animacy.

Base Generation Bias (BGB): All things being equal, the processor prefers a base generated analysis of any declarative sentence.

The BGB is proposed on the basis of processing simplicity, while the GAAP is motivated by the following contrasts. In general, the parser prefers to assign an agent role to the left-most argument in an English sentence, as shown in these examples:

(12) #Mary ruffled John's papers but it was on purpose.

(13) Mary ruffled John's papers but it was by accident.

In (12) the use of *but* signals that *it was on purpose* is in contrast with the proposition expressed by the first conjunct, and contrast leads to infelicity because *Mary* is posited to be an agent in control of her actions. In (13), *but* signals that *it was by accident* is in contrast to our agency assumptions, and the felicity of this contrast supports that *Mary* is considered to be an agent acting purposefully. However, there is no agent preference when the subject of the sentence is inanimate:

(14) #The wind ruffled John's papers but it was on purpose.

(15) #The wind ruffled John's papers but it was by accident.

By the logic of the previous set of examples, the infelicity of (14) should mean that *the wind* is posited as an agent. However, the infelicity of (15) shows that *the wind* cannot be an agent. These contrasts motivate the GAAP, in that inanimate subjects do not generate the same inferences about agency that animate subjects do.

In the default word order of English, the leftmost argument is the subject. Therefore, the parser can determine the subject before anything else in the sentence is encountered. In languages which are verb-initial and allow *pro*-drop of the subject, the parser cannot determine which argument is the subject until both arguments are encountered according to the BGB.

The following example comes from Mam (Minkoff 2010):

(16) # Ø-Ø-t-il tx'yan qya.
DC-ABS3S-ERG3S-see dog woman
'The dog saw the woman.'

Due to the fact that Mam is VSO and allows *pro*-drop of the subject, the subject cannot be identified until both DPs in the sentence are encountered. According to the GAAP, the processor prefers to assign the agent role to *woman* and not *the dog*. Whenever the object is more animate than the subject, the BGB and GAAP will make incompatible predictions, which leads to unacceptability.

In English, the left-most DP corresponds to the subject, so there is never a conflict between these two constraints because the parser identifies the subject as soon as it encounters the first DP. The BGB and GAAP never interact in an incompatible way in English, even if a sentence flouts (11). However, if the sentence flouts (11) in a verb-initial language, then the predictions are incompatible. Therefore, (11) does not have any independent status as a constraint in the grammar, but rather falls out naturally from innate preferences of the parser and syntactic properties of the language.

While this might give us an understanding the animacy component of the person-animacy hierarchy, this account cannot explain the person or pronoun components. If we accept this theory as an account of the animacy component, we might wonder

why it is called off in the cases of exceptional agreement as discussed above. In the case of the intransitive clause types, the passive and the antipassive, the verb indicates that the first DP is the subject. In the case of wh-agreement, a DP is fronted and the verbal agreement indicates its argument position. These three cases are consistent with Minkoff's account, in that the parser receives additional information that allows it to identify which argument is the subject before both arguments are encountered. However, the exceptional possessor agreeing verbs are not accounted for, because the verbal agreement carries the same information as canonical transitive agreement with regard to identification of the subject.

Additionally, transitive clauses in Chamorro are subject to an additional restriction—transitive clauses cannot have 3rd person plural non-pronoun subjects. This constraint does not fall out naturally from any interacting constraints like BGB and the GAAP and may give us cause to wonder whether we should try to understand the person-animacy hierarchy as a principled constraint, or an idiosyncratic property of the language.

1.5 Hard and Soft Constraints

The corpus data suggest that person-animacy hierarchy is an example of a hard language-specific constraint. Bresnan et al. (2001) argue that phenomena which are attributed to hard constraints in some languages show up as statistical preferences in other languages. For instance, some languages rank DPs according to their person or animacy features and require that the most prominent DP in a clause be in subject position. This configuration aligns grammatical function with other grammatical rankings of prominence, yielding a more optimal subject choice.

Lummi is an example of such a language. In Lummi, if a clause contains a local person pronoun and a 3rd person pronoun, the local person pronoun must be in subject position.

(17) 3rd person → {1st person, 2nd person} PASSIVE

(18) {1st person, 2nd person} → 3rd person ACTIVE

Bresnan et al. (2001) considered passivization rates in the Switchboard corpus of English. They found that when speakers described situations of local person entities acting on 3rd person entities, they used passives less often than when describing 3rd person entities acting on 3rd person entities. Conversely, speakers used more passives for situations of 3rd person entities acting on local person entities. These patterns are consistent with the constraints on passivization in Lummi. Evidence of this kind suggests that language-specific hierarchies could be a product of universal preferences. If this is so, can we find evidence that English respects a soft version of the person-animacy hierarchy?

English Argument Pairings

From our corpus of transitive clauses from the English translation of the Bible, we can look at the pairings of subject and object in English. In this data set, we exclude only the 1st person cases, since 1st person is unranked according to the Chamorro person-animacy hierarchy. Unlike our previous sample, we include relative clauses in this calculation because the morphological condition on specially marked verbs would not apply in English given that there is no special wh-morphology for English verbs. We also included clauses with 3rd person plural non-pronoun subjects since this is an independent restriction in Chamorro which also does not apply in English.

		OBJECT				
		2 PERS	3 ANIM *pron.*	ANIM	INANIM	
SUBJECT	2.PERS	1	9	16	52	78
	3.P ANIM *pronoun*	15	43	28	74	160
	ANIMATE	35	40	27	101	203
	INANIMATE	2	3	0	15	20
		53	95	71	242	N =461

Figure 1.2: Actual Distribution

		OBJECT				
		2 PERS	3 ANIM *pron.*	ANIM	INANIM	
SUBJECT	2.PERS	9.0	16.1	12.0	40.1	77.2
	3.P ANIM *pronoun*	18.4	33.0	24.6	84.0	160
	ANIMATE	23.3	41.8	31.3	106.6	202.9
	INANIMATE	2.3	4.1	3.1	10.5	20
		53	95	71	241	N =461

Figure 1.3: Predicted Distribution Under Independence Assumptions: (S=A, O=B) = P(A | S) x P(B | O)

Based on the differences between these tables, we can see that argument pairings do not cluster the way that they would by chance.

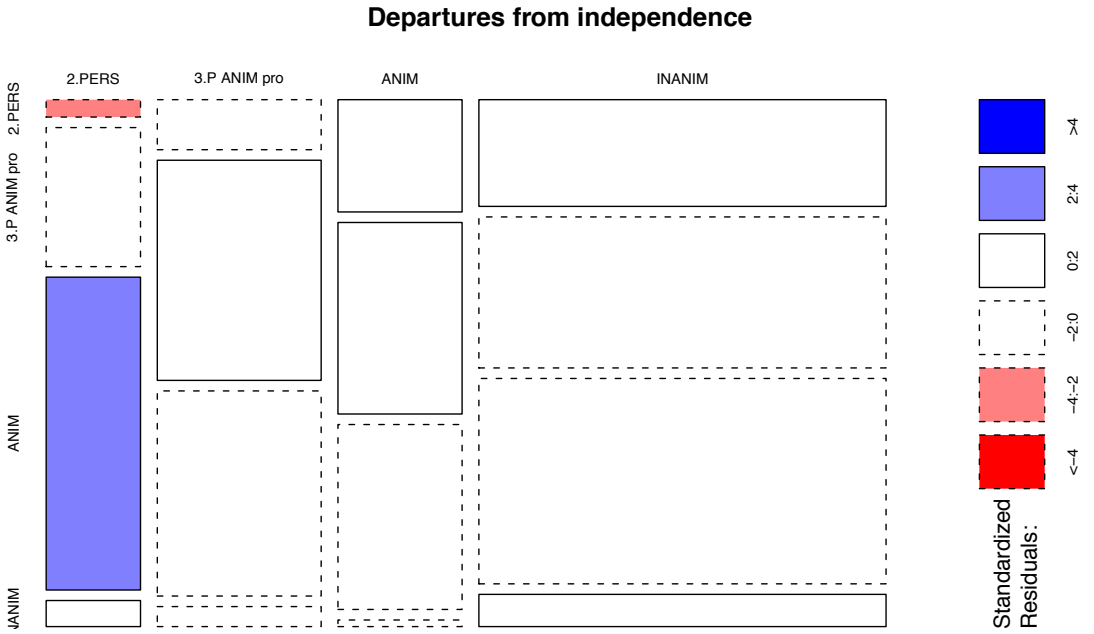


Figure 1.4: Departures from Independence Assumptions

We can make some generalizations from this data. Gray cells indicate argument pairings which violate the person-animacy hierarchy, and we find that there are not fewer of these violating pairings than we would expect by chance, suggesting that a person-animacy hierarchy overall is not active. Within these hierarchy-violating pairings, we observe in Figure 1.4 that there are significantly more animates acting on 2nd person than expected by chance. This pattern suggests that English is not sensitive to the ranking of 2nd person pronouns above all other DP types. However, we see that when 2nd person pronouns are objects, they usually have animate subjects. We could have imagined more inanimate subjects acting on 2nd person, but indeed we do not find this, despite finding slightly more inanimates acting on inanimates than we expect by chance.

As we interpret the differences between the actual and predicted distributions of English argument pairings, it is worth pointing out that the person-animacy hierarchy

can be thought of as consisting of three sub-hierarchies:

(a) pronoun >non-pronoun
(b) 2nd person >3rd person
(c) animate >inanimate

In an OT syntax view of the person-animacy hierarchy (Aissen 1999), we can derive this difference in strength between these sub-hierarchies from the ranking of the constraints of the grammar corresponding to these hierarchies. Since the animacy hierarchy is reflected here while the person hierarchy is not, the *inanimate >animate constraint corresponding to the animacy hierarchy is ranked higher than the *3 >2 constraint. This makes the prediction that other languages could have different constraint rankings, but more corpus data is needed to evaluate that prediction.

1.6 Conclusion

In this chapter, we considered corpus data as a means to probe what is in fact a hard constraint of the Chamorro language. Despite the fact that violating sentences are almost never spontaneously produced, the constraint is often violated in elicitation. By looking at this type of data, we have gained insight into a property of the Chamorro grammar which is hard to evaluate otherwise. However, more could be looked at in this data. To have more precise corpus statistics, we should account for the fact that the categories have overlap. For instance, 3rd person pronouns also fall under the category of animate, but they are not counted here because they are counted in the 3rd person pronoun category. Additionally, the metrics used by Bresnan et al. (2001) involved comparison of rates, choosing one as a baseline, and here we are only looking at counts based on frequency assumptions. Further comparisons could be made. Lastly, the corpus we are using is the New Testament of the Bible, which likely has inflated counts for 3rd person pronouns and 3rd person entities in general due to the nature of the discourse.

Chapter 2

Grammatical Person, Pronouns, and the Subject-Object Processing Asymmetry in Relative Clauses

2.1 Background

In the previous chapter, we considered evidence from corpus data indicating that the person-animacy hierarchy is a hard constraint of the Chamorro grammar.

The Person-Animacy Hierarchy

(19) 2nd person > 3rd person animate pronoun > animate > inanimate

Given that this is a hard constraint, we can wonder about its status in the grammar. Why would a language have such a complex constraint? As discussed in the previous chapter, we can think of this hierarchy as arising from the interaction of three sub-hierarchies:

(a) pronoun >non-pronoun
(b) 2nd person >3rd person
(c) animate >inanimate

Among these sub-hierarchies, we can see some divisions which determine grammatical prominence cross-linguistically and have consequences for sentence processing. In this section, we further explore the idea pursued by Bresnan et. al (2001) that hard con-

straints in some languages are mirrored by soft constraints in others. Support for this idea comes from the fact that the rates of passivization in English mirror those found in Lummi, which has a hard person constraint. We also find evidence for a pronoun hierarchy and for an animacy hierarchy from English relative clause production and relative clause comprehension. Gennari & MacDonald (2008) showed that there is a tendency to put human DPs in subject position in English, as evidenced by rates of passive and active clauses in a relative clause production study. Given two DPs and an experiencer-theme verb, participants were more likely to produce passive relative clauses than active relative clauses:

(20) The director that the movie pleased received a prize.

(21) The director that was pleased by the movie received a prize.

Because human experiencers are highly animate, they are prominent arguments and speakers want to put them in subject position. This finding suggests English respects the animate >inanimate hierarchy to some degree.

2.2 The Subject Advantage in Relative Clauses

There is considerable psycholinguistic literature documenting the asymmetry in processing of English relative clauses with subject gaps compared to object gaps.

(22) SUBJECT GAP The nurse that _ welcomed the mechanic with a smile ran a marathon.

(23) OBJECT GAP The nurse that the mechanic welcomed _ with a smile ran a marathon.

It has been shown that reading times are slower in object-extracted relative clauses (ORCs) than subject-extracted relative clauses (SRCs), and comprehension question accuracy is higher for SRCs compared to ORCs (Gordon et al. 2001). Though this finding is robustly attested, the source of the SRC preference is not fully understood.

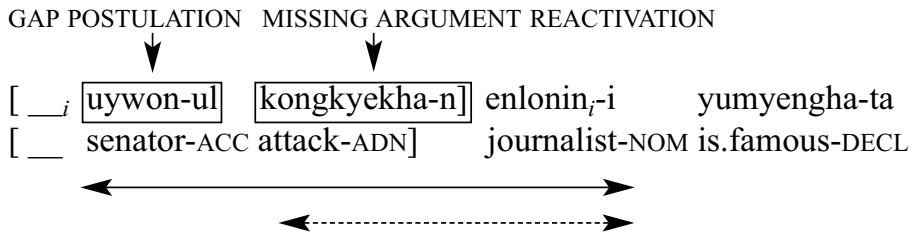
There are many theories about the factors underlying the preference, but some make identical predictions and are therefore hard to tease apart. It is also highly likely that there are interacting factors which modulate relative clause processing and that no single factor alone is responsible.

One account posits that memory limitations are responsible for the ORC processing difficulty. The comprehender has to hold onto two DPs before the wh-dependency can be resolved in the case of an object extraction, and these DPs interfere with one another in retrieval at the embedded verb when they have similar features (Gordon et al. 2001).

Another theory posits a cost of holding a DP filler in memory while integrating other structure (Gibson 2000). In the case of a subject extraction as in (22), the comprehender encounters the relative clause head followed almost immediately by its gap, and can therefore resolve the wh-dependency before encountering any complex verbal material or additional DPs. In an object extraction as in (23) however, the comprehender has to hold the relative clause head in memory, and the dependency is resolved only after encountering the subject DP and the embedded verb. An issue with an account based on linear distance and memory load is that it predicts a subject relative clause advantage in languages with postnominal relative clauses, but it does not predict this advantage with prenominal relative clauses.

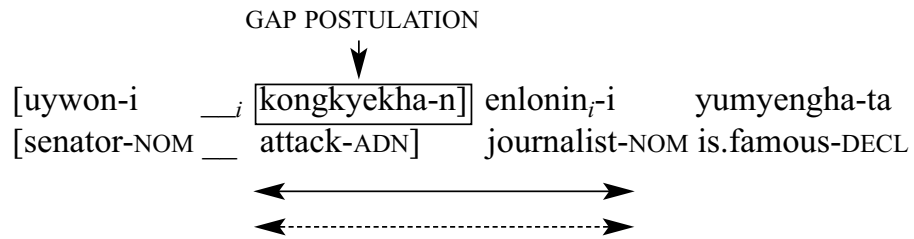
Korean is a language with prenominal relative clauses. The following diagram, taken from Kwon et al. (2010) provides a schematic for an SRC and ORC in Korean:

(15) Subject relative clause



'The journalist who attacked the senator is famous.'

(16) Object relative clause



'The journalist who the senator attacked is famous.'

Figure 2.1: Structural Distance of Prenominal Relative Clauses

Let's first consider the SRC. Given that the gap is sentence initial and phonologically silent, the parser does not posit a gap until the ACC-marked DP is encountered. The distance between the the gap postulation site and identification of its filler is marked with the solid line, and this distance is greater for SRCs. In contrast to the SRC, the ORC gap is not postulated until the embedded verb is encountered, and thus the distance between gap and filler is shorter in the ORC. Additionally, in order to interpret the relative clause, the first argument will have to reactivated at the embedded verb. The distance between this reactivation point and the filler is the same for both SRCs and ORCs, as indicated by the dashed line. According to these two distance metrics, ORCs should either be the as easy or easier than SRCs in Korean. Kwon et. al (2010) find that this prediction is not borne out, but in fact ORCs are more difficult to process. Thus, a distance-based theory of relative clause processing cannot account for the subject/object asymmetry cross-linguistically.

Another theory invokes a syntactic hierarchy. Based on typological variation in relative clauses, Keenan & Comrie (1977) propose the ACCESSIBILITY HIERARCHY OF RELATIVIZATION:

subject >direct object >indirect object >oblique

According to the hierarchy, if a language can relativize a DP occupying any of these positions, it can relativize a DP in all of the positions which are higher on the hierarchy as well.

Keenan & Comrie (1977) speculated that the typological variation may have a basis in processing—DPs on the hierarchy are ranked based on how difficult they are to relativize. Why is it harder to relativize DPs which are lower on the hierarchy? To say that the difficulty of relativizing a DP is because of its ranking would be circular reasoning. A view advanced by Kwon et al. (2010) is that object gaps are always more deeply embedded than subject gaps according to theories of phrase structure. If the subject/object asymmetry is ultimately grounded in inherent differences in phrase structure complexity between subject gaps and object gaps regardless of whether the language has prenominal or postnominal relative clauses, then this makes a connection between generative theories of syntax and the typological variation that we find cross-linguistically.

2.3 Elimination of the Subject/Object Asymmetry

It has also been shown that this asymmetry between object extractions and subject extractions disappears when the second DP encountered in the RC is a pronoun:

(24) SRC The nurse that _ welcomed you with a smile ran a marathon.

(25) ORC The nurse that you welcomed _ with a smile ran a marathon.

With this manipulation, differences in reading time and comprehension accuracy between SRCs and ORCs are neutralized. This pattern has been demonstrated for 1st

person and 2nd person pronouns (Gordon et al. 2001, Warren & Gibson 2005). Because only singular local person pronouns have ever been looked at in this regard, there are multiple possible reasons for the alleviation of the subject/object asymmetry.

This finding about pronouns is consistent with a similarity-based interference account. Full DPs are always 3rd person and in this experiment, they consisted of a determiner and a noun phrase. In contrast, local person pronouns are 1st or 2nd person and syntactically simplex. Because local pronouns have different features than full DPs in terms of person and structure, their representations should not interfere with those of full DPs like *the nurse* in memory and they are easier to encode and retrieve. Another possibility is that the surface differences between pronouns and full DPs are driving the reduction in the subject/object asymmetry. Having DP types with different features has been shown to improve the unacceptability of center embedding structures (Bever 1974):

(26) *The reporter the senator the photographer met knows trusts the president will resign.

(27) The reporter everyone I met knows trusts the president will resign.

When the DP types in (26) are mixed, the unparsable sentence becomes acceptable, as can be seen in (27). Pronouns have a different surface form than full DPs in that they are shorter and morphologically simplex. Just as (26) becomes easier to process by introducing different types of DPs, ORCs may be easier to process for the same reason.

Grammatical Hierarchy Hypothesis

In this experiment, we explore another potential explanation for why the object/subject asymmetry is reduced or eliminated when one of the DPs is a pronoun—the person-animacy hierarchy. Under one view, Chamorro obligatorily aligns grammatical person and grammatical role in canonical transitive clauses (Aissen 1999), and in an ORC like (28), the 2nd person pronoun is in subject position:

(28) The cook that you helped _ quit work after a month.

The processing of these ORCs should be facilitated because 2nd person is highest on the hierarchy and therefore a more optimal subject than a full DP like *the nurse*.

Previous research has shown that sentence comprehension is aided when there is alignment between linguistic hierarchies. Christianson & Cho (2009) investigate the interpretation of *pro* when the only context available for the listener is sentence-internal. They consider the Algonquian language Odawa. Odawa has two verb forms, called the direct and the inverse, which are used to focus the agent and patient of a transitive clause, respectively. The focused DP is the proximate, while the other DP in the clause is the obviative. The proximate can be *pro*. In direct constructions, the proximate is the agent while in inverse constructions, the patient is the proximate. Thus, there are three hierarchies operating in Odawa:

Thematic hierarchy: agent >patient
Animacy hierarchy: human >less-human
Obviation hierarchy: proximate >obviative

The authors used a self-paced listening methodology followed by a picture verification task. They found that the highest accuracy in the picture verification task was in the condition with an animal agent that was an overt argument while the other argument was *pro*. This condition collapses over the inverse and direct forms, but we can understand the accuracy advantage in terms of the satisfaction the largest number of hierarchies in this condition. In the direct form, the animacy hierarchy is satisfied because *pro* is human. In inverse sentences, having a human patient *pro* satisfies both the animacy hierarchy and the thematic hierarchy because inverse sentences focus the patient. This finding suggests that comprehension is aided when there is alignment between animacy, thematic, and obviation hierarchies.

If the degree of processing difficulty is correlated with the number of hierarchies violated, then we predict that ORCs with 2nd person pronoun subjects should be easier

to process than ORCs with full DP subjects. This is because both the 2 >3 and pronoun >non-pronoun hierarchies are satisfied when the subject is a 2nd person pronoun, making 2nd person the most optimal subject choice according to the person-animacy hierarchy. Full DPs are less optimal subjects, and thus parsing is not facilitated when these DPs are subjects.

2.4 Experiment

In the previous sections, we saw evidence in English for effects of the animate >inanimate and the pronoun >non-pronoun sub-hierarchies of the person-animacy hierarchy. This invites us to ask, do we find any any evidence for the 2 >3 hierarchy in English?

Chamorro Sub-Hierarchy	English
animate >inanimate	passive production rates
pronoun >non-pronoun	elimination of the S/O asymmetry
2nd >3rd	???

In the present study we explore 3rd person singular pronouns in addition to 2nd person pronouns to see if there is any evidence for 2nd person >3rd person hierarchy effects in English and to explore possible explanations of the reduction of the subject/object asymmetry when DP_2 is a pronoun.

Design

We look for 2 >3 effects in the relative clause environments where the processing of ORCs is facilitated by having a 1st or 2nd person pronoun subject. Previous research on the subject/object asymmetry in relative clauses with pronouns has been focused on the local persons only. This may be because the local person pronouns do not require any previous discourse to license them, which makes them easier to present in experimental settings. If we want to present a sentence with a 3rd person pronoun intended to refer to a different entity than the other DP in the sentence, and we want to ask a question about this pronoun, some discourse must be provided. We address this issue in our

experimental design.

The experiment consists of a 2 x 3 factorial design. The factors manipulated were DP type (2nd person, 3rd person, full DP) and relative clause extraction site (subject, object).

Sample Item Set

Full DP

(a) SRC The nurse that _ welcomed the mechanic with a smile ran a marathon during the month of July.

(b) ORC The nurse that the mechanic welcomed _ with a smile ran a marathon during the month of July.

2nd Person

(c) SRC The nurse that _ welcomed you with a smile ran a marathon during the month of July.

(d) ORC The nurse that you welcomed _ with a smile ran a marathon during the month of July.

3rd Person

(e) SRC The nurse that _ welcomed him with a smile ran a marathon during the month of July.

(f) ORC The nurse that he welcomed _ with a smile ran a marathon during the month of July.

Each item set has a modifier separating the embedded VP from the matrix VP to assess whether measures on the matrix verb from previous experiments are spillover measures from the embedded VP.

Procedure

The experiment was administered in Ibex using a self-paced reading methodology. There were 41 participants in the study, aged 18 to 60. Eighteen of these participants were recruited through Facebook and the rest were given course credit for their participation. There were 24 experimental item sets and 72 fillers, and a Latin square design was used to ensure that each participant saw different items.

Each trial was introduced by a transition "Your friend {John/Mary} tells you that" on the screen, followed by the sentence on a new screen without capitalization to indicate that the sentence was indirect discourse. Transitions were counterbalanced for *John* and *Mary*, and the 3rd person pronouns matched the DP in the transition in gender. The 3rd person pronouns were counterbalanced for gender using a scale from a gender stereotyping survey. Pronouns were selected to contrast in gender with the other DP in the sentence to minimize potential interference between the pronoun and DP due to the gender feature. Following each sentence, participants had to answer a comprehension question. These asked about the roles associated with the matrix verb, the roles associated with the embedded verb, and the modifiers in the sentences. An example comprehension question is *Was it with a smile that the nurse welcomed him?* Participants were given feedback on the incorrect trials.

Predictions

If the difference between pronoun and DP conditions found in previous studies is due to the fact that a pronoun has a different surface from than a full DP (Bever 1974), then there should not be a difference between 2nd and 3rd person pronouns. Accounts appealing to grammatical hierarchies and similarity-based interference both predict a weaker asymmetry between 2nd person RCs ((c) vs. (d)) than 3rd person RCs ((e) vs. (f)) but for different reasons.

If the difference between pronoun and DP conditions found in previous work is

driven by similarity-based interference, then 3rd person pronouns and full DPs have a dimension of similarity—the 3rd person feature—which may introduce interference in the 3rd person condition. Therefore, facilitation in ORCs compared to SRCs in the 3rd person condition should be less than the facilitation of ORCs compared to SRCs in the 2nd person condition because the 2nd person pronoun does not share any features with the other DP.

If a soft grammatical constraint like the person-animacy hierarchy is responsible for the asymmetry between SRCs and ORCs, we still expect the difference between 3rd person ORCs and 3rd person SRCs to be greater than the difference between 2nd person ORCs and 2nd person SRCs. ORCs in both the 2nd person and 3rd person conditions put a pronoun in subject position. According to the hierarchy, when a 2nd person pronoun is in subject position, both the pronoun >non-pronoun ranking and the 2nd person >3rd person ranking are satisfied. When the 3rd person pronoun is a subject, only the pronoun >non-pronoun ranking is satisfied. Thus, 2nd person pronouns are the most optimal subjects, and therefore facilitation in an ORC should be greatest in the 2nd person condition.

	2nd Person ORC	3rd Person ORC
	The nurse that you welcomed...	The nurse that he welcomed...
Pronominal Surface Form	no difference between 2nd and 3rd	no difference between 2nd and 3rd
Grammatical Hierarchy	according to the person-animacy hierarchy, 2nd person is the most optimal subject—2 >3	3rd person subjects are less optimal than 2nd person subjects—2 >3
Similarity-Based Interference	full DPs and 2nd person do not have overlapping person features—2 >3	full DPs and 3rd person pronouns share a 3rd person feature which may generate interference—2 >3

Table 2.1: Predictions Summary

In the present design, both the similarity-based interference hypothesis and the grammatical hierarchy account make the same predictions, and if this prediction is borne out, we will not be able to adjudicate between these hypotheses. An alternative design to probe this further is proposed in the last section.

2.5 Results

Reading time and comprehension accuracy data were modeled with mixed-effects regression with Helmert contrasts for DP type, first comparing the full DP condition to the pronoun conditions, and then 2nd person to 3rd person. It is worth noting that we made a methodological decision in our reading time analysis to compare reading times by the verbs in the SRC and ORC conditions, despite the fact that the verbs do not

have the same ordinal positions within the sentences in the respective conditions. While this may seem like a minor point, there is much variation in the literature as to how SRCs and ORCs are compared, with many authors choosing to compare words based on ordinal position in the sentence. However, it is not obvious that comparing a DP in one condition to the embedded verb in the other condition is the proper comparison to make.

Staub (2010) uses eye-tracking to analyze the subject/object asymmetry. He provides evidence that processing difficulty arises at both the ORC subject and the embedded verb, and these manifest as regressive saccades and higher first-pass reading times, respectively. Since processing difficulty in these two regions leads to different behavior, we have reason to believe that processing at the ORC subject and embedded verb are different and warrant comparisons by embedded verb, not ordinal position in the sentence.

Reading Times

The following graph compares word-by-word reading times for each condition. Boxed regions indicate the critical embedded verb. Comparing the boxed regions, we find that that the greatest difference between SRC and ORC is in the full DP plot.

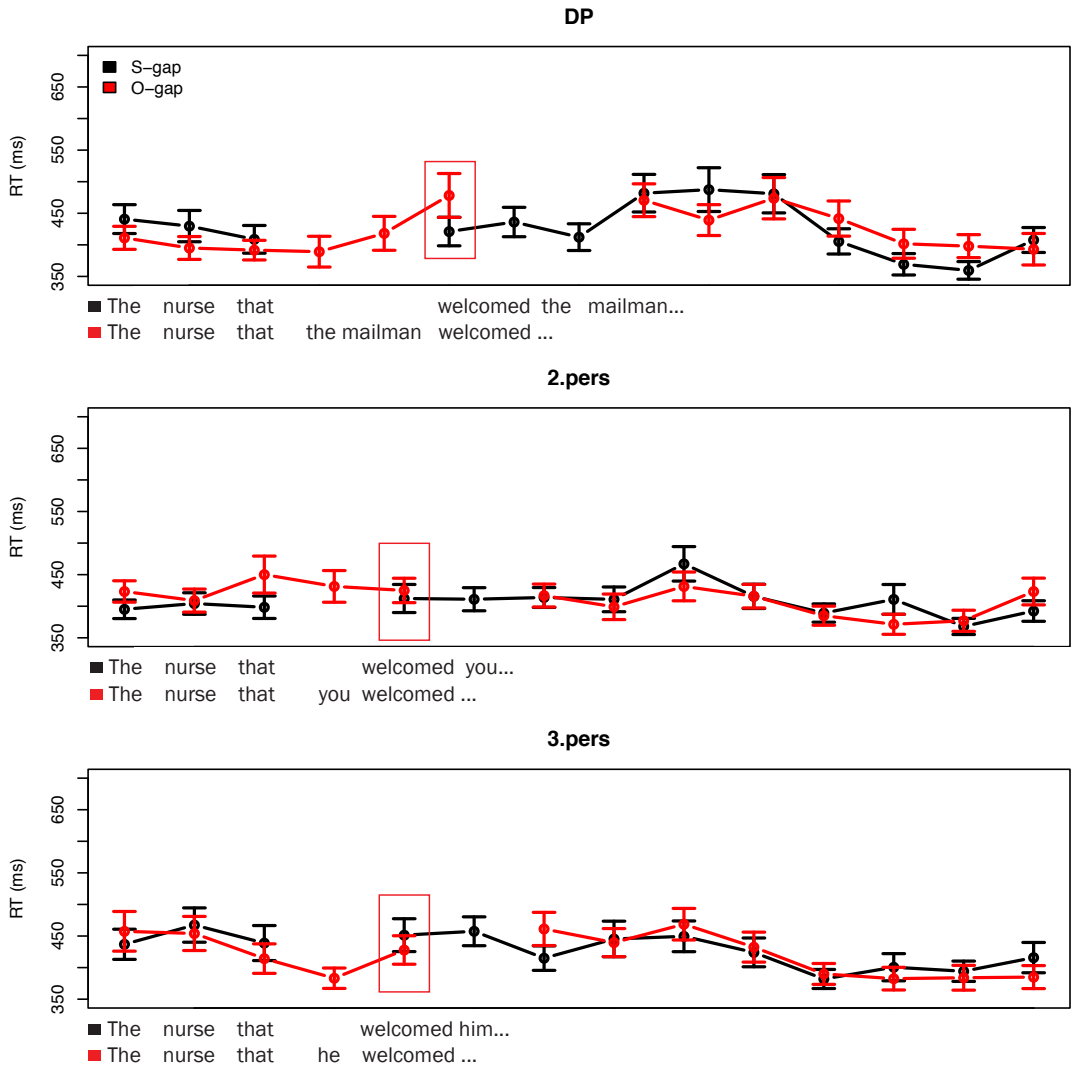


Figure 2.2: Overall Reading Times

Zooming in on the critical region, we find that ORCs in the pronoun conditions are read at the same rate as their SRC counterparts, while there is a significant slowdown for ORCs in the full DP condition compared to SRCs.

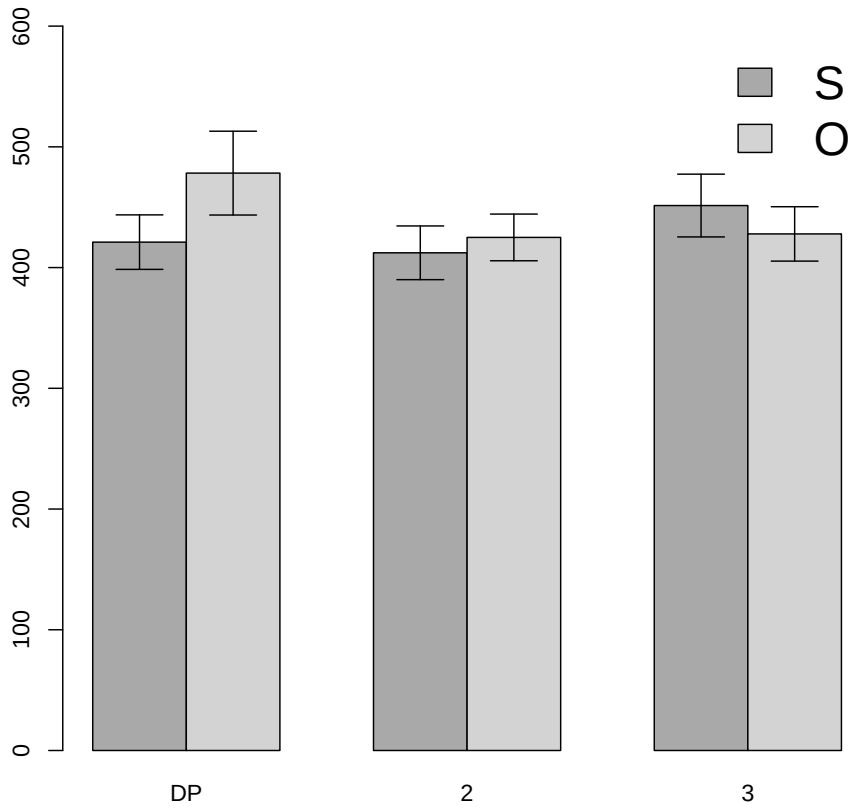


Figure 2.3: Reading Times at Embedded Verb

The interaction between DP type and gap type comparing DP vs. pronoun conditions is significant ($p = .04$). In this measure, we replicated the finding of Gordon et al. (2001) that object extractions cause reading times to be longer when both arguments in the relative clause are full DPs.

The subject/object difference is present in the full DP condition, but the pronoun conditions pattern alike in alleviating this difference. This suggests that there is not a difference between 2nd and 3rd person pronouns at the embedded verb. Thus, we replicate the findings of previous studies that local person pronouns eliminate the subject/object asymmetry, and we find that 3rd person pronouns pattern like 2nd person

pronouns in this regard. This novel finding supports the hypothesis that the surface form differences between pronouns and full DPs is driving the S/O asymmetry found in previous studies.

Comprehension Question Accuracy

The following graph shows the mean accuracy for each condition:

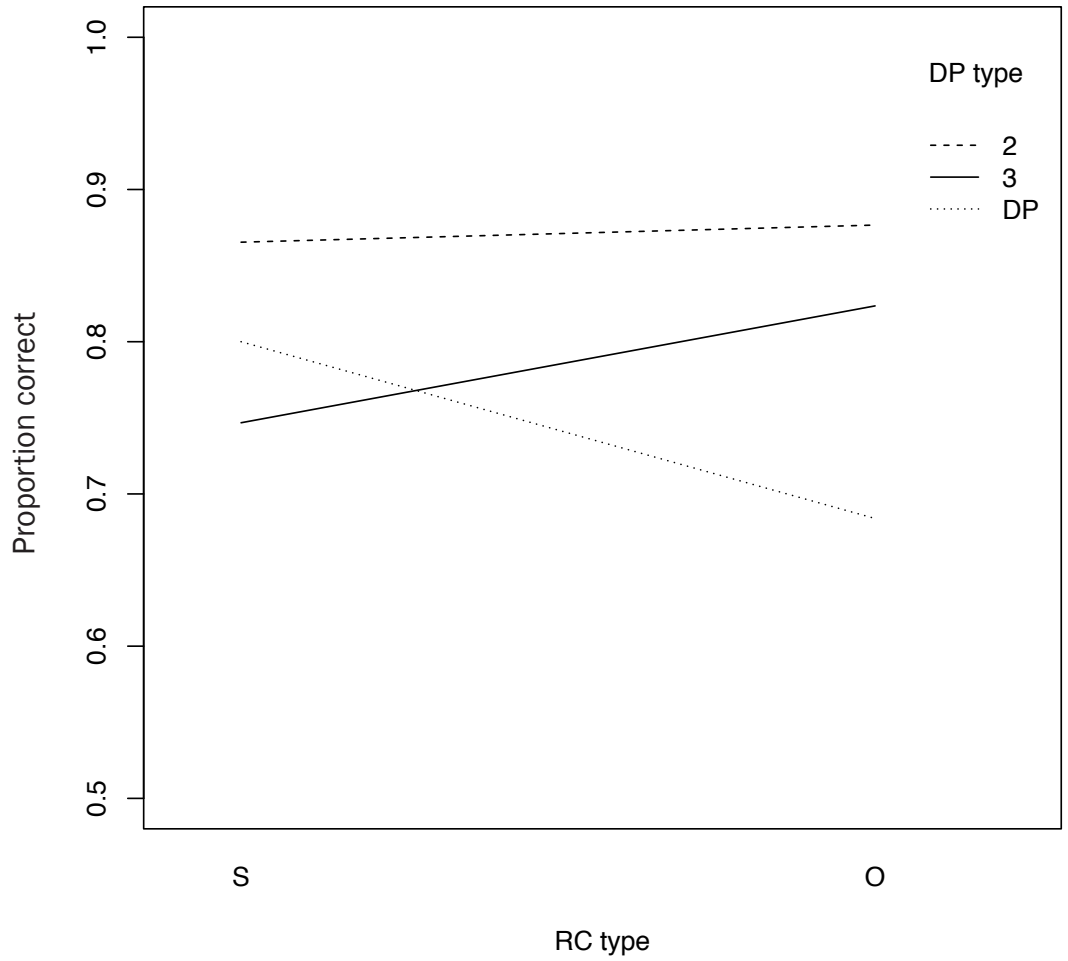


Figure 2.4: Comprehension Question Accuracy

As in the reading time data, the interaction between DP type and gap type comparing DP vs. pronoun conditions is significant ($p = .01$). This again suggests that the asym-

metry is present only in the full DP conditions. Here, in contrast to the reading time data, there is also a significant main effect of 3rd person vs. 2nd person ($p = .003$).

In contrast to the reading time data, we see that 3rd person does not pattern like 2nd here. Accuracy is lower in the 3rd person conditions than in the 2nd person conditions. Contrary to our expectations that ORCs should be harder to process than SRCs, we find that numerically, ORC comprehension is facilitated in the 3rd person condition, though this difference is not significant despite being persistent numerically. In other words, this difference has not reached significance despite being present at every sampling in the collection of this data. A power analysis suggests that many more participants are needed for this difference to reach significance if it is a small effect size.

2.6 Discussion

In the reading time data, we found the expected subject/object asymmetry on the embedded verb for the DP conditions, but not for the pronoun conditions. This finding supports the pronominal surface form hypothesis, and suggests that there is not an advantage for 2nd person pronouns compared to 3rd person pronouns in English.

We found a difference in overall accuracy between 2nd and 3rd person pronouns in the comprehension data. The fact that we find different results for the reading data and comprehension data for 3rd person is likely due to the fact that comprehension accuracy is an offline measure while reading time at the embedded verb is an online measure. Comprehension measures the accessibility of representations in memory, while reading time measures the construction and integration of these representations.

It is important to note that we did not have predictions about offline measures for the similarity-based interference or grammatical hierarchy accounts. Both of these accounts make predictions only about online processing. The finding that there is a decrement in accuracy in the 3rd person conditions compared to the 2nd person conditions suggests that person features contribute to the accessibility of representations

in memory (Ariel 1991). It is not surprising that 2nd person has a higher accuracy than 3rd person, since the referent of a 2nd person pronoun is given since it refers to the addressee, while the 3rd person pronoun has a referent that must be located elsewhere in the discourse context, outside of the dyad of speaker and addressee.

Case Marking

The numerical decrement in accuracy for the SRCs in the 3rd person condition was not predicted. One difference between 2nd person and 3rd person pronouns is that 3rd person pronouns are case-marked. It is possible that the case-marking on 3rd person pronouns is contributing to the numerical accuracy increase in ORCs. It has been argued that case-marking contributes to processing of pronouns (Warren & Gibson 2005) and the similarity of full DPs in languages which have case markers (Kwon et. al 2010).

Warren & Gibson (2005) look at object-extracted clefts, and demonstrate in their pronoun-pronoun condition that clefts with 2nd and 1st person pronouns do not show interference at the embedded verb. They account for this with case-marking on the 1st person pronoun, since case-marking is a cue that can be used to determine the grammatical roles of pronouns. These pronouns do show interference in the comprehension question accuracy, however, which suggests that the advantage of case-marking is negated by interference at the level of conceptual representation in this measure. In the present study, the condition where we may be seeing an advantage of case-marking contains a DP and a pronoun. It is possible that we could find effects of case-marking in the comprehension accuracy in this study because we are looking at comparisons between DP types different than those considered by Warren & Gibson.

If case-marking is a factor which contributes to the difference in accuracy between 3rd person SRCs and 3rd person ORCs, why is it that the ORC is easier than the SRC? As has been discussed, the ORC puts the 3rd person pronoun in subject position. Perhaps the facilitation of having the pronoun in subject position is increased when combined with the subject case-marking cue. We could think of the overall decrement

associated with the 3rd person conditions as being improved in the ORC case because of the additional cue.

If the difference between 3rd person SRCs and ORCs reaches significance in the comprehension accuracy data, we could interpret this as evidence for a NOM >ACC ranking. The 3rd person pronouns are case-marked whether they are objects or subjects, and this marking contributes to the representation of these DPs in memory. The fact that the ORCs have higher comprehension than the SRCs in the 3rd person condition could be interpreted as evidence that nominative case-marking, the case of subjects, is privileged over the accusative case-marking, the case of objects, in retrieval of these memory representations.

In 2nd person cases, we do not see a difference between SRCs and ORCs. This could be because 2nd person pronouns are not case-marked, or we would expect to see the ORCs be more accurate than SRCs for the same reason we see in the 3rd person cases. However, it could also be the case that we are observing a ceiling effect in the 2nd person cases. Since there aren't any overlapping features between 2nd person pronouns and full DPs, these conditions do not have any interference and therefore have maximum accuracy. Even if there were case-marking, it might not boost ORC accuracy in these cases.

Pronominal Surface Form

The subject/object processing asymmetry in relative clauses can be eliminated by manipulating DP_2. As discussed, much research has shown that the subject/object asymmetry is eliminated when DP_2 is a pronoun, and this finding was replicated in the current study for both 2nd and 3rd person pronouns. From this, we argued that it is the surface form of the pronoun that alleviates the processing asymmetry, and these pronouns are not subject to similarity-based interference. While we can attribute findings of this study to surface properties of pronouns compared to full DPs, we have not addressed whether the length of the word is the surface property that is driving the

reduction in processing difficulty.

Gordon et al. (2004) offer evidence that suggests that the subject/object asymmetry is not reduced based on the length of DP_2. In this study, the authors tested a variety of DP types—generics, definites, indefinites, names, quantifiers, and pronouns. They found that of these types, only the names, pronouns, and quantifiers reduced the subject-object processing asymmetry.

Based on the types of items which reduce the subject/object asymmetry, it appears that length—either by number of words or letters and syllables—is not responsible for the difference. Although the DP types which reduce the s/o asymmetry are each one word long, generic DPs which do not have a determiner did not reduce the asymmetry. While names and pronouns have short surface forms in both syllable and letter count, quantifiers have more syllables and letters, and yet they pattern like names and pronouns in reducing the asymmetry.

Similarity-Based Interference

Though we can establish that length of pronouns is not what makes them easier to process, Gordon et. al's findings do raise further questions about what quantifiers, pronouns, and names have in common such that they eliminate the subject/object processing asymmetry. All of these DP types have different surface forms than definite descriptions, consistent with the results of the present study. However, it is likely not just the surface forms of these DPs, but also the types of representations that are constructed for them in memory.

It is worth noting that the types of DPs which reduce the subject/object asymmetry all share the 3rd person feature. However, it is not unreasonable to think that the effect of having an overlapping 3rd person feature would be greater for pronouns than for quantifiers or names. This is because pronouns are variables which look for a referent with matching gender, number, and person features, while names map onto particular individuals and quantifiers have a more semantically complex representation

such that we might expect the 3rd person feature to be a less salient dimension for these other DP types than for pronouns.

If we expected to find an effect of the 3rd person feature for pronouns, why is it that 3rd person pronouns do not induce similarity-based interference despite having this overlapping feature with definite descriptions? It could be the case that this feature is not enough to induce interference. However, there is one other factor not yet discussed here—the frequency of ORCs with pronominal subjects. Experienced-based models predict that relative clause processing is facilitated by frequency. Since ORCs with pronominal subjects are more frequent than ORCs with full DP subjects, they are easier to process.

Experience-Based Models

The following figure from Reali & Christiansen (2006) shows the distribution of relative clauses containing pronouns:

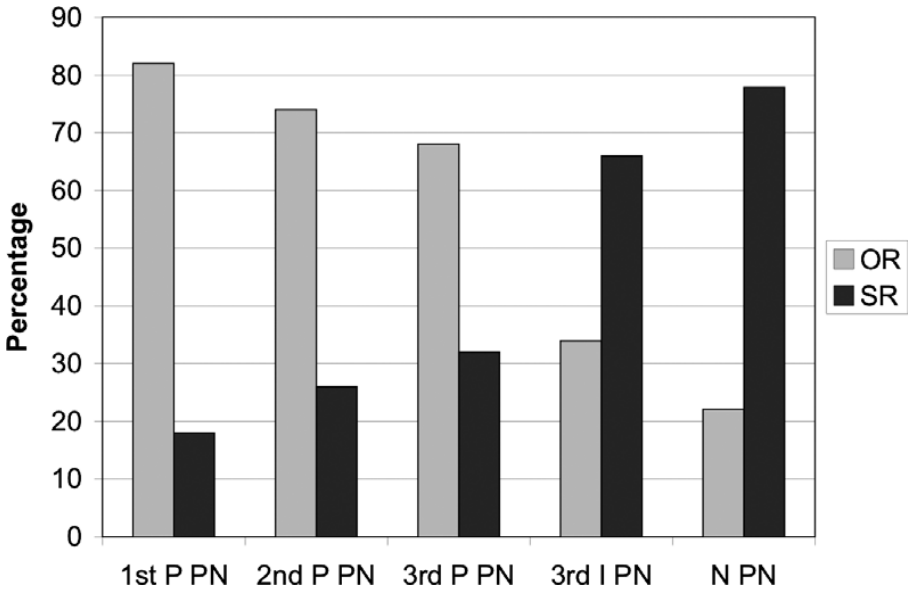


Figure 2.5: Relative Clause Frequency

This figure shows that the distribution of relative clauses with 2nd and 3rd person animate pronouns (3rd P PN and 2nd P PN) is very similar, and perhaps this is why we do not find a difference in the relative processing of 2nd and 3rd person pronouns. Further support for this hypothesis comes from the Reali & Christiansen's experiment using impersonal 3rd person pronouns (*it*). A surface form account and an interference account would predict that these ORCs with *it* should pattern like other pronominal ORCs in alleviating the subject/object asymmetry. However, this is not the case. An experience-based model predicts that these ORCs should not be facilitated compared to SRCs with *it* because they are so much less frequent, and this is indeed what was found in the experiment.

Discourse Models

Experienced-based models do not explain whether structures are more frequent because they are easier to process or if they are easier to process because they are more frequent. Fox & Thompson (1990) argue for a discourse-based analysis of the frequency of pronominal ORCs. Based on the function of relative clauses in discourse, they argue that when an ORC has an inanimate head noun, it is very like to have a pronominal subject. This is a result of the way in which people talk about inanimate objects. If a person talks about an inanimate object, they want to express a human's relation to it. Thus, the inanimate object becomes anchored by relation to a given human discourse referent expressed as a pronoun.

Another theory of discourse not specific to relative clauses is Centering Theory (Groz et. al 1995). According to this theory, coherence of a text is increased by having the same subject in many adjacent sentences. This theory may provide an explanation for the 3rd person ORC comprehension accuracy advantage. The embedding context for each trial of the experiment was *Your friend* {*John/Mary*} *tells you that....* In this embedding sentence, *John* is the subject. In the ORC condition, the 3rd person pronoun is also the subject. Given that sequences of sentences with the same subject are

preferred, the ORC condition provides a more natural continuation of the embedding context because of the syntactic parallelism between *he* and *John*. The application of Centering Theory will be explored in more depth in the next chapter.

2.7 Conclusion

The present study explores the differences between 2nd and 3rd person pronouns with respect to the subject/object processing asymmetry in relative clauses. Both the reading time and comprehension question accuracy indicate that the subject/object asymmetry is uniformly alleviated in the pronoun conditions. If English respected a soft version of the 2nd >3rd hierarchy, we would expect incremental parsing to be facilitated more in the 2nd person condition than in the 3rd person condition. This prediction was not borne out. The difference between the 2nd and 3rd person conditions in overall accuracy suggests that person features contribute to the accessibility of representations in memory (Ariel 1991). Given that 2nd person referents are given, this finding is not surprising.

Some open questions remain. What is driving the ORC advantage in the 3rd person condition? If it is case-marking, the fact that we see a numerical trend for higher accuracy in the ORCs for 3rd person suggests that the alignment of grammatical hierarchies could boost memory representations, since there is a higher accuracy when the pronoun subject also has subject case-marking. Future work could explore a possible account related to Centering Theory and the interaction of discourse-level parsing heuristics in individual sentence trials to account for this small yet persistent difference by manipulating the syntax of the embedding context.

At the outset, we asked if language-specific constraints are instantiations of universal linguistic preferences. In the case of the $2 > 3$ hierarchy, we did not find evidence in support of this idea. This finding does not, however, invalidate the claim of Bresnan et al. (2001). Their claim was based on aggregated statistical preferences in corpora,

which is a very different measure than those associated with online processing and comprehension.

We can also consider some aggregated corpora statistics from English argument pairings in transitive clauses in the New Testament, repeated here from Chapter 1:

SUBJECT \ OBJECT	2 PERS	3 ANIM *pron.*	ANIM	INANIM	
2.PERS	1	9	16	52	78
3.P ANIM *pronoun*	15	43	28	74	160
ANIMATE	35	40	27	101	203
INANIMATE	2	3	0	15	20
	53	95	71	242	N =461

Figure 2.6: Actual Distribution

SUBJECT \ OBJECT	2 PERS	3 ANIM *pron.*	ANIM	INANIM	
2.PERS	9.0	16.1	12.0	40.1	77.2
3.P ANIM *pronoun*	18.4	33.0	24.6	84.0	160
ANIMATE	23.3	41.8	31.3	106.6	202.9
INANIMATE	2.3	4.1	3.1	10.5	20
	53	95	71	241	N =461

Figure 2.7: Predicted Distribution Under Independence Assumptions: (S=A, O=B) = P(A | S) x P(B | O)

Considering these distributions, we find significantly more animates acting on 2nd person than expected by chance. We also find marginally fewer 2nd person pronouns acting on 3rd person pronouns than expected by chance. These patterns suggest that the animate > inanimate hierarchy is respected more than the 2 > 3 hierarchy in English. This could be due to differences in ranking of the various components of the person-

animacy, consistent with an OT theory of subject choice (Aissen 1999). These findings suggest that online measures and corpora statistics can vary, and that hierarchies can also vary in their ranking. This makes the prediction that a language can have a person hierarchy ranked higher than an animacy hierarchy, but corpus data from more languages is needed to evaluate that prediction.

2.8 Future Work

Filled-Gap Effect

A corollary to the subject relative clause preference is that all fillers are initially posited as subjects. Perhaps the subject/object asymmetry is about the ease of reanalysis. Do we find a difference between DPs and pronouns with respect to the filled-gap effect?

If it is the case that the surface form is driving the difference between DPs and pronouns in the present study, then pronouns are distinct enough that they should act as a greater error signal and cause a larger filled-gap effect. Conversely, if the parser independently prefers pronominal subjects, then it will be easier to resolve the unexpected filled gap and pronouns will lead to a smaller filled-gap effect. Based on the frequency of ORCs with pronominal subjects, an experience-based model predicts that there should be a smaller filled-gap effect for these ORCs.

Example Item Set

(a) Susan asked whether, until recently, the baker has been fighting with the plumber.

(b) Susan asked which plumber, until recently, the baker has been fighting with _.

(c) Susan asked which plumber, until recently, you have been fighting with _.

(d) Susan asked which plumber, until recently, he has been fighting with _.

Though we did not find any evidence in favor of the 2 >3 hierarchy in relative

clause processing, as alluded to above, this design could adjudicate between similarity-based interference and grammatical hierarchy accounts. Because similarity-based interference is about the integration of representations in online structure building, at the point of the filled-gap, these DPs are just being encountered and are not yet integrated into the structure. Thus, if we find evidence for the 2 >3 hierarchy in this measure, we can attribute it to a grammatical hierarchy account because similarity-based interference has not yet come into effect. Additionally, this design could help us differentiate the predictions of an experience-based account with those of a grammatical hierarchy account. Given that the distribution of 2nd and 3rd person ORCs is quite similar, we might not expect a a difference of approximately 5% to have a large effect on the relative processing of 2nd person pronouns compared to 3rd person pronouns. If we find a large advantage for 2nd person pronouns in this measure, we could attribute it to a grammatical hierarchy account.

Chapter 3

The Distribution of Nominal Expressions in Personal Narratives

3.1 Introduction

At the end of the experiment discussed in the previous chapter, we were left with some evidence in favor of the Accessibility Hierarchy (Ariel 1991) and Centering Theory (Grosz et al. 1995). Recall that comprehension question accuracy was highest in the conditions with 2nd person pronouns, and we attributed this to the fact that 2nd person pronouns do not introduce discourse referents, but rather their referent is given in the dyad of speaker and addressee. We also found that comprehension accuracy in the 3rd person pronoun condition was elevated in ORCs compared to SRCs, and hypothesized that this could be related to discourse coherence as predicted by Centering Theory. In the current chapter, I explore the predictions of Centering Theory and the Accessibility Hierarchy in a corpus study of English weblogs, a type of discourse which approximates spoken language.

Much of the data in the human sentence processing literature comes from experimental trials consisting of sentences in isolation. But in practical use, language consists of many utterances strung together to form coherent discourses. This unit of linguistic structure could provide insight into the phenomena we observe in human sentence processing. If the predictions of linguistic theories about the structure of discourse and

the realization of nominal forms are on the right track, these preferences for coherence and the use of pronouns could be applied to human sentence processing to account for observations about the processing of pronouns in self-paced reading studies, for instance.

In this chapter, we consider the distribution of nominals in personal narratives, a genre which has properties of spoken language and is therefore a good candidate for approximating natural speech. Though weblogs are written and therefore are subject to revision in a way that spoken language is not, these discourses are examples of spontaneous descriptions of situations. In psycholinguistic research, experimenters use tasks to elicit descriptions of this type. By testing the predictions of Centering Theory and Accessibility, we can evaluate the contributions of syntactic features and discourse features in accounting for the distribution of nominals in natural language.

Centering Theory

Previous research has shown the distribution of DP types in natural language is sensitive to the discourse context. One theory that attempts to capture this is Centering Theory Grosz et al. (1995), which forms the foundation of entity-based coherence models in computer science Elsner et al. (2007).

Centering Theory was originally proposed to account for the incoherence of certain discourses like the following:

(29) Susan gave Betsy a pet hamster.

(30) She reminded her that such hamsters were quite shy.

(31) She told Susan that she really liked the gift.

In Centering Theory, the resolution of pronouns is based on the syntactic position and pronominal form of preceding DPs. The final sentence (31) is incoherent because the object of (30), which refers to *Betsy*, is now the pronominal subject. In both (29) and (30), the subject is *Susan*, and in (30) this subject is a pronoun. The configuration in (31) increases the load on the hearer to disambiguate the pronoun because the hearer relies

on cues such as syntactic position of a DP across utterances. Based on discourses like this one, Grosz et al. (1995) propose a set of constraints and rules about the distribution of nominals in discourse.

According to Grosz et al. (1995), a discourse is organized into segments which consist of utterances. Each utterance introduces a set of CFs, or *forward-looking centers*, which essentially correspond to discourse entities. For each utterance that introduces a set of CFs, the next utterance will contain a CB which is the highest ranking entity in the set of CFs from the preceding utterance.

Constraint 1

All utterances of a single segment except for the first have exactly one CB.

Rule 1

If any CF is pronominalized, the CB is.

Entity-based approaches to coherence have systematically ignored the component of Centering Theory related to pronouns since it is not feasible for multi-document summarization tasks, for instance.

Accessibility Hierarchy

Ariel (1991) describes a hierarchy of nominal types which are ordered with respect to their Accessibility. Accessibility is derived by the interaction of several factors, including the informational content and uniqueness of an expression and an entity's salience in the discourse.

full name <definite description <last name <first name <pronouns

The Accessibility of a given DP will be modulated by mentions in the discourse. For instance, the first mention of a DP might consist of a full name, but subsequent mentions will likely be pronominalized.

3.2 Related Work

In the original proposal of Centering Theory, the central notions were never formally defined, but rather presented as open to further investigation. In a computational study, Poesio et al. (2004) determine the parameters which most closely capture the distribution of nominals in museum descriptions and patient leaflets. For this project, we adopt their method for testing the predictions of Centering Theory and model our annotation scheme after theirs.

3.3 Method

Annotation

For this task, we annotated 40 personal narratives from the collection used by Gordon and Swanson (2009) for nominals and their properties, as well as features of the discourse[1]. We are interested in this type of discourse because it is a more natural use of language than the texts analyzed by Poesio et al. (2004). Museum object descriptions and patient leaflets are more technical forms of writing, and do not necessarily parallel spoken speech, whereas personal narratives a closer approximation to everyday speech.

Additionally, museum descriptions and patient leaflets are likely to be poor candidates for evaluating predictions of the Accessibility Hierarchy. When referring to artists, last names are used by convention, whereas people discussed by the narrator of a weblog will probably not be referred to in this way.

We used the annotation guide supplied by Poesio et al. (2004) to determine the features we will annotate. For nominals, we identify their form (focusing on names, pronouns, and definite descriptions) and their grammatical features (syntactic position, animacy, number, person).

[1]I thank Valery Vanegas for assistance with annotating a portion of these weblogs.

Discourse Segments

For discourse level annotations, we consider discourse segments and utterances. Discourse segments are groups of utterances which together form a coherent section of the discourse. An example of a segment break is given here:

(u1) *Thank goodness I'm drinking some hot coco and nibbling on sugar free Girl Scout cookies and while watching this stuff.* (u2) *When I was a child, I yearned for snow days to escape the humdrum and loneliness of school.*

In (u1), the narrator is describing some actions that they are currently doing. In (u2), the narrator is talking about events that happened long ago in their childhood. The reader can perceive a shift in the discourse between these two utterances. While identification of discourse segmentation in this case is very clear due to the action/backgrounding distinction, some cases are harder to identify.

Utterances

An utterance is a clausal unit such as a main clause or a relative clause. We identified three types of utterance in our data–main clauses, relative clauses, and adjuncts.

An issue that immediately presents itself in adopting this annotation scheme is the one of hierarchical structure. In Centering Theory, where utterances are defined as u_n and u_{n+1}, how do we count embedded clauses? For the research presented here, we annotate all clauses linearly. This means that in a sentences like *We keep a bowl for Benjamin to drink out of*, we treat the matrix clause as an utterance and the relative clause as an utterance.

We could have treated utterances of this type as singleton clauses, but this treatment fails to capture parallelism relationships between matrix and embedded clauses within a complex sentence. Such a treatment, however, does make the same predictions about the centers of utterances immediately following complex sentences.

Another place where the hierarchical structure issue comes about is in complex

sentences in which the matrix verb selects a sentential complement. For example, in a sentence like *I think she's still getting used to us*, there is a matrix verb *think* and an embedded clause *She's still getting used to us*. We extended our treatment of embedded relative clauses to cases like these as well, treating the matrix and embedded clauses as separate clauses, though it is unclear whether this was the best treatment of such cases.

In this sentence, there is some parallelism between the matrix and embedded clause if we count *us* as a realization of *I*. Treating these as separate clauses highlights this connection between the extradiegetic opinions of the narrator and the narrator's actions as a participant in the story. The narrator could have said *I think she likes it* where there is no realization of of the matrix subject in the embedded clause. Both types of clauses exist in our data set but given that we are looking at personal narratives with lots of extradiegetic clauses, we wanted to highlight instances of parallelism of arguments between matrix and embedded clauses.

Backward-Looking Centers

We also identify by hand the CB for each utterance in the narrative. Consider the following discourse excerpt:

(u1) *His bowl has become a very popular site.* (u2) *Throughout the day, many birds drink out of it and bathe in it.* (u3) *Squirrels also come to drink out of it.*

In (u1), the subject of the sentence is *his bowl*. In both (u2) an (u3), the pronominalized element is *it*, which refers back to *his bowl*. Since the only element of the CF set in (u1) which is realized in (u2) is *it*, this is the CB in both (u2) and (u3).

When more than one of the CFs in an utterance is realized in the next utterance, determining the CB becomes more difficult:

(u1) *I got completely trashed for my birthday and asked one of my friends out* (u2) *but he of course didn't take me seriously* (u3) *because I was drunk.*

In this example, two CFs in (u1) are realized in (u2). Since *I* is the subject in (u1),

it is the highest-ranking Cf. Therefore, *me* is the Cb in (u2) by virtue of realizing *I*. Again, because *I* is the subject in (u3), it is again the Cb.

We allow indirect realizations of forward-looking centers, meaning that we allow references like bridging, where an object like *the vase* can be realized by a nominal which refers to a subpart of it such as *the handle*. It was pointed out to me by Ellen Riloff (p.c.) that bridging can go in the opposite direction, where a DP is used to refer to an entity, and then a broader DP is used in the bridging reference. For example, in this data, we found the definite description *the frog* was used in one utterance, and then in the next utterance, *the amphibian* is used to refer to the same frog.

Complex DPs raise some issues, particularly for the genre of personal narrative. Since we are looking at 1st person narratives, plural 1st person pronouns are frequent in the data. If the narrator uses *I* in an utterance, we considered *we* to be a realization of this Cf in the subsequent utterance. Similarly, we considered *I* to be a realization of *we* in a previous utterance.

Another type of complex entity affected by indirect realization are possessed DPs. Considering an entity like *her car*, would we count *she* as a possible indirect realization? For consistency, we adopted the same treatment for DPs of this type as we did for plurals and bridging, in that we permitted singleton entities within a larger entity to count as an indirect realization of that entity, and conversely larger entities to count as realizations of singleton entities.

In this research, we made some decisions about how to annotate relative clauses which do not necessarily conform to all analyses of relative clauses. According to some theories, relative clauses always have a backward-looking center that is the trace of the relativized entity. For this project, we annotate the relative clauses ignoring traces as possible candidate Cb's.

In a previous version of our annotation scheme, we did not include null Cb's. However, we decided to make such a category based on examples like the following:

I checked my bumper where she hit me. Scraped up.

It is clear that *scraped up* has some sort of null implicit argument, and so we identify null arguments as possible Cb's.

Some Issues

The following example comes directly from Grosz et al. (1995), with the lists of Cf's annotated by the authors:

(a) John has been having a lot of trouble arranging his vacation.

(b) He cannot find anyone to take over his responsibilities, (he = John) Cb = John; Cf = {John}

(c) He called up Mike yesterday to work out a plan, (he = John) Cb = John; Cf = {John, Mike}

According to our annotation scheme, these example sentences are inherently confounded in why they predict that backward-looking centers that they do. Under a view where the sentences in these examples are monoclausal, *John* is the backward-looking center of (b) because he is the only Cf in (a). Similarly, *John* is the Cb in (c) because he is the highest-ranking Cf introduced in (b).

Since we segment all clauses and acknowledge null centers, our segmentation would find two clauses in (b), *he cannot find anyone* and a relative clause *to take over his responsibilities*. We also consider a possessed DP like *his responsibilities* as a possible indirect realization of *John*. Therefore, because *he* is the highest-ranking Cf in the first utterance, *his responsibilities* makes *John* the Cb in (b), which makes *he* the Cb of (c) as well.

Predictions

The Accessibility Hierarchy and Centering Theory make different predictions about the importance of discourse segments. The Accessibility Hierarchy is based on the salience of an entity in the discourse, which is less sensitive to sentence-level boundaries. As mentioned above, the first instance of a DP might include a first and last name, but it is

unlikely that this referring expression will be used again. Regardless of the segmentation of the discourse, since this entity has been brought to salience at the onset of the discourse, the entity will likely remain accessible enough that the last name will not be necessary to refer to it subsequently.

If other entities are mentioned, first names will be necessary to disambiguate pronoun references and it will certainly be sensitive to discourse segmentation according to Centering Theory. In the case of definite descriptions, subsequent uses of this nominal form will probably be dependent on the subject matter of the discourse and how important that entity is in the story.

Evaluation

Poesio et al. (2004) tabulate the number of utterances which have a CB, the utterances which do not have a CB but are segment-initial, and those which neither have a CB nor are segment-initial in their study. They also calculate the proportion of utterances which violate Rule 1. We will calculate similar metrics using utterances from as many discourse segments in a narrative as possible.

For every coreference chain we find in a narrative, where chain is defined as all of the expressions used to refer to a particular entity, we will look at the forms of the DPs in that chain. Because we are choosing a subset of the nominal forms listed on the Accessibility Hierarchy, it is certain that these chains will contain nominal types that we are not annotating specifically. We will consider the relative ranking definite description <first name <pronouns. We will consider these chains in full, and also break them according to discourse segmentation to see if the discourse mentions are sensitive to this unit of analysis. We will calculate the proportion of chains which violate the hierarchy.

3.4 Results

Constraint 1 Violations

With our annotation scheme, we collected main clauses, relative clauses, and adjunct clauses.

Cb	Segment Initial	No Cb	Total
304	113	152	569

Table 3.1: All Clauses

Looking at all clause types, we find that 152/569 (26.7%) of clauses lack a backward-looking center. There are generalizations to be made about the violating cases, which will be discussed in the next section.

Cb	Segment Initial	No Cb	Total
292	111	133	536

Table 3.2: No Relative Clauses

Removing the relative clauses, we find that 133/536 (24.8%) of clauses lack Cb's. When we consider relative clauses, we find that there are 33 total in the data set, and 12 have backward-looking centers while 19 do not. These figures shed some light on our assumption that relative clauses can have Cb's that are not null traces.

Approximately 1/3 of the time, relative clauses have non-trace Cb's. This suggests that parallelism between the preceding utterance and the relative clause is a strategy used sometimes, but usually these clauses do not share arguments with their preceding clause, and excluding relative clauses ultimately leads to slightly fewer violations.

Cb	Segment Initial	No Cb	Total
250	109	120	479

Table 3.3: No Relative Clauses or Adjuncts

When we exclude relative clauses and adjuncts, we find that 120/479 (25.1%) of clauses lack a backward-looking center. While the number of clauses is reduced, the proportion of violating clauses changes only slightly.

General Discussion

It should be noted that the corpus used in this study consisted of weblogs, which often contain text mixed with photos. Therefore, there are many sentences in this data which we annotated but which probably should be excluded due to their non-story content.

Overall, we find that Constraint 1 is largely obeyed. However, there are some types of constructions which systematically violate Constraint 1. For instance, many of the violations come from sentences with non-thematic subjects:

I checked and there wasn't anything there.

As was pointed out to me by Pranav Anand (p.c.), discourses which conform completely to Centering Theory are often judged as boring, and using existential constructions is a way to break up long sequences of sentences with repeated subjects.

When sentences of this type intervene, the last entity to serve is the backward-looking center is often continued in the next utterance. This is reminiscent of the finding of Walker et al. (1998) that a Cb can be referenced across intervening discourse segments. It suggests that there should be some finer-grained level of annotation than the discourse segment, since existentials like this one are clearly part of the same segment as the utterance that precedes them, but behave as if they are not present for the purposes of tracking backward-looking centers.

Rule 1 Violations

Given that Rule 1 is intended to apply to 3rd person entities, it seems incorrect to assign violations of this rule if the pronominalized Cf is 1st or 2nd person. This follows from the fact that these entities are always pronominalized. Applying this method of

assigning violations leads us to find 0 violations of Rule 1 in our data set.

This finding seems to be genre-dependent. In this data set, Cb's which are not pronouns always occur in sentences with 1st person pronouns. Poesio et al. (2004) find that between 13.4% and 23% of utterances violate Rule 1, depending whether Cb's are directly or indirectly realized. However, in the genres they analyze, there are no 1st person pronouns used, though the patient leaflets contained some 2nd person pronouns. Therefore, non-pronoun Cb's are almost guaranteed to co-occur with 3rd person entities when they occur.

Accessibility Hierarchy

In this research, we annotated names, definite descriptions, and pronouns. Since definite descriptions are the least accessible according to the hierarchy, we expect that once an entity is referred to with a definite description, a referring expression that is higher on the hierarchy, like a pronoun, will be used to refer back to this entity. We expect a similar pattern for names. Once a name is used, we expect pronouns to be used for subsequent mentions of the same entity. However, while these predictions are on the right track, the application of these principles interacts with discourse segmentation and the number of entities in the discourse, as we will see.

We consider the data from 68 3rd person chains, where a chain is defined as all of the expressions used to refer to an entity throughout a narrative. The following table counts the number of chains where a name is used to refer to an entity, and subsequent mentions of that entity also use a name.

Total	Violations	% Violations
68	7	10.3

Table 3.4: Name Violations

Discussion

In the chains which contain subsequent name mentions after a name was used, we find that discourse segmentation and competing entities are factors which modulate the type of mention. Here is an example within a discourse segment where multiple mentions of the same entity use a name:

Do you think we should sit Kyle down and tell him that Cassy is cheating on her boyfriend with him? Because we all realise if she is willing to cheat on her boyfriend with Kyle, she will also be willing to cheat on Kyle.

In this context, there are multiple male entities within the same discourse segment, and therefore the name *Kyle* is used to avoid ambiguity associated with the pronoun *him*. Note that only pronouns are used to refer to *Cassy* after the first mention in this segment because she is the only female entity and ambiguity is avoided.

In the absence of multiple entities within the same discourse segment that a pronoun could refer to, the domain of application of the Accessibility Hierarchy appears to be the discourse segment. This means that if a name is used in discourse segment, subsequent mentions of that entity will be pronouns and the hierarchy will not be violated. In some stories, the narrator discusses interactions with multiple characters. Even if all the characters have different animacy and gender features such that pronouns are unambiguous, names are often reused at the onset of discourse segments as the relative salience of the characters shift throughout the narrative.

There are some narratives in which the narrator describes interactions they have with one other character. In these stories, the other character will first be referred to with a name, and no matter how the discourse is segmented, subsequent mentions of this character will only be pronouns. This is because these pronouns are unambiguous in their reference, and the entity is highly salient and accessible due to it being the only character besides the narrator.

The following table counts the number of chains where a definite description is

used to refer to an entity, and subsequent mentions of that entity also use a definite description.

Total	Violations	% Violations
68	11	16.2

Table 3.5: Definite Description Violations

Discussion

Of the coreference chains where definite descriptions are used for subsequent mentions, the majority of the definite descriptions are inanimate, suggesting these entities are less accessible throughout a discourse. For the animate entities, the types of violations that we see here are similar to those for the names. Consider the following examples from different narratives:

(32) *Then in went the frog! It sat, barely moving, on one of the larger rocks before realizing that water was near. The frog scurried into the water and proceeded to float.*

(33) *We put some holes in the lid so the frog could breathe, then some rocks and grass and bark and water inside. Then we put the frog in.*

In both examples, the definite description *the frog* is used in the first sentence, and this definite description is used again in the last sentence. In both examples, there are also intervening inanimate DPs like *one of the larger rocks* and *the water*.

If we remove the sentences containing these additional nouns, use of a pronoun to refer to *the frog* becomes more acceptable:

(34) *Then in went the frog! It sat, barely moving, before realizing that water was near. It scurried into the water and proceeded to float.*

(35) *We put some holes in the lid so the frog could breathe. Then we put it in.*

Contexts like this highlight that the definite descriptions are chosen when there are other possible antecedents in the same segment, just as we saw for names.

There is one narrative in which the narrator refers to her husband as *my husband* throughout the story, until she uses his name for the first time in the penultimate sentence of the narrative. Though many subsequent mentions of the husband are made using the definite description and it feels a little strange that the name is not used for so long in the story, the definite description is not used again after the name. This suggests that when two non-pronoun expressions are used for the same entity, the one lower on the Accessibility Hierarchy is not used again after a higher one is used. This was also seen in the story where the narrator has a bunny named *Devi*, and once this name is used in the first sentence, the definite description is not used again.

Within a discourse segment, when there are no competing entities that a pronoun could refer to, as with names we again find that pronouns will be used for subsequent mentions and the low Accessibility definite description will not be used to refer back to an entity, as in this example from a story discussed above:

First let me start by saying, I love my husband. I really, really do. If I didn't wouldn't I have married him and I wouldn't want to start a family with him and grow old together and all that good mushy stuff.

3.5 Conclusion

These data suggest that numerically, Centering Theory constraints are largely obeyed. However, there are clear cases where the constraints are intentionally evaded because long strings of sentences with parallel subjects are repetitive. It suggests that constraints of Centering Theory are not absolute, but can be used to characterize parts of some discourses. It also raises the question of how intervening existentials should be treated, since Cb's act often act as if they are not there.

Additionally, our annotation scheme and treatment of embedded clauses are likely

affecting these results, as is the fact that we annotated non-story content which occurs in weblogs. Given our definition of violations for Rule 1, we did not find any violations of this rule. However, if the genre did not contain so many 1st person entities, which are obligatorily pronominalized, we would expect our data to pattern more like texts analyzed by Poesio et al. (2004)

In our evaluation of the Accessibility Hierarchy, some generalizations emerged. Crucially, this theory interacts both with discourse segmentation and the number of entities in the discourse. Within a discourse segment, subsequent mentions of an entity are consistent with the Accesibility Hierarchy in that high Accessibility markers are used when there are no competing entities. When there are competing entities, lower Accessibility markers are used to disambiguate the referent where a pronoun would be ambiguous. In a discourse with several characters and many segments, lower Accessibility markers are used at the onset of segments. If there is only one character and the narrator, pronouns are used after the first mention of the character, regardless of discourse segmentation.

3.6 Future Work

This research is inspired by Poesio et al. (2004), in that we are interested in how changing annotation parameters affects the quantitative results, but also in understanding why these parameters contribute in the way that they do. Whatever annotation scheme is adopted will lead to some cases being evaluated as violations. A qualitative analysis is necessary to understand the parameters which are driving these figures.

There were many choices we made in our annotations which likely affected our results, and it would be interesting to compare alternatives. For instance, we decided to segment our clauses based on the matrix and embedded clauses, and it would be interesting to see the degree to which parallelism between these clause types is exhibited, how it relates to the diegetic/extradiegetic perspective of the narrator, and what the

numerical outcomes would be if we annotated them differently.

Hu and Pan (2001) argue that the definition of CB as defined in Centering Theory does not take into account the discourse segment topic, which leads to incorrect productions about what the CB of an utterance is. Since we only looked at realizations of DPs in the surface syntax of adjacent utterances, we did not take into account factors like topic, which surely would lead to different Cb's.

General Discussion & Conclusion

In this research, I have attempted to show that the distribution of nominal expressions in natural language is not random, but is in fact sensitive to grammatical hierarchies and the information structure of a discourse. In Chapter 1, I showed that the person-animacy hierarchy of Chamorro is a hard constraint which prohibits certain combinations of subjects and direct objects in transitive clauses. In Chapter 2, I presented data from English relative clause processing which does not suggest that a 2 >3 hierarchy is active in English, though there is evidence for pronoun >non-pronoun and animate >inanimate hierarchy effects in other domains. In Chapter 3, I presented corpus data that shows that English weblogs strongly respect the Accessibility Hierarchy and largely respect Centering Theory.

In Chamorro, the person-animacy hierarchy determines the types of DPs which can co-occur within a transitive clause. While this hierarchy is not strictly mirrored in English, there are statistical tendencies in the types of arguments that co-occur in transitive clauses in English. There is also evidence from relative clause processing and production that English speakers are sensitive to combinations of subjects and objects within relative clauses. Though the extent to which the preference for animate subjects over inanimate ones and pronominal subjects over non-pronominal ones is respected by the grammar varies by language, these same grammatical categories affect processing cross-linguistically.

I have made these claims by leveraging computational tools and supplementing linguistic data with evidence from corpora statistics and behavioral measures. In Chapter 1, we saw that this method of using corpus data allowed us to understand the extent

to which the person-animacy constraint is respected in Chamorro in a way that is not possible from elicitation data alone. In Chapter 2, I supplemented self-paced reading data with corpus statistics to find differences in the strengths of the sub-hierarchies of the person-animacy hierarchy as reflected in English. Single sentence trials in experimental settings are one way to examine these preferences, but aggregated distributional patterns over many sentences provide another way to probe these questions.

In weblogs, a corpus type which approximates speech, the types of referring expressions are modulated by structural considerations and the Accessibility of a DP. In the results of the experiment presented in Chapter 2, we found evidence for the Accessibility of 2nd person pronouns compared to 3rd person pronouns, and we observed a slight advantage in comprehension for sentences which obeyed Centering Theory. This data suggests that discourse-level hierarchies and structural preferences are relevant even in sentence-level trials like those most commonly used in psycholinguistics. Sentences do not exist in a vacuum, but are the units of structure which build discourses and constraints which operate at the discourse level should be considered when constructing experimental stimuli.

Unlike the Accessibility Hierarchy and Centering Theory, the person-animacy hierarchy constraint targets clause-sized linguistic structures, and violating the hierarchy leads to ungrammaticality. In contrast, the Accessibility Hierarchy and Centering Theory are theories of discourse coherence and how referential expressions are used with respect to the changing topic matter and number of entities under discussion. Given the different domains of application of these hierarchical rankings of grammatical prominence and referential accessibility, we could conceive of a situation where person-animacy hierarchy is called off in favor of discourse considerations, such as when an animate non-pronoun is the topic of the discourse. The topic construction has a null pronoun, and therefore avoids the person-animacy hierarchy by elevating an animate non-pronoun to a pronoun status (Chung 1998). This construction is usually used at the beginning of paragraphs (Cooreman 1987), and this perhaps illustrates a case where the person-

animacy hierarchy interacts with the discourse structure. Further interactions of this type between the types of data presented in Chapter 1 and Chapter 3 could be explored.

Appendix A

Experiment Stimuli

1. The nurse that welcomed {the mechanic/you/him} with a smile ran a marathon during the
2. The nurse that the {the mechanic/you/he} welcomed with a smile ran a marathon during
3. The paramedic that assisted the lifeguard with urgency went on vacation before it started
4. The paramedic that the lifeguard assisted with urgency went on vacation before it started raining.
5. The butcher that interrupted the electrician at the party got a promotion on the first of the month.
6. The butcher that the electrician interrupted at the party got a promotion on the first of the month.
7. The mailman that visited the preacher with a parcel fed the birds on Friday morning.
8. The mailman that the preacher visited with a parcel fed the birds on Friday morning.
9. The doctor that challenged the engineer in a competitive way won an award after healing from an injury.
10. The doctor that the engineer challenged in a competitive way won an award after healing from an injury.
11. The agent that surprised the clerk during the holidays bought a house at the end of the year.
12. The agent that the clerk surprised during the holidays bought a house at the end of the year.
13. The actor that questioned the pilot over the phone got a new car before school started.
14. The actor that the pilot questioned over the phone got a new car before school started.

15. The musician that greeted the realtor with a handshake went on a cruise in the middle of October.
16. The musician that the realtor greeted with a handshake went on a cruise in the middle of October.
17. The anthropologist that appreciated the director with enthusiasm went to the library before returning home.
18. The anthropologist that the director appreciated with enthusiasm went to the library before returning home.
19. The senator that dismissed the chef with boredom went grocery shopping after the game.
20. The senator that the chef dismissed with boredom went grocery shopping after the game.
21. The ranger that complimented the tailor very politely had a birthday before the election.
22. The ranger that the tailor complimented very politely had a birthday before the election.
23. The poet that befriended the runner in the park had a picnic in the middle of summer.
24. The poet that the runner befriended in the park had a picnic in the middle of summer.
25. The architect that called the firefighter last night got a new dog at the start of last week.
26. The architect that the firefighter called last night got a new dog at the start of last week.
27. The trainer that mentored the magician on the job left for a road trip on New Year's Eve.
28. The trainer that the magician mentored on the job left for a road trip on New Year's Eve.
29. The professor that phoned the administrator in desperation hiked a trail after a long week.
30. The professor that the administrator phoned in desperation hiked a trail after a long week.
31. The locksmith that respected the barber immensely taught a class at the end of the month.
32. The locksmith that the barber respected immensely taught a class at the end of the month.

33. The teacher that contacted the banker after many years threw a party before the sun set.
34. The teacher that the banker contacted after many years threw a party before the sun set.
35. The journalist that approached the chauffeur over the weekend built a treehouse on Tuesday afternoon.
36. The journalist that the chauffeur approached over the weekend built a treehouse on Tuesday afternoon.
37. The chemist that liked the gymnast since high school climbed a mountain over the course of two weeks.
38. The chemist that the gymnast liked since high school climbed a mountain over the course of two weeks.
39. The composer that praised the coach in adoration moved away during the holidays.
40. The composer that the coach praised in adoration moved away during the holidays.
41. The hairdresser that recognized the photographer across the street finished a project at the start of the day.
42. The hairdresser that the photographer recognized across the street finished a project at the start of the day.
43. The plumber that trusted the librarian with confidence had a baby during the weekend.
44. The plumber that the librarian trusted with confidence had a baby during the weekend.
45. The principal that addressed the lawyer with skepticism went kayaking before attending a show.
46. The principal that the lawyer addressed with skepticism went kayaking before attending a show.
47. The optometrist that overheard {the explorer/you/him} at the store quit his job a long time ago.
48. The optometrist that {the explorer/you/he} overheard at the store quit his job a long time ago.

Bibliography

J. Aissen. Markedness and subject choice in optimality theory. Natural Language & Linguistic Theory, 17(4):673–711, 1999.

M. Ariel. The function of accessibility in a theory of grammar. Journal of Pragmatics, 16(5):443–463, 1991.

T. G. Bever. The ascent of the specious, or theresa lot we dont know about mirrors. Explaining linguistic phenomena, pages 173–200, 1974.

S. E. Brennan, M. W. Friedman, and C. J. Pollard. A centering approach to pronouns. In Proceedings of the 25th annual meeting on Association for Computational Linguistics, pages 155–162. Association for Computational Linguistics, 1987.

J. Bresnan, S. Dingare, and C. D. Manning. Soft constraints mirror hard constraints: Voice and person in english and lummi. In Proceedings of the LFG01 Conference, pages 13–32, 2001.

B. T. A. Camacho. Nuebu Testamento. Diocese of Chalan Kanoa, 2007.

K. Christianson and H. Y. Cho. Interpreting null pronouns (pro) in isolated sentences. Lingua, 119(7):989–1008, 2009.

S. Chung. The design of agreement: Evidence from Chamorro. University of Chicago Press, 1998.

S. Chung. On reaching agreement late. Proceedings of CLS, 48, 2012.

A. M. Cooreman. Transitivity and discourse continuity in Chamorro narratives. Number 4. Walter de Gruyter, 1987.

M. Elsner, J. L. Austerweil, and E. Charniak. A unified local and global model for discourse coherence. In HLT-NAACL, pages 436–443, 2007.

S. P. Gennari and M. C. MacDonald. Linking production and comprehension processes: The case of relative clauses. Cognition, 111(1):1–23, 2009.

E. Gibson. The dependency locality theory: A distance-based theory of linguistic complexity. Image, language, brain, pages 95–126, 2000.

J. J. Godfrey, E. C. Holliman, and J. McDaniel. Switchboard: Telephone speech corpus for research and development. In Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on, volume 1, pages 517–520. IEEE, 1992.

A. Gordon and R. Swanson. Identifying personal stories in millions of weblog entries. In Third International Conference on Weblogs and Social Media, Data Challenge Workshop, San Jose, CA, 2009.

P. C. Gordon, R. Hendrick, and W. H. Levine. Memory-load interference in syntactic processing. Psychological science, 13(5):425–430, 2002.

P. C. Gordon, R. Hendrick, and M. Johnson. Effects of noun phrase type on sentence complexity. Journal of Memory and Language, 51(1):97–114, 2004.

B. J. Grosz, S. Weinstein, and A. K. Joshi. Centering: A framework for modeling the local coherence of discourse. Computational linguistics, 21(2):203–225, 1995.

J. R. Hobbs. Resolving pronoun references. Lingua, 44(4):311–338, 1978.

J. Hu and H. Pan. Processing local coherence of discourse in centering theory. In Proceedings of the 15 th Pacific Asia Conference on Language, Information and Computation. Hong Kong: City University of Hong Kong. Citeseer, 2001.

E. L. Keenan and B. Comrie. Noun phrase accessibility and universal grammar. Linguistic inquiry, pages 63–99, 1977.

N. Kwon, P. C. Gordon, Y. Lee, R. Kluender, and M. Polinsky. Cognitive and linguistic factors affecting subject/object asymmetry: An eye-tracking study of prenominal relative clauses in korean. Language, 86(3):546–582, 2010.

M. W. Lowder and P. C. Gordon. Effects of animacy and noun-phrase relatedness on the processing of complex sentences. Memory & cognition, 42(5):794–805, 2014.

W. M. Mak, W. Vonk, and H. Schriefers. The influence of animacy on relative clause processing. Journal of Memory and Language, 47(1):50–68, 2002.

New American Standard Bible. Anaheim: Foundation, 1997.

M. Poesio, R. Stevenson, B. Di Eugenio, and J. Hitzeman. Centering: A parametric theory and its instantiations. Computational linguistics, 30(3):309–363, 2004.

F. Reali and M. H. Christiansen. Processing of relative clauses is made easier by frequency of occurrence. Journal of Memory and Language, 57(1):1–23, 2007.

A. Staub. Eye movements and processing difficulty in object relative clauses. Cognition, 116(1):71–86, 2010.

P. Stenetorp, S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou, and J. Tsujii. Brat: a web-based tool for nlp-assisted text annotation. In Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics, pages 102–107. Association for Computational Linguistics, 2012.

S. A. Thompson. The passive in english: A discourse perspective. In Honor of Ilse Lehiste, pages 497–511, 1987.

R. S. Tomlin. On the interaction of syntactic subject, thematic information, and agent in english. Journal of Pragmatics, 7(4):411–432, 1983.

M. A. Walker, A. K. Joshi, and E. F. Prince. Centering in naturally occurring discourse: An overview. Centering theory in discourse, 128, 1998.

T. Warren and E. Gibson. Effects of np type in reading cleft sentences in english. Language and Cognitive Processes, 20(6):751–767, 2005.