MEASURING EMPIRICAL BIAS TOWARD ERGATIVITY AND ACCUSATIVITY

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

ÇAĞRI ŞAKİROĞULLARI

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
COGNITIVE SCIENCE

MAY 2019

Approval of the thesis:

MEASURING EMPIRICAL BIAS TOWARD ERGATIVITY AND ACCUSATIVITY

submitted by **ÇAĞRI ŞAKİROĞULLARI** in partial fulfillment of the requirements for the degree of **Master of Science in Cognitive Science Department, Middle East Technical University** by,

Prof. Dr. Deniz Zeyrek Bozşahin
Dean, **Graduate School of Informatics Institute, METU** _____

Prof. Dr. Cem Bozşahin
Head of Department, **Cognitive Science, METU** _____

Prof. Dr. Cem Bozşahin
Supervisor, **Cognitive Science, METU** _____

Examining Committee Members:

Prof. Dr. Deniz Zeyrek Bozşahin
Cognitive Science, METU _____

Prof. Dr. Cem Bozşahin
Cognitive Science, METU _____

Assist. Prof. Dr. Burcu Can Buğlalılar
Computer Engineering, Hacettepe University _____

Assist. Prof. Dr. Murat Perit Çakır
Cognitive Science, METU _____

Assist. Prof. Dr. Umut Özge
Cognitive Science, METU _____

Date: **03 May 2019**

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name: ÇAĞRI ŞAKİROĞULLARI

Signature :

ABSTRACT

MEASURING EMPIRICAL BIAS TOWARD ERGATIVITY AND ACCUSATIVITY

Şakiroğulları, Çağrı

M.Sc., Department of Cognitive Science

Supervisor : Prof. Dr. Cem Bozşahin

May 2019, 48 pages

Starting with six word order possibilities for a basic transitive clause, exposed data can bias English children to the point of making a categorial distinction for SVO. However, for an SVO language there are two categorial possibilities, an ergative and an accusative one. The acquisition of ergativity and accusativity is a complex phenomenon, since they both have similar phonological structure for the baby. We show, using Eve database of CHILDES, that these possibilities can be acquired from exposed data, to the extent that for any transitive clause there are actually eight possibilities available to the child. We do so using the radically lexicalized and probabilistically trainable grammar formalism of Combinatory Categorial Grammar.

Keywords: Combinatory Categorial Grammar, Ergativity, Accusativity, PCCG, Language acquisition

ÖZ

DİLDE DENEYİMSEL KILIMLILIK VE BELİRTMELİLİK EĞİLİMİNİN ÖLÇÜMÜ

Şakiroğulları, Çağrı
Yüksek Lisans, Bilişsel Bilimler Bölümü
Tez Yöneticisi : Prof. Dr. Cem Bozşahin

Mayıs 2019 , 48 sayfa

Yalın bir geçişli cümle için altı farklı olası sözdiziminden yola çıkıldığında maruz kalınan örnekler İngiliz bebeklerde zamanla belirgin bir şekilde Özne-Yüklem-Nesne (ÖYN) dizilimini yeğleme eğilimi oluşturmaktadırlar. Ancak ÖYN dizilimli bir dil için kılımlı (ergative) ve belirtmeli (accusative) olmak üzere iki farklı sözdizimi olasılığı bulunmaktadır. Kılımlılığın ya da belirtmeliliğin edinimi, yüzeyde ikisi de benzer yapılara sahip oldukları için daha karmaşıktır. Bu tezde CHILDES Eve veritabanı kullanılarak bebeğin maruz kaldığı örnekler üzerinden sekiz sözdizimsel olasılığı da değerlendirip doğrusunu edinebildiği gösterilecektir. Bunu yaparken de olasılıksal öğretime açık, dağarcık odaklı birleşimsel ulamsal dilbilgisi (CCG - Combinatory Categorial Grammar) kullanılacaktır.

Anahtar Kelimeler: Bileşimsel Ulamsal Dilbilgisi, Kılımlılık, Belirtmelilik, PCCG, Dil edinimi

to my small family of wonderful people
and to my greater family of wanderers in an unexplored universe...

ACKNOWLEDGMENTS

I would first of all thank my parents Gölay Şakiroğulları and A. Gökay Şakiroğulları for their unconditional support through my life of approximately 10000 days ($\pm$50 days) and for training the first set of entries in my lexicon. They also got me in contact with countless lexicon trainers since the early times of my life and flared up my enthusiasm for lexicon expansion.

The experiments in this thesis would not be possible without the supervision of my adviser Cem Bozşahin. He is indeed an extraordinary adviser by providing his students with inspiring ideas and instructive tools like CCGlab.

Omri Abend contributed to this thesis not only by their laborious work published in 2017, but also by providing me with their datasets.

I also thank to all my professors that have introduced me to different aspects of languages in the last three years: Umut Özge, Cem Bozşahin, Ceyhan Temürcü, Deniz Zeyrek Bozşahin, Cengiz Acartürk, Burcu Can Buğlalılar, Ayşenur Birtürk, Duygu Özge, Martina Gračanin Yüksek.

Members of the Examining Committee of this thesis (Deniz Zeyrek Bozşahin, Cem Bozşahin, Umut Özge, Burcu Can Buğlalılar and Murat Perit Çakır) have contributed with their alternative and constructive approaches.

My department fellows Samet Albayrak, Gizem Özen, Tunç Güven Kaya, Fırat Öter, Tzu-Ching Kao, Alaz Aydın, Ahmet Üstün, Faruk Büyüktekin, Mustafa Özaydın, Şükrü Bezen, Efe Can Yılmaz, Emre Erçin, Hüseyin Aleçakır, Borabay Kadirdağ, Oğuzhan Demir, Enis Dönmez, Zuhal Ormanoğlu, Yasemin Göl, Nihan Soycan, Maani Tajaldini, Sara Razzaghi, Kerem Usal, Çağatay Taşçı, Umutcan İpekoğlu, Arzu Burcu Güven, Salih Canpolat, Jan Watson (and probably others) have all been great subjects of habitual talks about deep matters of Cognitive Science with their diverse backgrounds.

I would also thank our department's student affairs officer M. Hakan Güler for always being able to find a solution to our problems.

Last of all, I would like to thank my university METU for being an inspiring environment filled with wanderers of all ages.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

CCG	Combinatory Categorial Grammar
PCCG	Probabilistic Combinatory Categorial Grammar
CKY	Cocke-Kasami-Younger algorithm
lf	Logical Form
SVO or AVO	Subject-Verb-Object syntactic order for a transitive sentence (accusative)
SVO′ or AVO′	Subject-Verb-Object syntactic order for a transitive sentence (ergative)
OVS	Object-Verb-Subject syntactic order for a transitive sentence
A	Agent (Subject of a transitive sentence)
O	Object of a transitive sentence
S	Subject of an intransitive sentence
P	Patient (Affected entity by an action)

CHAPTER 1

INTRODUCTION

Human language acquisition is a challenging procedure that mostly takes place in the first few years after birth. Even though various models were proposed to explain this phenomenon (Chomsky, 1969, Abend et al., 2017), the mechanisms making this procedure successful for children in normal course of development have not yet been well defined. Children learn to distinguish different phonemes, morphemes and words, acquiring syntactic structures of the language they are exposed to.

A major aspect of language learning is the acquisition of the syntactic word order. Syntactic word orders are generally represented with the order of the constituents in a neutral transitive sentence. The main constituents in a transitive sentence are the Subject (S), the Object (O) and the Verb (V). The verb is generally considered as the element that syntactically and semantically relates to the other two constituents. Therefore in the surface, there can be six different syntactic word orders in a language: SVO, SOV, VSO, VOS, OVS, OSV. Semantically, the entity that does the action is considered the Agent (A) and the entity that is affected by the action is called the Patient (P).

Abend et al. (2017) have demonstrated the acquisition of syntactic structures of the exposed language in presence of distracting meanings. Assuming that children can analyze their surroundings using an innate faculty of concepts present in human languages, the proposed Probabilistic Combinatory Categorial Grammar (PCCG) model can be trained to gradually link the surface forms that children hear from language speakers to their corresponding logical forms as well as to "acquire" the language by assigning a very high probability to the target structures, while minimal probabilities to others.

However, Abend et al. (2017) acknowledge that their model has forgone an account of ergativity in SVO and OVS languages. Depending on the first constituent the verb combines with, two different syntactic categories can be proposed. The verb that first attaches to the Object results in an *accusative-nominative* alignment (denoted as simply SVO or OVS in the thesis) and the verb that first combines with the Subject results in an *ergative-absolutive* alignment (denoted as simply SVO' or OVS' in the thesis).

In this study, I will explore this point omitted in Abend et al. (2017). With a similar approach, I will train a PCCG grammar in which SVO and SVO'

verbs are initiated with equal parameters with Eve dataset of CHILDES Corpus and its logical forms. Eve Corpus includes child-directed utterances in English (SVO) in the language acquisition period of a child and is frequently used in language acquisition studies, including Abend et al. (2017). For the experiment in this study, I did not substantially change the corpus and I used it to measure the categorial bias against wrapping categories for transitive verbs, which is the first step in understanding the acquisition of ergativity or accusativity. In order to make a more precise comparison between accusativity and ergativity, similar experiments need to be done after switching word order in intransitive sentences of the corpus.

CHAPTER 2

BACKGROUND

2.1 Ergative-absolutive Languages and Syntactic Ergativity

Ergativity is a type of morphosyntactic alignment for transitive sentences. Ergative-absolutive structures are generally considered to contrast accusative-nominative structures (Dixon, 1994). In accusative structures, the subject of intransitive sentences is aligned with the agent of a transitive sentence using either a case marking or through word order. Alternatively, ergative structures align the subject of an intransitive sentence with the object of a transitive sentence and generally require a separate case for the agent. Examples of ergativity can be found in a variety of languages, but ergative-absolutive structures generally coexist with accusative-nominative structures in the same language (*split ergativity*). The following examples (Ex. 1) of ergativity in Dyirbal are taken from Dixon (1994).

(1) Intransitive sentences

a. ŋ*uma* *banaga-n^{y}u*
Father return-NONFUT
S

Father(S) returned.

b. *yabu* *banaga-n^{y}u*
Mother return-NONFUT
S

Mother(S) returned.

(2) Transitive sentences

a. ŋ*uma* *yabu*-ŋ*gu* *bura-n*
Father mother-ERG see-NONFUT
P A

Mother(A) saw father(P).

b. *yabu* ŋ*uma*-ŋ*gu* *bura-n*
Mother father-ERG see-NONFUT
P A

Father(A) saw mother(P).

In Dyirbal, S and O are in the same case while A is marked by a case marker as shown in Example (2). Ergative-absolutive alignment in Dyirbal is demonstrated clearly in examples with coordination or subordinating conjunction of intransitive verbs with transitive verbs, shown in Example (3).

(3) Coordination

a. ŋ*uma* *banaga-n^{y}u* *yabu*-ŋ*gu* *bura-n*
Father return-NONFUT mother-ERG see-NONFUT
S (P) A

Father(S) returned, mother(A) saw father(P).

b. ŋ*uma* *yabu*-ŋ*gu* *bura-n* *banaga-n^{y}u*
Father mother-ERG see-NONFUT return-NONFUT
P A (S)

Mother(A) saw father(P), father(S) returned.

An important point to note in the Example (3) is that the translation in English required two sentences since the accusative-nominative alignment of English does not allow coordination between *Mother saw X* and *X returned* without repeating *X*.

Dyirbal is considered to have ergative *syntax* in Dixon (1994). There are also *semantically* ergative languages mentioned in Dixon (1994), such as Manipuri and Folopa, but they are not considered in the scope of this thesis.

2.2 Combinatory Categorial Grammar (CCG)

This section gives a brief introduction to Combinatory Categorial Grammar, described in detail by Steedman (1996, 2000). Generative grammars are based on phrase structure rules and parse strings top-down. The phrase structure rules given in Figure 2.1a are used to parse the example string *Eve loved Jack* in Figure 2.1b. Elements of the lexicon are considered as *terminal nodes* and parsing continues until all branches lead to a terminal node.

(a)
$S \rightarrow NP\ VP$
$VP \rightarrow V\ NP$
$NP \rightarrow$ Eve, Jack
$V \rightarrow$ loved
(b)

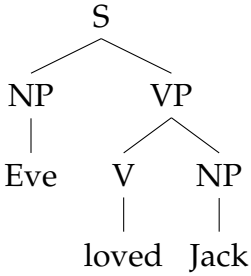


Figure 2.1: Generative grammar and top-down parse example

In contrast to generative grammars, combinatory categorial grammars include the syntactic role (syntactic *category*) and the semantics *in the lexicon* and parse strings bottom-up. This approach facilitates the computation and reduces the search space for parsing. CKY parsers are a common solution used in parsing strings with CCG. It also facilitates learning from supervision pairs of sentences and logical forms because all items are equipped with a logical form in the grammar and training data.

In CCG style grammars, every surface form is assigned a category with a syntactic type and a logical form. Syntactic categories of functors are generally written in a result-first-argument-last notation with slashes specifying the direction of the arguments. All categories can be defined as a combination of S (sentence), NP(noun phrase) and slashes. As an instance, the syntactic category S/NP is given to a surface form that is applied to a noun phrase to its right and forms a sentence.

Logical forms for functors in categories are described using lambda calculus (λ-calculus). Lambda calculus is a useful tool for describing functions and is used widely in computation of semantics. A basic expression in lambda calculus is shown in Figure 2.2a. This representation corresponds to a function with input x (next to the lambda) and replaces all instances of it with the input in the main expression (between paranthesis), resulting in two times the input. This evaluation procedure is called the β *reduction*, shown in Figure 2.2b. Lambda expressions with multiple inputs are generally simplified to have a lambda-variable cluster on the left side of the dot and function expression to

the right of it as shown in Figure 2.2c.

(a) $\lambda x.(2 \times x)$
(b) $(\lambda x.(2 \times x))5 \rightarrow_\beta 10$
(c) $\lambda x \lambda y.x + y$

Figure 2.2: Lambda calculus expressions and reductions

An example grammar is given in Figure 2.3. Each of the entries have a surface form (Eve, Jack, loved), a syntactic category (NP or S\NP/NP) and a logical form representing their meanings (*Eve'*, *Jack'*, $\lambda x \lambda y.$ *loved'xy*). Note that as a convention, the doer of the verb is always assigned the outermost position and objects are assigned inner positions.

Eve := NP : *Eve'*
Jack := NP : *Jack'*
loved := S\NP/NP : $\lambda x \lambda y.$ *loved'xy*

Figure 2.3: Example CCG lexicon

The most basic rules of the syntax-semantics interface in CCG are function application rules shown in Figure 2.4. Once the syntactic category of two adjacent strings allow the application, the semantics also can be computed.

(a) $X/Y : f \quad Y : a \quad \Rightarrow \quad X : fa$ (Forward Application : >)
(b) $Y : a \quad X\backslash Y : f \quad \Rightarrow \quad X : fa$ (Backward Application : <)

Figure 2.4: Functional application rules

Using the syntactic category and corresponding semantics included in the example lexicon in Figure 2.3, the meaning of a given surface form (*Eve loved Jack*) can be parsed bottom-up as shown in Figure 2.5.

$$\dfrac{\dfrac{\text{Eve}}{\begin{array}{c}NP \\ : Eve'\end{array}} \quad \dfrac{\dfrac{\text{loved}}{\begin{array}{c}S\backslash NP/NP \\ : \lambda x \lambda y.loved'\,x\,y\end{array}} \quad \dfrac{\text{Jack}}{\begin{array}{c}NP \\ : Jack'\end{array}}}{S\backslash NP \ : \ \lambda y.loved'\,Jack'\,y}\,{>}}{S \ : \ loved'\,Jack'\,Eve'}\,{<}$$

Figure 2.5: CCG bottom-up parse example

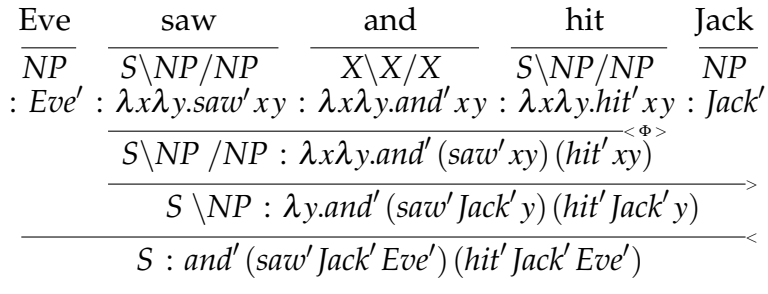
2.2.1 Coordination in CCG

Coordination is modeled to take two arguments of same type from each side of the conjunction and to result in a category of same type. This is represented as X\X/X, where X represents any category.

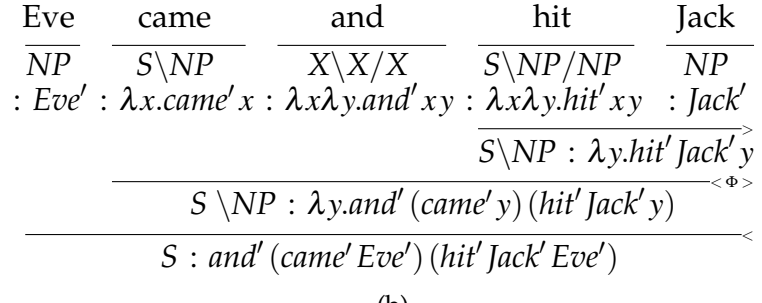
and := (X\X)/X : $\lambda p \lambda q \lambda x.\ and'\ (px)(qx)$ (Coordination : Φ)

Figure 2.6: Lexical entry for coordination

The definition of coordination seen above makes coordination of intransitive and transitive verb phrases unparseable with the intended semantics. Parses in Figure 2.7 show us an example of accusativity of verbs in English. The parse in Figure 2.7b would not be possible in ergatively aligned verbs since the coordination is made between an intransitive verb and a transitive verb.



(a)



(b)

Figure 2.7: CCG parses for some sentences with coordination

Some more complex structures observed in languages can be parsed using **B** (composition), **T** (type-raising) and **S** (substitution) *combinators* . Here, I will mention a few relevant combinators and give their definition: Combinator **B** (Combination) and Combinator **T** (Type-raising).

Combinator **B** *composes* two functions as shown in Figure 2.8. An example use of the combinator **B** for English is encountered in the parsing of modal verbs as shown in Figure 2.9

(a) $X/Y : f \quad Y/Z : g \quad \Rightarrow_{\mathbf{B}} \quad X/Z : \lambda x.f(gx)$ (Forward Composition : >**B**)
(b) $Y\backslash Z : f \quad X\backslash Y : g \quad \Rightarrow_{\mathbf{B}} \quad X\backslash Z : \lambda x.f(gx)$ (Backward Composition : <**B**)

Figure 2.8: Basic composition rules

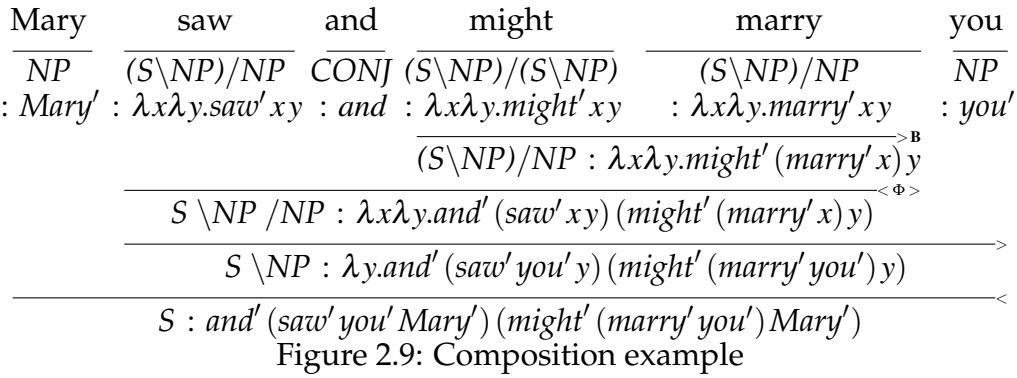


Figure 2.9: Composition example

Type-raising turns arguments into functions and are defined as shown in Figure 2.10.

(a) $X : a \quad \Rightarrow_{\mathbf{T}} \quad T/(T\backslash X) : \lambda f.fa$ (>**T**)
(b) $X : a \quad \Rightarrow_{\mathbf{T}} \quad T\backslash(T/X) : \lambda f.fa$ (<**T**)

Figure 2.10: Type-raising rules

More combinators (Substitution combinator **S**, generalized and cross versions of **B** and **S**) are defined in Steedman (2000), however, they are not used in scope of this thesis since the utterances in Eve Corpus have relatively simple or simplified structures.

2.3 Probabilistic CCG (PCCG)

In Probabilistic CCG, apart from the three parts mentioned above (surface form, syntactic category and logical form) every lexical entry is also assigned a likelihood parameter (Zettlemoyer and Collins, 2005). This is not the parameter in the sense proposed in Principles and Parameters approach (Chomsky et al., 2002). In PCCG grammars, *every entry* in the lexicon has a separate parameter, no matter how similar their categories are. In fact, even the same surface form (intransitive *read* as in *I read.* and transitive *read* as in *I read a book.*) may have different parameters for different lexical entries.

Using this parameter, we can calculate the likelihood of some logical form, and the derivation (parse tree) that resulted in it, $P(L,T|S)$ for each utterance, string or sentence. Utterances may result in multiple logical forms or may end up in the same logical form over different trees. The multitude of (L,T) pairs corresponds to ambiguities that can be ranked by their likelihood using the probability parameters of the lexicon entries that resulted in that sentence. The ambiguity generally occurs when a string has multiple categories in the lexicon. Ambiguity may also occur when a logical form is the result of multiple parse trees, generally referred as *spurious ambiguity*.

$P(L,T|S)$ is defined using the log-linear model described in Clark and Curran (2003). The function $\bar{f}$ maps (L,T,S) triplets to feature vectors in $\mathbb{R}^d$, assuming

d features. Even though defining complex features is possible, it is simply taken as the number of times a lexical entry is used in a tree T in Zettlemoyer and Collins (2005)

$$P(L,T|S;\bar{\theta}) = \frac{e^{\bar{f}(L,T,S)\cdot\bar{\theta}}}{\sum_{(L,T)} e^{\bar{f}(L,T,S)\cdot\bar{\theta}}} \tag{2.1}$$

Parsing with PCCG, we calculate the most probable logical form for a string S with given parameters $\bar{\theta}$. When computing the most probable logical form, the probability of different trees that yield the same semantics are summed.

$$arg_L max P(L|S;\bar{\theta}) = arg_L max \sum_T P(L,T|S;\bar{\theta}) \tag{2.2}$$

Apart from ranking the parses depending on their likelihood, parameters of a PCCG grammar can be updated. One needs surface forms and their "correct" logical forms (called *training pairs* and shown as (S,L) from now on) to train a PCCG lexicon. Training a PCCG grammar refers to estimating the optimal parameters and updating them to those values using the gradient ascent algorithm described below with provided training pairs. When estimating parameters according to given (S,L) pairs, the parse tree becomes a hidden variable. The log-likelihood of the training set is:

$$O(\bar{\theta}) = \sum_{i=1}^{n} log\, P(L_i|S_i;\bar{\theta}) \; = \sum_{i=1}^{n} log(\sum_T P(L_i,T|S_i;\bar{\theta}) \tag{2.3}$$

The derivation of equation 2.3 with respect to a parameter θ_j yields:

$$\frac{\partial O}{\partial \theta_j} = \sum_{i=1}^{n}\sum_T f_j(L_i,T,S_i)P(T|S_i,Li;\bar{\theta}) - \sum_{i=1}^{n}\sum_{L,T} f_j(L,T,S_i)P(L,T|S_i;\bar{\theta}) \tag{2.4}$$

This derivative is calculated using an adopted version of inside-outside algorithm (Baker, 1979) and the finalized gradient ascent algorithm for parameter estimation becomes:

Set $\bar{\theta}$ to some initial value
for $k = 0...N-1$ **do** ▷ N passes over training data
  for $i = 1...n$ **do** ▷ n pairs in training data set
    $\bar{\theta} = \bar{\theta} + \frac{\alpha_0}{1+c(i+kn)} \frac{\partial log P(L_i|S_i;\bar{\theta})}{\partial \bar{\theta}}$ ▷ α_0 and c are learning rate parameters
  end for
end for

Figure 2.11: Gradient ascent algorithm for parameter estimation

0 @Loc: Eng-NA/Brown/Eve/010600a.cha
1 @PID: 11312/c-00034743-1
2 @Begin
3 @Languages: eng
4 @Participants: CHI Eve Target_Child , MOT Sue Mother , COL Colin Investigator , RIC Richard Investigator
5 @ID: eng|Brown|CHI|1;06.00|female|||Target_Child|||
6 @ID: eng|Brown|MOT|||female|||Mother|||
7 @ID: eng|Brown|COL|||||Investigator|||
8 @ID: eng|Brown|RIC|||||Investigator|||
9 @Date: 15-OCT-1962
10 @Time Duration: 10:00-11:00
11 *CHI: more cookie . [+ IMP]
12 % mor: qn|more n|cookie .
13 % gra: 1|2|QUANT 2|0|INCROOT 3|2|PUNCT
14 % int: distinctive , loud
15 *MOT: you 0v more cookies ?
16 % mor: pro:per|you 0v|v qn|more n|cookie-PL ?
17 % gra: 1|2|SUBJ 2|0|ROOT 3|4|QUANT 4|2|OBJ 5|2|PUNCT
18 *MOT: how_about another graham+cracker ?
19 % mor: pro:int|how_about qn|another n|+n|graham+n|cracker ?
20 % gra: 1|3|LINK 2|3|QUANT 3|0|INCROOT 4|3|PUNCT
21 *MOT: would that do just as_well ?
22 % mor: mod|will& COND pro:dem|that v|do adv|just adv|as_well ?
23 % gra: 1|3|AUX 2|3|SUBJ 3|0|ROOT 4|5|JCT 5|3|JCT 6|3|PUNCT
24 *MOT: here .
25 % mor: adv|here .
26 % gra: 1|0|INCROOT 2|1|PUNCT
27 *MOT: here you go .
28 % mor: adv|here pro:per|you v|go .
29 % gra: 1|3|JCT 2|3|SUBJ 3|0|ROOT 4|3|PUNCT
30 *CHI: more cookie . [+ IMP]
31 % mor: qn|more n|cookie .
32 % gra: 1|2|QUANT 2|0|INCROOT 3|2|PUNCT
33 % int: distinctive , loud
34 *MOT: you have another cookie right on the table .
35 % mor: pro:per|you v|have qn|another n|cookie adv|right prep|on
36 det:art|the n|table .
37 % gra: 1|2|SUBJ 2|0|ROOT 3|4|QUANT 4|2|OBJ 5|6|JCT 6|2|JCT 7|8|DET 8|6|POBJ 9|2|PUNCT
39 *CHI: more juice ?
40 % mor: qn|more n|juice ?
41 % gra: 1|2|QUANT 2|0|INCROOT 3|2|PUNCT
42 *MOT: more juice ?
43 % mor: qn|more n|juice ?
44 % gra: 1|2|QUANT 2|0|INCROOT 3|2|PUNCT

Figure 2.12: Extract from original Eve Corpus

2.4 CHILDES Database and Brown Eve Corpus

CHILDES database is a collection of transcribed child-directed speech in various languages (Macwhinney, 2000). For this study I used the Eve set (Brown, 1973) of Brown Corpus, which consists of the transcriptions of the utterances recorded during 20 sessions of about an hour each conducted as she is at the age of 18 to 27 months old. Brown (1973) describes Eve as "linguistically precocious child" and indicates that her speech developed rapidly over the nine months these sessions took place, between October 15, 1962 (first session) until July 23, 1963 (last session). These sessions, together with the other two sets (Adam and Sarah) and their transcriptions, were included in CHILDES database available online. The methodology of transcription is explained in detail in the CHAT Manual included in Macwhinney (2000). An extract from the transcription of the first session is shown in Figure 2.12.

The transcriptions include the utterances of the child, mother and investigators. Sessions take about one hour each. The utterances were annotated by investigators with respect to morphology, but annotators also noted phonological character of some utterances and a dependency parse was also included.

As indicated in the CHAT manual Macwhinney (2000), there were several issues in transcribing the recordings. Letter spellings(*it is e v e .* $\rightarrow$ *it is e@l v@l e@l* , repetitions (*milk milk milk milk*), assimilations (*going to, give me* $\rightarrow$ *gonna, gimme*), baby talk (*choochoo*), unidentifiable words (shown as *x*) are all among the features that needed special attention in the transcription. All utterances ended with one of three basic utterance terminators: period (.), question mark (?) or exclamation point (!).

2.5 Previous Studies about Language Acquisition with PCCG

CHILDES Corpus has been used to conduct experiments about language acquisition as it provides researchers with both the utterances of the child as well as the utterances they are exposed to. One of those studies making use of the CHILDES Corpus is the study conducted by Abend et al. (2017).

Kwiatkowski (2012) has used the dependency representations created by Sagae et al. (2010) with CHILDES Corpus and semi-automatically turned them into Davidsonian-style meaning representations. Abend et al. (2017) used these representations for their experiments. Since the main stimulus the child is exposed to during the language acquisition period is the mother's utterances, marked by MOT.

Abend et al. (2017) considered the possible ambiguity in child's environment and conducted all their simulations in four scenarios with 0, 2, 4 and 6 distractors apart from the correct meaning. These distractor interpretations are chosen randomly and do not need to have a conceptual similarity with the utterance. For example, the utterance *where 's your cup ?* is assigned two dis-

tractors in Figure 2.13b, one for the utterance *more juice ?* and the other one for *I took it ..* Abend et al. (2017) made use of the CHILDES Corpus's Eve set and the meaning representations done by Kwiatkowski (2012). These pairs were used to train their PCCG lexicon using a Bayesian learning algorithm. In total there were 5123 utterances in their training pairs. This number is 41% of the complete Eve dataset since they discarded very long sentences (more than 10 words), one word interjections (*hmm*, *yeah*) etc. They divided this corpus of 5123 utterances into a training set of 4915 utterances and a test set of 208 utterances.

Sent: where 's your cup ?
Sem: lambda $0_e.eqLoc(pro:poss:det | your($1,n | cup($1)),$0)
example_end

(a) Corpus entry with no distractors

Sent: where 's your cup ?
Sem: lambda $0_ev.Q(qn | more($1,n | juice($1)),$0)
Sem: lambda $0_e.eqLoc(pro:poss:det | your($1,n | cup($1)),$0)
Sem: lambda $0_ev.v | take& PAST(pro | I,pro | it,$0)
example_end

(b) Corpus entry with two distractors

Sent: where 's your cup ?
Sem: lambda $0_ev.v | go(pro | you,$0)
Sem: lambda $0_ev.Q(qn | more($1,n | juice($1)),$0)
Sem: lambda $0_e.eqLoc(pro:poss:det | your($1,n | cup($1)),$0)
Sem: lambda $0_ev.v | take& PAST(pro | I,pro | it,$0)
Sem: lambda $0_ev.not(adj | sure(pro | I),$0)
example_end

(c) Corpus entry with four distractors

Sent: where 's your cup ?
Sem: lambda $0_ev.adv:loc | here($0)
Sem: lambda $0_ev.v | go(pro | you,$0)
Sem: lambda $0_ev.Q(qn | more($1,n | juice($1)),$0)
Sem: lambda $0_e.eqLoc(pro:poss:det | your($1,n | cup($1)),$0)
Sem: lambda $0_ev.v | take& PAST(pro | I,pro | it,$0)
Sem: lambda $0_ev.not(adj | sure(pro | I),$0)
Sem: lambda $0_ev.Q(aux | be& PRES(part | say-PROG(pro | you, n:prop | Fraser,$0),$0),$0)
example_end

(d) Corpus entry with six distractors

Figure 2.13: Examples with various number of distractors used in Abend et al. (2017)

Since they used the Eve database and their meaning representations in the simulations, their main interest was not how the infant could separate the utterance into lexical items or how the infant had the structured representa-

tion present in logical forms. However, their probabilistic model and learning algorithm took into account novel words and sentences.

The significance of the study conducted by Abend et al. (2017) is that their experiments on language acquisition using PCCG resulted in interesting learning curves. They interpret these curves to be similar to aspects of the phenomena observed in language acquisition (vocabulary spurt, learning of nouns before verbs etc.).

As for the learning of syntactic word order, Abend et al. (2017) considered six syntactic word order possibilities: SVO, SOV, OVS, OSV, VSO, VOS. Relative probabilities of SVO word order prevails in all settings with 0, 2, 4, 6 distractors as shown in Figure 2.14.

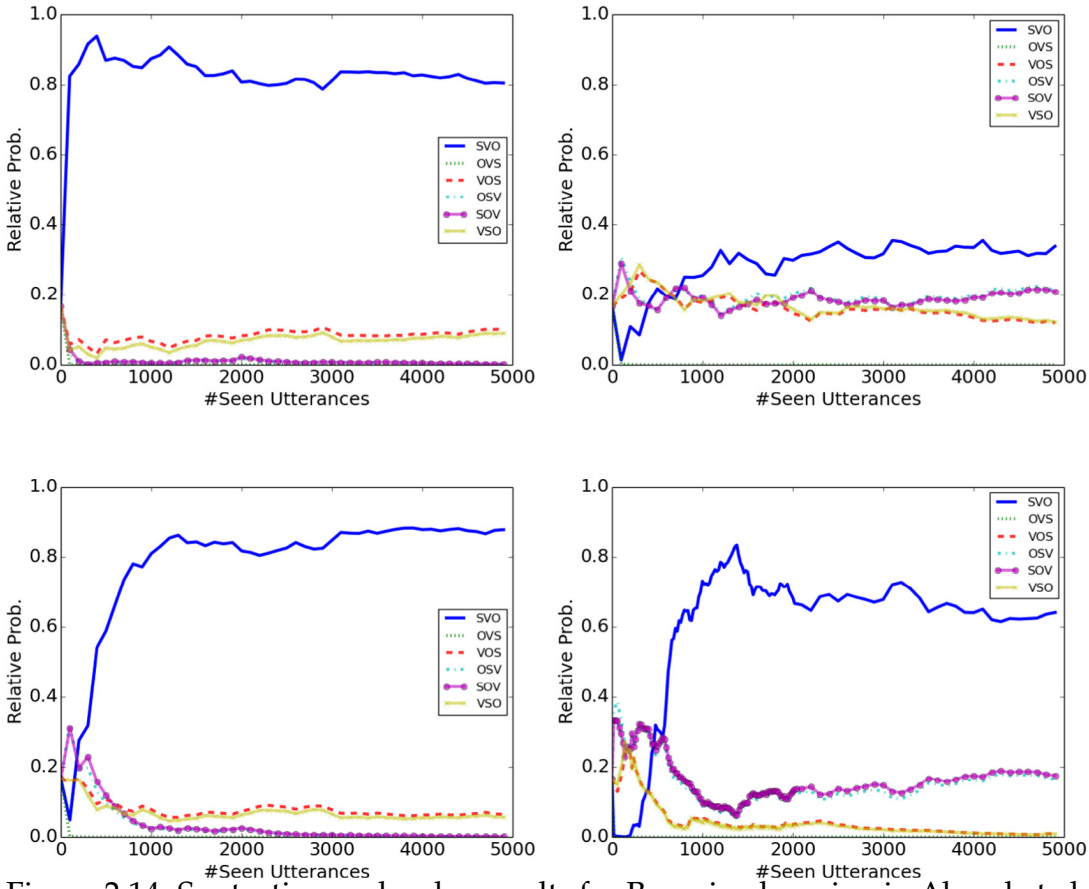


Figure 2.14: Syntactic word order results for Bayesian learning in Abend et al. (2017) with 0, 2, 4 and 6 distractors (left to right, top to bottom order)

Considering that they used an English corpus, SVO order prevailed rapidly after less than 500 training pairs in case there is no distractor as shown in Figure 2.14. Similarly, in case of 4 and 6 distractors, SVO prevailed after approximately 1000 training pairs. Only in the 2 distractors setting the SVO word order does not get a significant difference from others (even though it is the most probable word order as well). Abend et al. (2017) explains this with using a very steep learning curve and conducts further experiments with a more gradual learning rate and the resulting relative probability graph is given in Figure 2.15.

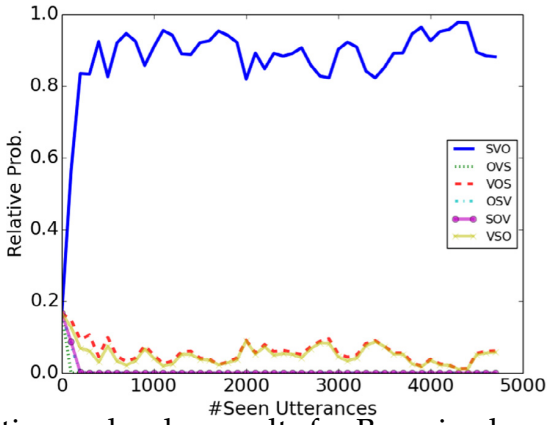


Figure 2.15: Syntactic word order results for Bayesian learning in Abend et al. (2017) with 2 distractors and a more gradual learning curve

Abend et al. (2017) mentions that there are two alternative versions for transitive verbs (shown in Figure 2.16 in verb medial syntactic orders (SVO and OVS) that can lead to the same final logical form. However, they assume that the verb in these cases attaches first to the semantic object and do not consider the categories for verb medial ergative transitive verbs in their experiments.

like := (S\NP)/NP : $\lambda x \lambda y.$ *like(y,x)*

(a) Verb medial *accusative* entry

like := (S/NP)\NP : $\lambda y \lambda x.$ *like(y,x)*

(b) Verb medial *ergative* entry

Figure 2.16: Two alternative categories for verb medial transitive verbs mentioned in Abend et al. (2017)

With their Bayesian learning algorithm, Abend et al. (2017) have shown the acquisition of many features in the language, including syntactic disposition of determiners, noun/verb distinction, novel words, nonce verbs and the most important of all, syntactic word order. They have considered six possible word orders (SVO, SOV, VSO, VOS, OSV, OVS) and assumed accusativity in SVO and OVS order.

For this thesis, I obtained the Eve data set used in Abend et al. (2017). I made minor corrections and changes in the utterances explained in Chapter 3 and listed in Appendix C. Since I divided some entries (like *milk salt egg*) into multiple utterances, resulting data set contains 5134 utterances.

In following chapters "corpus" refers to this modified version of the original Eve set collected and first published by Brown (1973), revised by Macwhinney (2000), morphosyntactically annotated by Sagae et al. (2010), semantically parsed by Kwiatkowski (2012) and used in Abend et al. (2017). Though, for several reasons explained in Chapter 3, I did not use the Davidsonian style logical forms of Kwiatkowski (2012).

Statistics for the utterances (number of utterances by sentence type and main verb type in each set) are given in Table 2.1. There were multiple verbs in some of the utterances (e.g., *go and get them* and *what do you want me to do ?*). All the verbs are included in the main verb count but the auxiliary verbs (e.g., be, have, do) are not counted.

Table 2.1: Statistics for the Eve data set used in this study

Set no.	Total	Sentence type			Main verb type				
		Informative	Question	Imperative Non-sentential	None	Intransitive	Be	Transitive	Ditransitive
1	384	141	162	81	83	39	115	139	13
2	372	152	143	77	60	54	111	135	13
3	134	59	49	26	23	23	38	39	11
4	315	146	92	77	69	33	74	131	9
5	323	105	127	91	80	37	94	108	9
6	188	66	97	25	32	32	53	71	1
7	296	113	126	57	39	31	81	128	18
8	405	132	202	71	65	28	114	191	9
9	159	75	54	30	21	13	32	90	3
10	172	68	69	32	26	21	52	63	13
11	169	64	73	32	22	20	54	71	3
12	221	99	92	30	26	23	81	85	6
13	152	71	48	33	33	13	41	61	4
14	182	68	84	30	39	18	40	83	2
15	319	162	97	60	48	52	84	131	5
16	274	127	94	53	61	23	62	125	3
17	373	132	161	80	79	34	97	158	9
18	309	131	111	67	55	43	98	108	8
19	178	85	67	26	26	27	68	60	1
20	209	68	90	51	48	24	61	70	7
Total	5134	2064	2038	1029	935	588	1450	2047	147

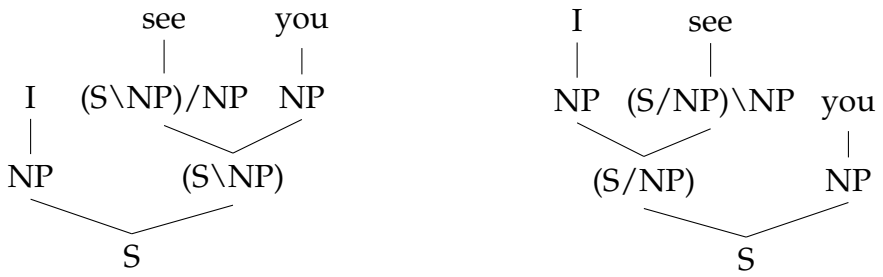
CHAPTER 3

DERIVING LOGICAL FORMS FOR EVE

While measuring the bias toward accusativity and ergativity, the PCCG model I have used in the experiment of this thesis (explained in Section 2.3) requires training with utterance-meaning pairs. Therefore, just as Abend et al. (2017) used the pairs generated by Kwiatkowski (2012), I have generated logical forms for utterances obtained from the Eve Corpus.

I decided not to use the logical forms of Kwiatkowski (2012) because the Davidsonian style event-entity representations were not fit for my purposes in this thesis and generating logical forms turned out to be an easier task than translating Davidsonian style event-entity representations to the CCGlab format.

The ergative English model in this thesis is based on the attachment order of the arguments for verb entries: Intransitive verbs attach only to the subject to the left. Accusative transitive verb entries first attach to the object to its right (forming a "verb phrase") and then attaches to the subject as shown in 3.1a. Ergative transitive verb entries first attach to the subject and then attach to the object, forming the tree in 3.1b.



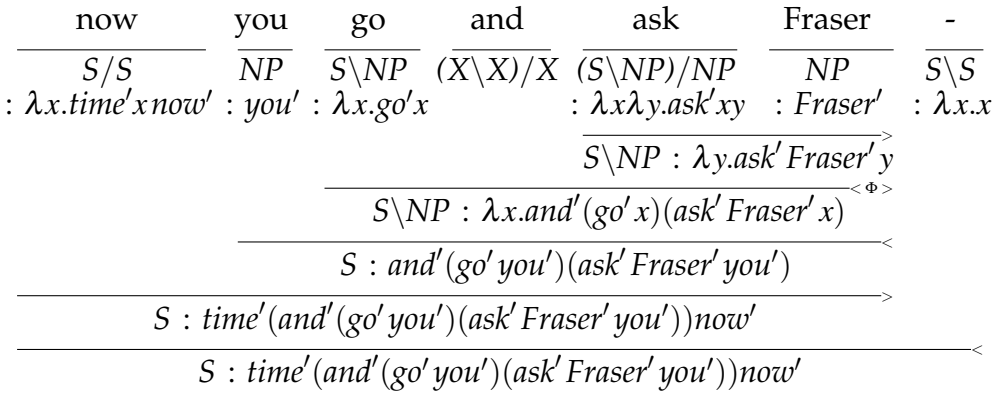
(a) Accusative transitive verb (b) Ergative transitive verb

Figure 3.1: Parse trees for different transitive verbs

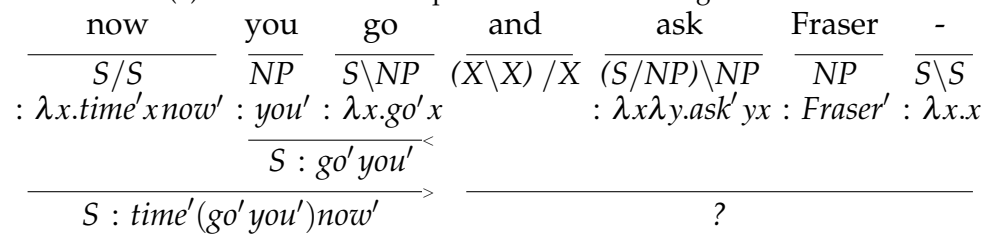
The speculated English is *syntactically* ergative and I assumed that no matter which syntactic tree the baby assumes (accusative or ergative), it reaches the same logical form with A or S in the outermost position.

This change in the attachment order makes some structures unparseable for ergative entries. Most common examples in the corpus are imperative sen-

tences, coordination between transitive and intransitive verbs, use of adverbs. As seen in Figure 3.2, the coordination between the transitive clause and the intransitive clause parses successfully with the accusative entry while the ergative transitive entry makes the parse impossible using the coordination model in Figure 2.6.



(a) Coordination example with accusative English entries



(b) Same entry with speculated ergative English entries (does not project)

Figure 3.2: Coordination example parsable only with accusative entries

3.1 Preprocessing

In order to facilitate parsing and to simplify the grammar, abridged instances like "*won ′t*", "*shouldn ′t*", "*doesn′ t*", "*you ′re*" are changed into their unabridged versions "*will not*", "*should not*", "*does not*", "*you are*" and so on with the exception of "*′s*" that marks possession. Also, all periods (.) were replaced by hyphens (-) to facilitate working with Common LISP reader. Also, along the annotation, I have noticed some corpus entries had unintended repetitions, missing verbs, mistaken punctuation etc. A list of corrected entries can be found in Appendix C.

3.2 Logical Forms

Logical forms for the utterances in the corpus are generated according to the guidelines described below. Each utterance in the corpus is assigned a logical form. Logical forms contain information about the tense, aspect, subject, object(s), type of sentence, modals and adjuncts.

The representational model does not consider complex linguistic features like conditionals or intrasentential conjunctions because very long sentences were either omitted by Abend et al. (2017) or divided into two utterances by Brown (1973). There were no conditional structures in the corpus.

Since being child directed speech, sentences are mostly simple and there are single word utterances and repetitions. Non-sentential utterances were assigned a sentence-like logical form. Also, pragmatic and contextual references (*there*, *this*, *it*) are frequently used in the corpus. I evaluated each utterance singularly and did not seek to solve pragmatic inferences between utterances or to the visual context. Example logical forms are available in Appendix B.

3.2.1 Informative Sentences

Every sentence and utterance is parsed to a logical form beginning with a tense-aspect group. In this group abbreviations for aspects are *simp*′ for simple, *cont*′ for continuous and *prft*′ for perfect. Tenses are *pst*′ for past, *prt*′ for present, *ftr*′ for future, *gng*′ for going-to or gonna. Also, modals (can, must, would...) begin with (*simp*′ <*modal*>′) and to-do infinitives have (*simp*′ -′) at the beginning. Copula is represented by *eq*′.

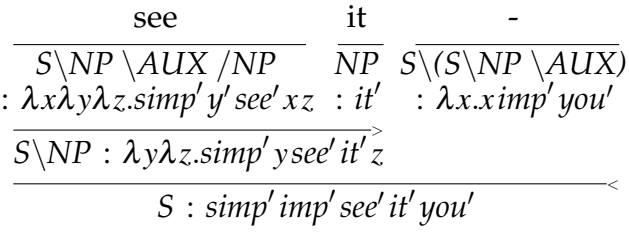
Negative sentences have a *not*′ in their tense-aspect group, except if it is formed by "no" as in *you have no pockets ?*. Some examples are available in Table B.1.

3.2.2 Question Words and Questions

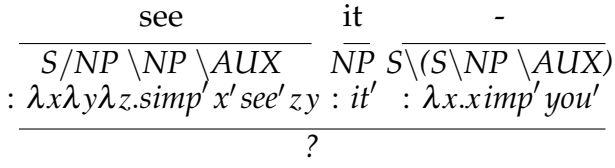
Question words are assigned a NP[type=qw] category and a logical form that begins with Q'. Questions are represented similar to Karttunen (1977). Yes-no questions then have the logical form of the informative sentence. Wh-questions have the logical form of the informative sentence and the questioned element replaced with a logical form including Q' sign. Some examples can be found in Table B.2.

3.2.3 Imperative Sentences

Imperative sentences have (*simp*′ *imp*′) at the beginning, and *you*′ is added by the period (.) at the end of the utterance. Some examples can be seen in Table B.4



(a) Accusative entry



(b) Ergative entry

Figure 3.3: Parse for imperative sentence

3.2.4 Non-sentential Utterances

A significant portion of the utterances in Eve Corpus consists of non-sentential utterances, mostly of a noun or a noun phrase with a punctuation (period or question mark). I assumed that these utterances refer to the visual or verbal context and therefore I have entered a logical form corresponding to *it is* <*X*>. for a N or a NP X. Similarly for questions, the logical form corresponds to a yes-no question *is it* <*X*>*?*. Examples are available in Table B.3.

3.3 Grammars

3.3.1 Nouns, Noun Phrases, Pronouns, Proper Names

Nouns are syntactically annotated according to whether they are countable and according to their count (singular or plural).

Uncountable nouns, plurals, proper names and some other types of nouns (e.g., reflexive nouns - yourself, one as in *you have one .* etc.) are directly assigned a syntactic category of NP with corresponding lexical constraints.

(a) duck n ⊢ N[type=count, count=sg] : *duck′*
(b) coffee un ⊢ NP[type=uncount] : *coffee′*
(c) crayons pln ⊢ NP[type=count, count=pl] : *pl′ crayon′*
(d) Fraser pn ⊢ NP[type=proper] : *Fraser′*
(e) him pro ⊢ NP[type=pronoun] : *he′*
(f) yourself np ⊢ NP : *yourself′*

Figure 3.4: Some examples for nominal lexical entries

Compound nouns like *grape juice* are unified in the corpus to form a single

symbol *grape-juice* and a separate entry was added to the lexicon for *grape-juice*. A complete list of the compound nouns modified this way can be found in Table C.2

3.3.2 Adjectives, Predicates, Determiners

Determiners, adjectives and predicates are syntactically np/n and np/np. Some lexicalized adjectives (e.g., *brand new*) are united with a hyphen (-) as in compound nouns.

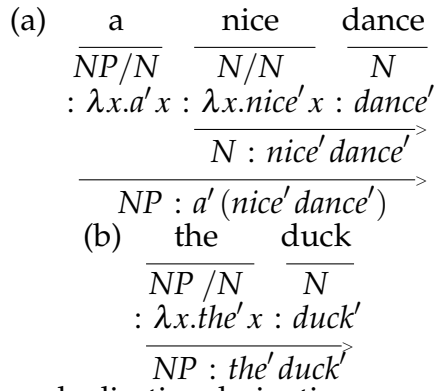


Figure 3.5: Determiner and adjective derivation examples
(lexical constraints omitted for sake of brevity)

3.3.3 Prepositions

Prepositional phrases are represented as noun phrases with lexical constraints in the syntactic category and as a logical unit with the preposition at the head.

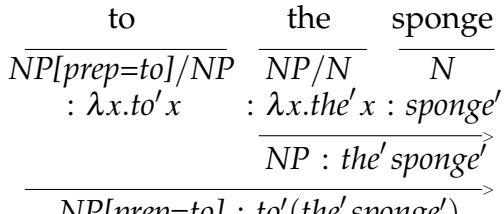


Figure 3.6: Prepositional phrase derivation example

Some phrasal verbs (e.g., *look at*) in the corpus require prepositional phrases with a particular preposition. This constraint is handled in the lexical entries for phrasal verbs.

3.3.4 Adjuncts, Adverbs and Sentential Adverbs

Temporal, instrumental, locative and causal adjuncts are represented similar to prepositional phrases that result in S\S instead of a NP. Some temporal and

locative adjuncts (e.g., now, tomorrow, here, there) are directly assigned an adjunct category. Adjuncts are parsed similarly to the prepositional phrases 3.7 except that they result in a (S\S) category at the end.

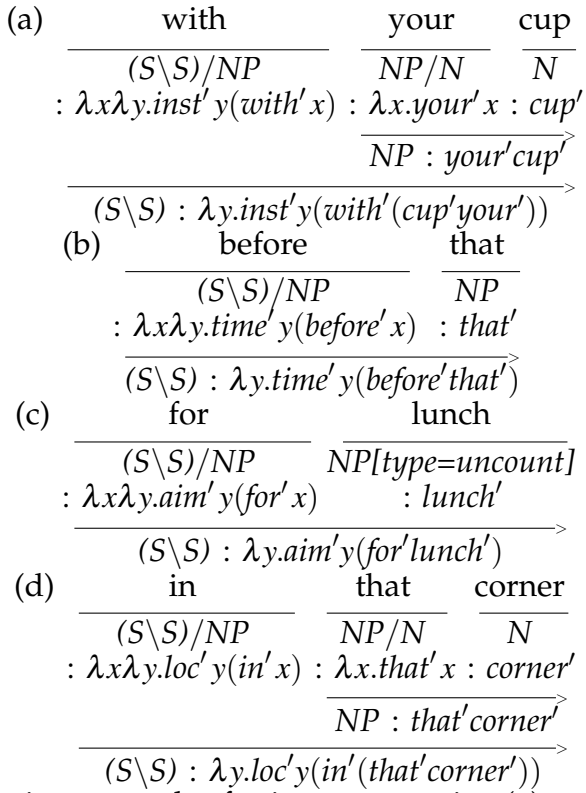
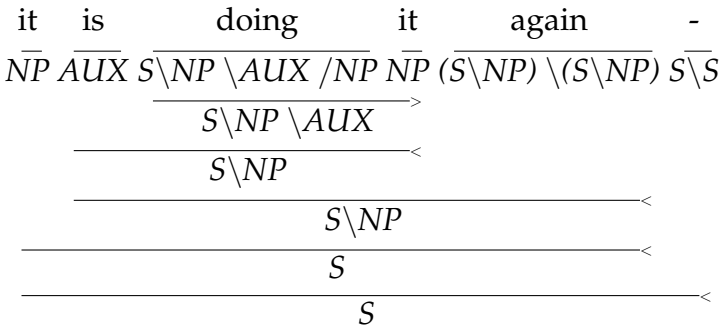


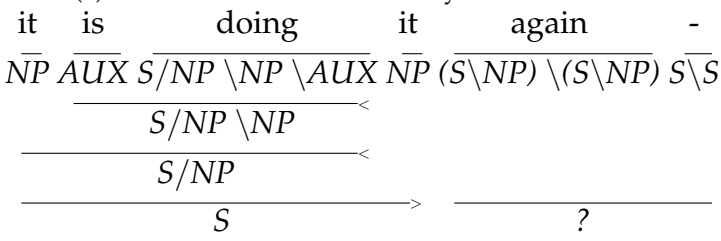
Figure 3.7: Derivation examples for instrumentative (a), temporal (b), causal (c) and locative (d) adjuncts

Sentential adverbs are assigned a syntactic category of (S\S) and (S/S) similar to adjuncts. They can be parsed with both accusative and ergative main verbs since they do not take a (S\NP).

On the other hand, adverbs are assigned a (S\NP)\(S\NP) or (S\NP)/(S\NP) category. This creates a distinction between ergative and accusative entries for some utterances in the corpus as shown in Figure 3.8, since it assumes transitive verbs to attach first to the NP to their right side (Object).



(a) Adverb with accusative entry of transitive verb



(b) Adverb with ergative entry of transitive verb

Figure 3.8: Examples for adverbs with accusative and ergative entries

3.3.5 Verbs

Since the main research question of the thesis is ergativity and accusativity, verbs are the main parsing elements holding the syntactic information for all types of sentences. Each verb (in basic form) has at least three lexical entries: two for informative sentences (with or without auxiliary), one for a yes-no question. Continuous (-ing) and past participle forms similarly have three entries but require a be or have auxiliary. Verbs in past form (e.g., *went*) have less entries since they do not form questions as they are. Transitive and ditransitive verbs have an accusative variant taking arguments first from right side (as in English and in Abend et al. (2017)) and speculated ergative entries taking object arguments first from left side. They all result in the same logical form. Examples for various forms in each verb category are available in Appendix A

3.3.5.1 Intransitive verbs

Intransitive sentences in basic form have the three above-mentioned entries since they do not have ergative/accusative distinction.

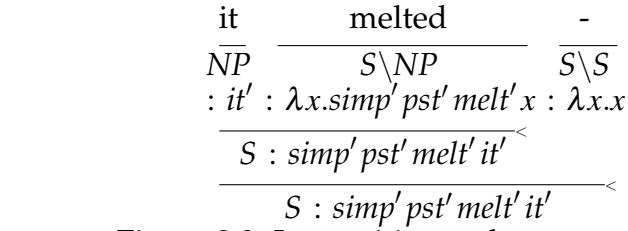
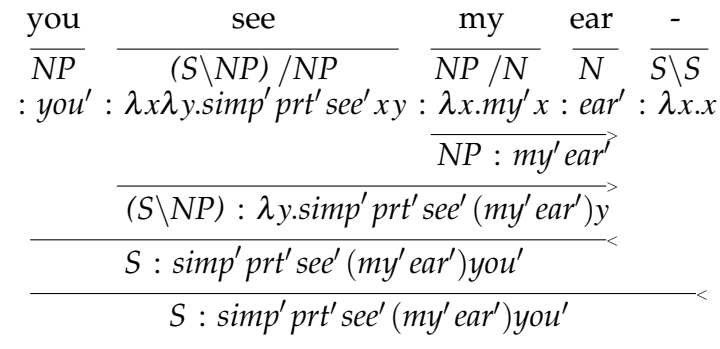


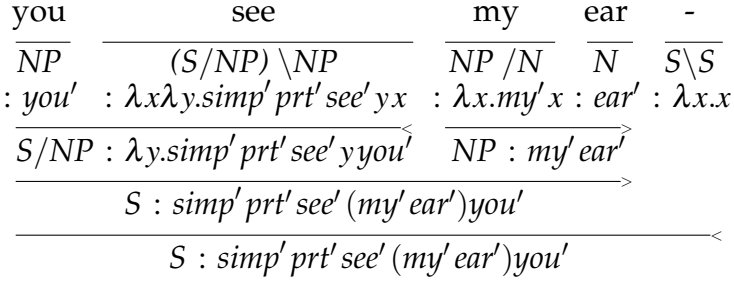
Figure 3.9: Intransitive verb parse example

3.3.5.2 Transitive and ditransitive verbs

Transitive verbs are modeled depending on the order they take their arguments in. Accusative verbs are assigned a syntactic category that takes arguments corresponding to the object(s) from right and ergative verbs first take the argument from left (agent). They both result in the same logical form so that given the meaning representation both structures can create a valid tree for most cases. Transitive and ditransitive verbs also have an extra entry for wh-questions. However, since the SVO order is not preserved in wh-questions, there is no speculated ergative or accusative variant of the entry.



(a) Accusative version



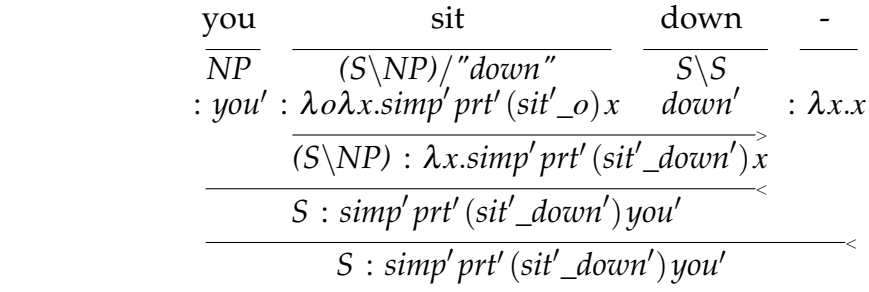
(b) Ergative version

Figure 3.10: Transitive verb parse example

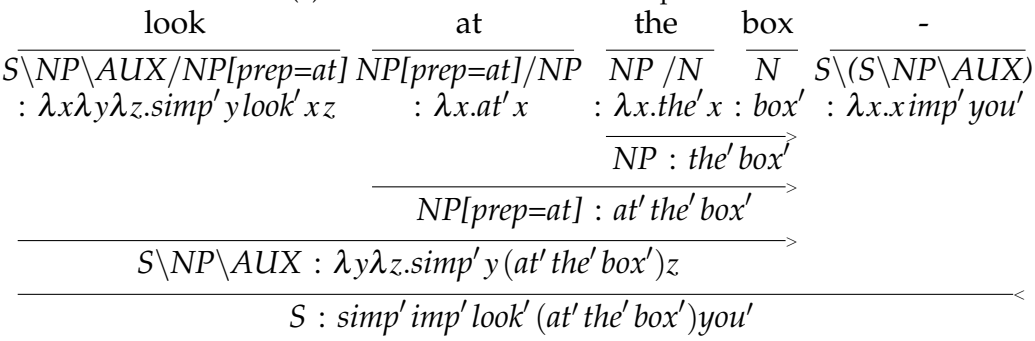
Another issue in handling transitive and ditransitive verbs has been the auxiliary verbs. Since the main aim of this thesis is to train verbs based on their transitivity, auxiliaries are placed in the syntactic categories of the main verbs in a way that does not hinder the parsing of either alignment models.

3.3.5.3 Phrasal verbs

Phrasal verbs are frequently used in the corpus and their categories are created in parallel to their transitive or intransitive counterparts. They are handled in two different categories. The first group is the phrasal verbs with motion-through-location or terminus features, formed by combining with adverbial particles like *down*, *up*, *in*, *off*, *back*, *on*, *out* as proposed in Bolinger (1971). In the syntactic category of this group the particles were included in the syntactic category of the verb entry as proposed in Bozsahin and Guven (2018). The logical form of the verb is also modified to include the semantic change. The parse of a verb of this group (e.g., *sit down*) can be seen in Figure 3.11a. The second group of phrasal verbs are the verbs that require prepositional phrases formed by a particular preposition as argument (e.g., *look at*). The syntactic categories of these verbs were modified to accept noun phrases with the required preposition, but the verb in the logical form was left as it is (Figure 3.11b).



(a) Phrasal verb with adverbial particles



(b) Phrasal verb with a prepositional phrase (accusative version)

Figure 3.11: Phrasal verb parse example

3.3.6 Auxiliaries

Since main objective of the research question is ergativity and accusativity in transitive sentences, auxiliary verbs have unconventional syntactic categories in order to put the argument structure of the main verb into focus. They are modeled as transitive verbs (with accusative and ergative variants) in case they are the main verb of the sentence.

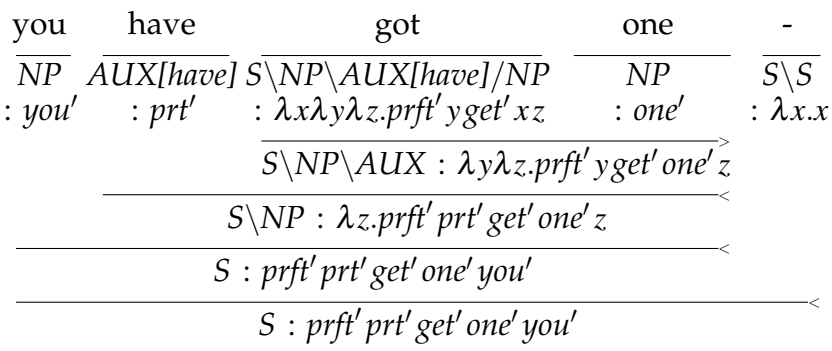
3.3.6.1 be (am, is, are, was, were, be, been)

Auxiliary verb to be and its variants (*am*, *is*, *are*, *was*, *were*, *been*) are assigned a category AUX with lexical constraints [type=be] and agreement. Since they are given lexical constraints, they can only be used in the presence of verbs in continuous (-ing) form. Their logical form is the corresponding tense (prt', pst', ftr' or gng').

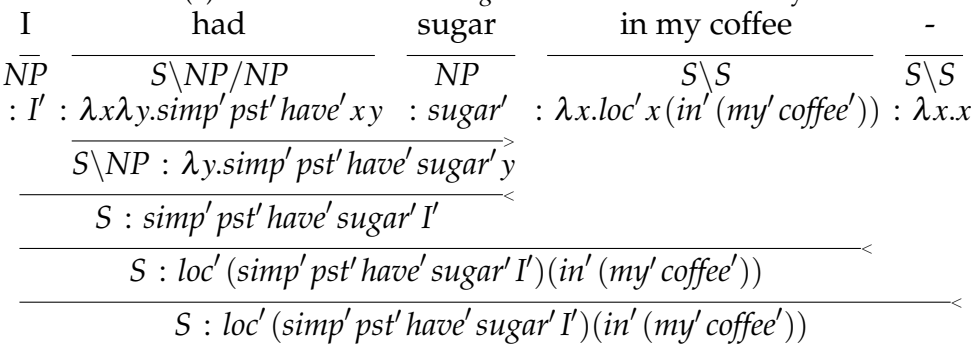
Verb to be is also used as the main verb in transitive affirmative sentences and questions. Therefore all different forms are also assigned a transitive syntax and logical structure with eq'.

3.3.6.2 have (has, have, had)

To have and its variants (*has*, *had*) are assigned similarly an AUX category with lexical constraints [type=have] and agreement. The corresponding logical form is their tenses. They are used in the derivations only if a past participle is present. An example sentence with *have* auxiliary is given in Figure 3.12a (lexical constraints are simplified).



(a) Accusative version of *got* used with *have* auxiliary



(b) *Have* used in an utterance as the main verb - accusative

Figure 3.12: Example uses of *have*

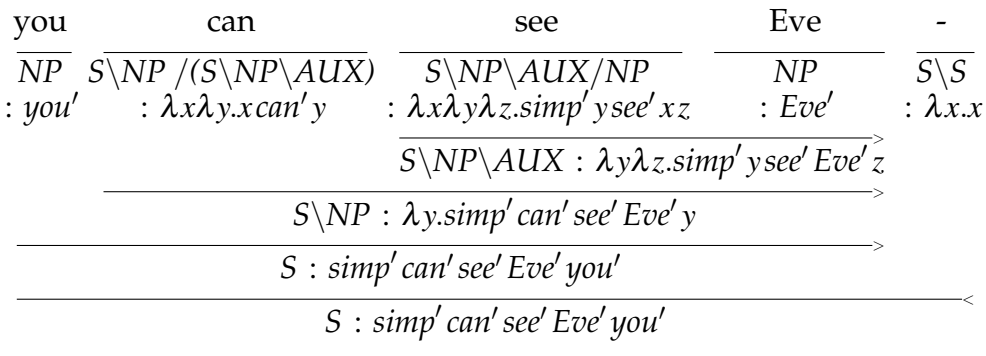
To have is also used as a transitive main verb in the corpus as seen in Figure 3.12b.

3.3.6.3 temp (do, does, did, will, going-to, gonna)

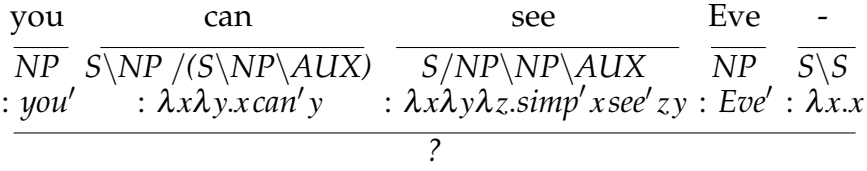
These auxiliaries are used to indicate the tense of the sentence mostly in questions. Therefore, they were assigned AUX category with lexical constraints [type=temp] and agreement; and a logical form of tense. *do*, *does* and *did* also have transitive main verb entries.

3.3.7 Modals

Modals (e.g., *can*, *would*, *shall*) take a verb phrase and a noun phrase to form a sentence.



(a) Accusative version



(b) Ergative version

Figure 3.13: Modal verb parse example

Note that the assumption that the modal verbs take a "verb phrase" makes it unparseable for ergative entries as shown in Figure 3.13b. The accusative version of the same verb parses seamlessly (Figure 3.13a)

3.3.8 Question Words

Question words include a Q' symbol as a requirement of the question representation. Also, they have some other symbols to indicate the type of the question (e.g., (Q' *person′*) for "who"). Question words are parsed with a particular entry that does not classify in ny of the accusative or ergative groups because the SVO order is disturbed. They are assigned a syntactic category of NP[type=qw] but they are also type-raised when necessary with a lexical rule. An example parse can be seen in Figure 3.14 (lexical constraints are simplified to the relevant ones).

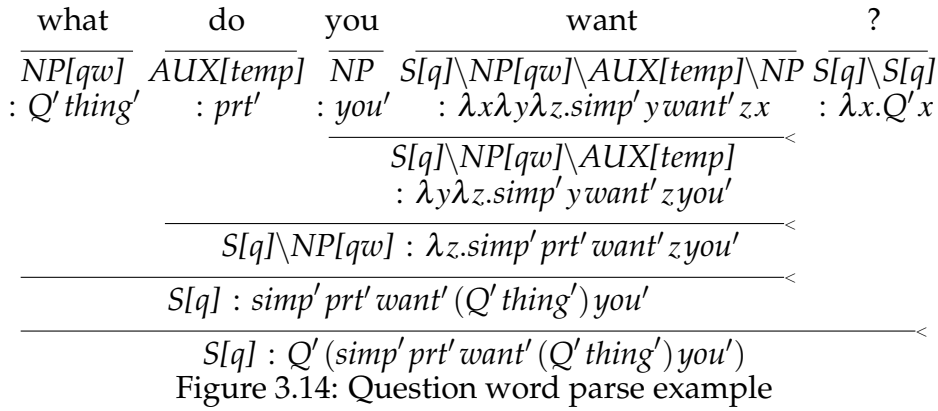


Figure 3.14: Question word parse example

3.3.9 Punctuation

In CHILDES Eve Corpus, there are only two punctuation marks: period(.) and question mark(?). They are both used to mark the end of sentences and non-sentential utterances. Lexical entries for the punctuation can be seen in Figure A.4 in Appendix A

For sentential utterances, period does not make any change in the meaning, while question mark adds the Q' at the beginning to indicate that the utterance consists of a question. This also helps turning the informative sentences turned into questions by intonation.

Non-sentential utterances (e.g., *lunch .* or *supper ?*) are represented with a dummy "it is" that I assume corresponds to some event or entity in the verbal and visual context of the child. Also, in case the utterance consists of a verb phrase with a continuous form (e.g. *eating what ?*), the question mark adds "you are" instead of "it is" because the child directed speech implies this meaning. I did not further resolve the pragmatic references. This way, the utterance *good .* is assigned a logical form that corresponds to *(it is) good .*, *a hammer ?* is assigned a logical form that corresponds to *(is it) a hammer ?*, *eating what ?* is assigned a logical form that corresponds to *(you are) eating what ?*, *eating popcorn .* is assigned a logical form that corresponds to *(you are) eating popcorn .*, *paper-clip ?* is assigned a logical form that corresponds to *(is it a) paper-clip ?* and so on.

In the experiments, I assume that the child can semantically interpret pauses between utterances as the end and the rising intonation in the utterance as a question. The punctuation marks in the corpus corresponds to these phonetic features.

CHAPTER 4

EXPERIMENT AND RESULTS

4.1 Aim and Assumptions

The aim of this simulated experiment is to measure the parameter difference that occurs after the training of a PCCG grammar, in which SVO-SVO' entries are assigned equal initial parameters, with a corpus of a SVO language. However, language acquisition of babies is a challenging process that involves many more steps. By focusing on accusative-ergative alignment, I made following assumptions:

- The main verbal input for the infant's language acquisition is the utterances of the mother and this input is sampled in the CHILDES Eve dataset.
- The child has already acquired the ability to divide mother's utterances into lexical units and can identify these units in her acquired lexicon.
- The infant has an innate or acquired faculty of semantic concepts like agent, object, past, continuing action etc. and can represent events and entities in her environment using these concepts. Logical forms used in the experiment are an approximation for the mental representation of the infant.
- The child can resolve the intonational clues in the mother's utterances and understand whether the utterance is a question or not with the help of these clues. The punctuation (.) and (?) represents this information in the corpus.
- The mother's utterances are relevant to the events occuring in the infant's environment. (The mother takes something when she says *I took it*.) There might be events irrelevant to the mother's utterances, but this possibility is modeled well in Abend et al. (2017) by adding distractors. So I have not added distractors in my training pairs.
- Learning of novel words, parsing of novel sentences, noun-verb distinction, acquisition of SVO among six syntactic word orders possible on the surface (SVO, SOV, VSO, VOS, OVS, OSV) are all modeled in Abend et al. (2017).

This experiment is about acquiring one of the two alternative syntactic categories for transitive/ditransitive verbs having the same surface form and leading to the same semantics through different syntactic trees.

4.2 CCGlab

CCGlab is a Common LISP based CKY parser for experimenting with CCG and PCCG grammars (Bozsahin, 2019). CCGlab can be used for grammar testing, model testing and parameter estimation in a model. CCGlab's main input files are .ccg grammar file for grammar testing, .ind and .sup file for parameter estimation. .ind file initiates initial parameters for parameter estimation and .sup file contains the training pairs. The notation in .ccg grammar is similar to the ones in academic papers as shown in Figure 4.1. Logical forms are curried by default if not stated otherwise. Lexical constraints are indicated in square brackets (e.g., [agr=3s]).

(a) An example ccg style lexical entry from Abend et al. (2017)
like $\vdash$ (S\NP)/NP : $\lambda x \lambda y.like(y,x)$
(b) Same entry in (a) in .ccg notation
like := (S\NP)/NP : \x\y. !like y x ;

Figure 4.1: .ccg file notation for a lexical entry

For the experiment in this thesis I used CCGlab version 5.2 that introduced normal form parsing. For parameter estimation, CCGlab uses the PCCG paradigm introduced by Zettlemoyer and Collins (2005) and similarly uses the number of times a lexical entry is used in a tree as the only feature.

4.3 Experimental Materials

In order to simulate incremental language learning of the infant, I will assume that the child has access to the surface structure of the utterance as the phonetic/orthographic input, and a structured semantics constructed through observations and an innate faculty of structured representation. Accordingly, I had to generate a meaning representation for all entries in the Corpus in order to train the initially SVO-SVO' neutral grammar. Logical forms for the utterances and the neutral grammar are constructed according to the methods explained in Chapter 3. For the training, all lexical items in the lexicon were assigned an initial parameter of 0.5.

4.4 Experiment (Training)

The algorithm described in Figure 2.11 was applied with learning parameters $N = 10, 20, 60, 80$ and 100, $\alpha_0 = 1.0$ and $c = 1.0$ (note that $n = 5134$ is the size of

the unified corpus). Experiment was conducted using CCGlab with a SBCL (Steel Bank Common LISP) environment.

4.5 Results

After conducting the training with N=10, 20, 60, 80 and 100 final parameters for verbal entries that are different from 0.5 (initial value) are seen in Figure 4.3. As N increases, final parameters get further away from the initial 0.5 line. The change in distance for each iteration decreases as N approaches 100. Therefore we can say that the gradient has decreased and is about to settle in the values in Figure 4.3.

Using the final parameters for N=100, I have conducted a binary logistic regression test with 0=Ergative and 1=Accusative using SPSS (Version 25). In this analysis, I only included final parameter values different than 0.5 for accusative (marked with >) and ergative (marked with <) entries of transitive and ditransitive verbs. The class (accusative/ergative) of the lexical entries predicted their final parameter, $b = 0.084$, Wald $\chi^2(1) = 59.269$, $p < 0.01$. However, the regression model increased the accuracy from the baseline 56.4% to 58.4%. Variables in the equation are given in Table 4.1.

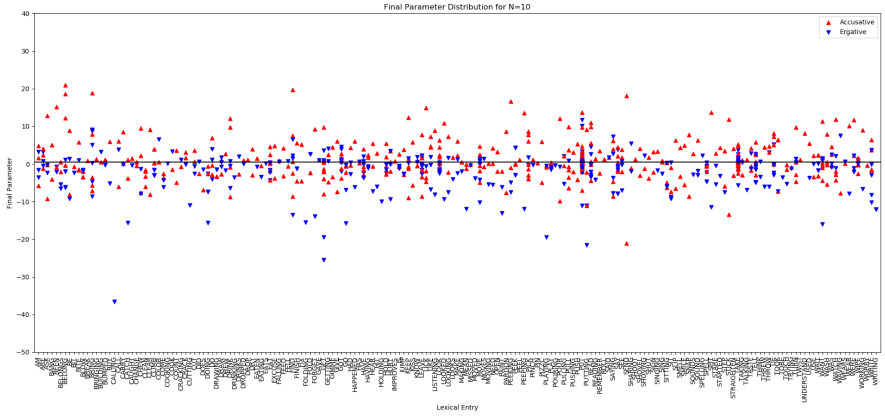
Table 4.1: Binary logistic regression analysis results for N=100 training

		95% CI for Odds Ratio		
	B (SE)	Lower	Odds Ratio	Upper
Included				
Constant	0.333 (0.069)			
Intervention	0.084 (0.011)	1.065	1.088	1.112

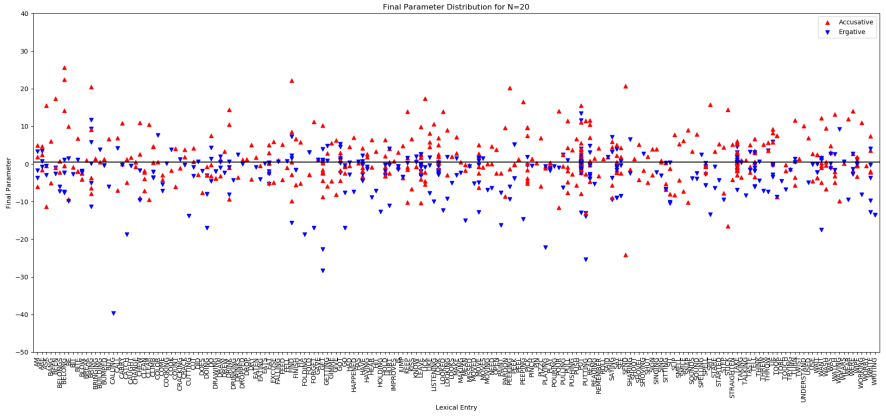
Note: R^2=.076 (Cox& Snell), .102 (Nagelkerke). Model $\chi^2(1)$=74.593, p<0.01

In Figure 4.3, the final parameters for accusative entries are indicated with upward (red) triangles and those for their ergative counterparts are indicated with downward (blue) triangles. Initial parameter 0.5 is indicated with a solid horizontal line. The figure shows that top entries for most of the verbs are accusative ones while the bottom entries are for the ergative ones. We can also see that there is a cluster of accusative and ergative lexical items in the middle around the initial value that needs further research. Top accusative cluster and bottom ergative cluster were expected since I was training the neutral grammar with a corpus of an accusative language. There might be several explanations for the central cluster.

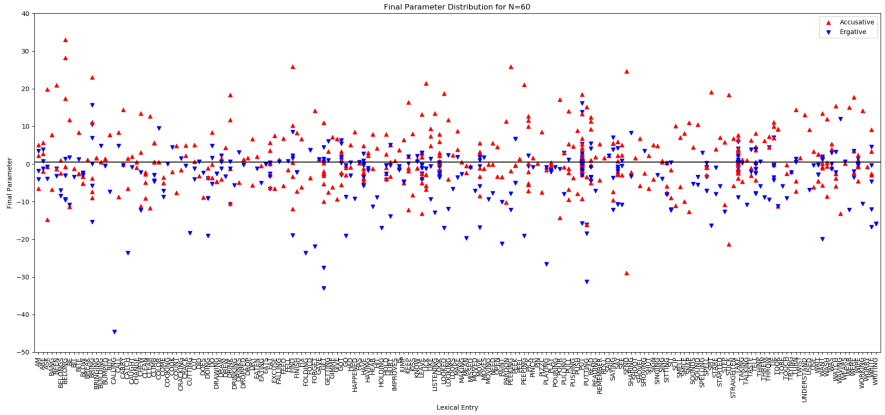
The first possible reason might be the insufficiency of the corpus size. The number of entries that can only be parsed with accusative entries might be insufficient to create a significant distinction in the entries of the central cluster. Since I used the corpus used by Abend et al. (2017) (that is about the half of the original corpus after omissions) and most of the coordination examples were divided into two different corpus entries, using the original corpus and uniting coordination and/or subordinating conjunction examples might help in getting a more decisive result.



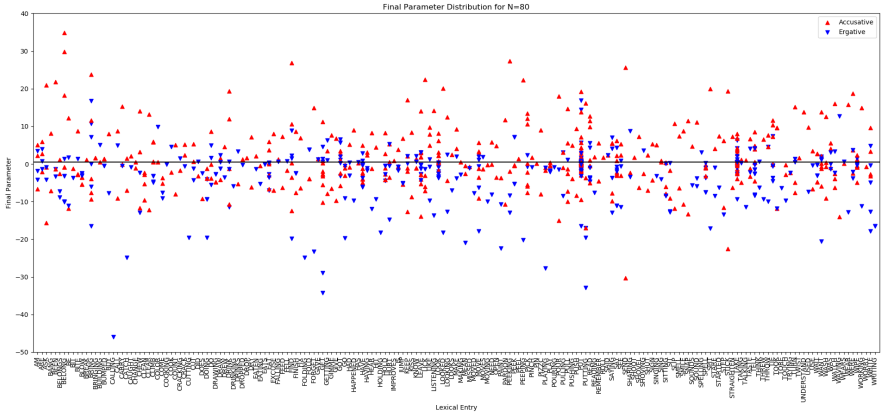
(a) Final parameters for N=10



(b) Final parameters for N=20



(c) Final parameters for N=60



(d) Final parameters for N=80

Figure 4.2: Final parameters for various values of N

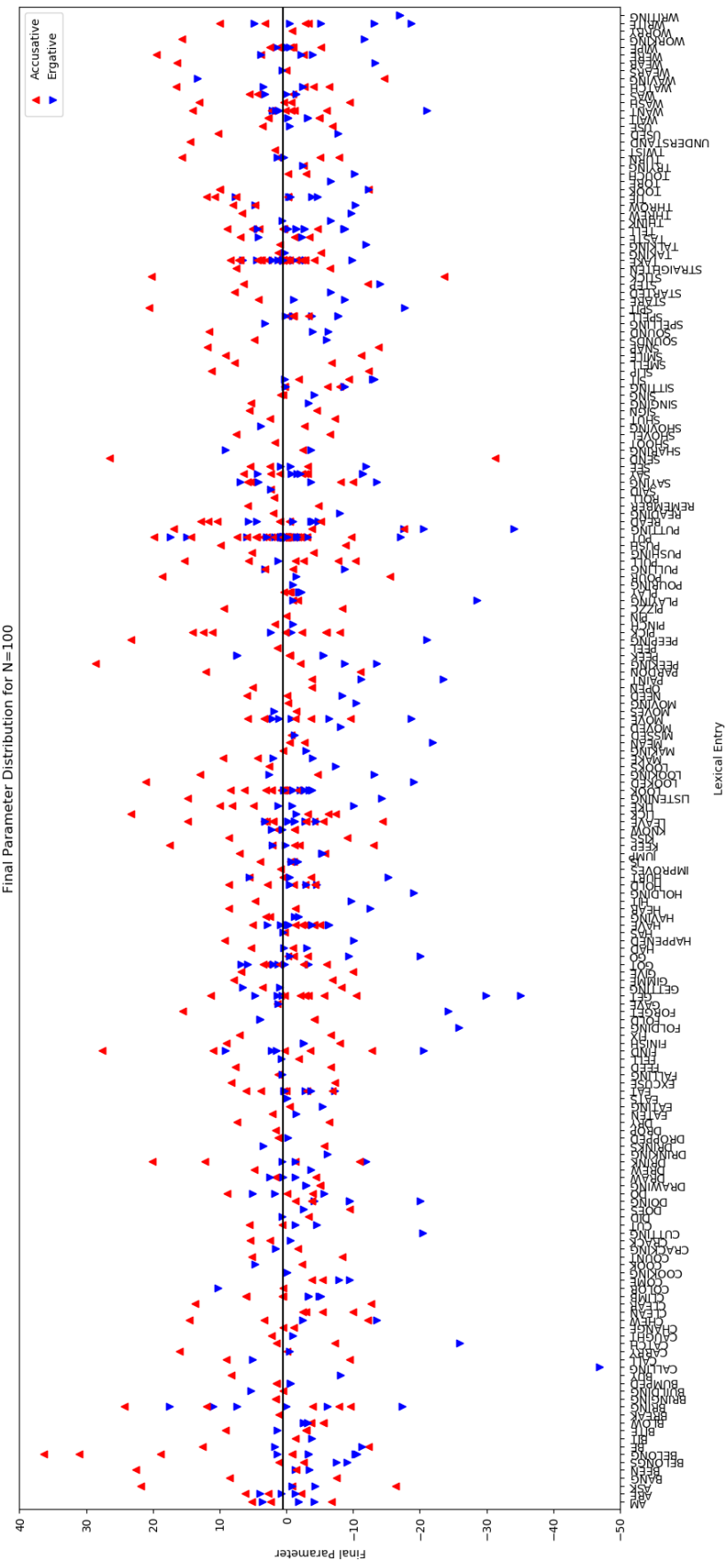


Figure 4.3: Final parameter distribution after training with N=100

The second possible reason might be the use of intransitive/transitive/ditransitive versions of the same verb. Ellipsis and similar pragmatic omissions of arguments were frequently used in the corpus. However, since our training model was based on utterance-meaning pairs, I did not seek to solve these pragmatic inferences and considered all entries in isolation. For instance, the verb *read* was used in all three versions of argument structure and therefore it had 26 entries with its intransitive, transitive, ditransitive, past form and phrasal verb entries. Similarly, the verb *put* was used mostly in phrasal verbs (like *put in*, *put away*) and in transitive and ditransitive versions. The multitude of the forms and phrasal verbs for the same symbol might have created unintended changes in the parameter estimation.

Third, for transitive and ditransitive verbs, there were separate accusative/ergative couples for three types of sentences to reduce the effects of the accusativity assuming structures of English (like auxiliaries): informative sentences with auxiliary, informative sentences without auxiliary and yes-no questions. Depending on the frequency of these types of sentences for each transitive verb, the parameter of *the couple* increases or decreases together. This creates the possibility that the ergative entry for informative sentences with auxiliary (very frequent in corpus) gets a higher final parameter than the accusative entry for yes-no questions (less frequent) for the same symbol, even though it is still less than the final parameter of the accusative entry for informative sentences with auxiliary. In the analysis of the results, all ergative and accusative entries are evaluated collectively. So, marking different accusative-ergative couples for the same string distinctively for the analysis could allow a more contrastive evaluation.

Another point worth mentioning is that the ergative English model I used in the grammars is (naturally) not completely ergative. Ergative alignment is defined based on the word order of *both* intransitive and transitive sentences. Considering that there is no case marking for noun phrases in English and that the case (or syntactic role) is solely determined based on the word order, one would need to switch the order of intransitive sentences to VS while preserving the AVO order for transitive sentences to create a truly ergative alignment for English. Similarly for the lexicon, even though the attachment order of transitive and ditransitive verb entries is modified, lexical entries for intransitive verbs do not have ergative counterparts reversed to VS because there is no utterance with VS order in the corpus. This experimental setup with a modified ergative English corpus in which intransitive sentences are reversed to VS order can lead to more interesting results and give more insight about the acquisition of ergative or accusative alignment. Since these further experiments require more time and substantial changes in the corpus, they are not included in the scope of this thesis.

CHAPTER 5

CONCLUSION AND FURTHER RESEARCH

Ergative-absolutive alignment is a syntactic feature in contrast to nominative-accusative alignment that aligns the subject of intransitive verbs with the object of transitive verbs. Ergative-absolutive structures are present in about 25% of world languages, but there is no language that is fully ergative, these languages are said to have *split ergative* systems (Dixon, 1994). There are two types of ergativity: syntactic ergativity and semantic one. I have considered syntactic ergativity in this thesis.

Learning of the ergativity or accusativity in verb-medial (SVO and OVS) languages poses a more serious challenge to the infant since the surface form is the same and the infant gets much less clues about the syntactic structure. However, there are competing models in the literature to explain the mechanisms underlying in the occurrence of syntactic features (Chomsky, 1993, Abend et al., 2017). In this thesis I have considered this issue using PCCG formalism and set up an experiment in order to measure the bias of an infant between accusative and ergative variants of transitive and ditransitive verbs as she is exposed to English sentences in the CHILDES Eve Corpus (Brown, 1973).

For the experiment, I made use of Probabilistic Combinatory Categorial Grammar defined in Zettlemoyer and Collins (2005), which is a variant of the Combinatory Categorial Grammar formalism (Steedman, 1996, 2000). Even though the acquisition of language-specific features has been experimented using CCG formalism (Abend et al., 2017), they have noted down the ergative versions in verb-medial syntactic orders and considered only six possible word orders in their experiments. Their experiments with Bayesian learning demonstrated that after less than 1000 utterances, the partial probability of the syntactic order the child is exposed to (SVO) prevails while all others are reduced, even in presence of distracting meanings around.

In my experiment, empirical bias of the infant's assumed SVO-SVO' neutral grammar after being exposed to the same corpus resulted in a cluster of high parameter SVO entries and another cluster of low parameter SVO' entries as well as a mixture of those in the middle. Even though an indirect correspondence is apparent in the resulting lexicon, the presence of the central cluster and singular inconsistencies justifies need for further research about the possible reasons and to reach a more decisive result.

There might be multiple reasons leading to this result: The size of the corpus might be insufficient to create a significant split. Also the way the corpus was transcribed (by separating coordination and subordinating conjunction examples in two entries) may have had an effect. Similarly, competition between intransitive, transitive and ditransitive entries and phrasal verb entries for the same verb could have interfered in the parameter estimation algorithm. Apart from these methodological issues, the occurrence of ergative-absolutive alignment may as well be the result of other linguistic mechanisms and not decided by parameters in the PCCG sense.

For further research, I can consider using a greater data set of the CHILDES corpus, experiment with other SVO languages, unite methodologically separated corpus entries with coordination or subordinating conjunction, and reconduct similar experiments while observing gradual changes in the parameters of competing accusative and ergative entries in order to reach more decisive and comprehensive results. Similarly, finding child-directed speech corpus for a verb medial language with ergatively aligned syntactic structures or modifying the Eve dataset to be truly ergative (with VS intransitive sentences) can be very helpful in understanding the nature of the acquisition of ergativity.

Bibliography

Abend, O., Kwiatkowski, T., Smith, N. J., Goldwater, S., and Steedman, M. (2017). Bootstrapping language acquisition. *Cognition*, 164:116–143.

Baker, J. K. (1979). Trainable grammars for speech recognition. *The Journal of the Acoustical Society of America*, 65(S1):S132–S132.

Bolinger, D. L. M. (1971). *The phrasal verb in English.* Cambridge, Mass., Harvard University Press, 1971.

Bozsahin, C. (2019). CCGlab manual. `https://github.com/bozsahin/ccglab/blob/master/docs/CCGlab-manual.pdf`. Last accessed on May 03, 2019.

Bozsahin, C. and Guven, A. B. (2018). Paracompositionality, MWEs and argument substitution. *CoRR*, abs/1805.08438.

Brown, R. (1973). *A First language: the early stages.* Cambridge, MA: Harvard University Press.

Chomsky, N. (1969). *Aspects of the Theory of Syntax*. The MIT Press. MIT Press.

Chomsky, N. (1993). *Lectures on Government and Binding : The Pisa Lectures.*, volume 7th ed of *Studies in Generative Grammar*. De Gruyter Mouton.

Chomsky, N., Belletti, A., Rizzi, L., and Chomsky, N. (2002). *On nature and language.* Cambridge ; New York : Cambridge University Press, 2002.

Clark, S. and Curran, J. R. (2003). Log-linear models for wide-coverage ccg parsing. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 97–104. Association for Computational Linguistics.

Dixon, R. M. W. (1994). *Ergativity*. Cambridge Studies in Linguistics. Cambridge University Press.

Karttunen, L. (1977). Syntax and semantics of questions. *Linguistics and Philosophy*, 1(1):3–44.

Kwiatkowski, T. M. (2012). *Probabilistic grammar induction from sentences and structured meanings*. PhD thesis, University of Edinburgh.

Macwhinney, B. (2000). The CHILDES project: tools for analyzing talk. *Child Language Teaching and Therapy*, 8.

Sagae, K., Davis, E., Lavie, A., MacWhinney, B., and Wintner, S. (2010). Morphosyntactic annotation of childes transcripts. *Journal of Child Language*, 37(3):705–729.

Steedman, M. (1996). *Surface structure and interpretation*. MIT press.

Steedman, M. (2000). *The syntactic process*, volume 24. MIT press.

Zettlemoyer, L. S. and Collins, M. (2005). Learning to map sentences to logical form: structured classification with probabilistic categorial grammars. In *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence*, pages 658–666. AUAI Press.

Appendix A

GRAMMARS

As the core of this thesis is the alignment of verbs, there are multiple lexical entries for all forms of each verb. Additionally, to reduce the effect of auxiliary verbs, main verb categories are designed to be the main "parser" of the sentences, taking auxiliary verbs to form informative sentences and yes-no questions. Intransitive verbs do not have accusative-ergative alternative categories but they have different entries for informative sentences with auxiliary, informative sentences without auxiliary, and yes-no questions. Transitive and ditransitive verbs have accusative (marked with >) and ergative (marked with <) entries for these settings. They also have an extra entry for wh-questions, but this entry does not have any accusative or ergative bias since the verb-medial syntactic order is not preserved.

Syntactic categories of other word types (nouns, adjectives, pronouns, prepositions, modal verbs, phrasal verbs etc.) are described in Chapter 3.

```
jump iv1 := s[type=inf]\np\aux[type=temp] : \x\y. !simp x !jump y;
jump iv1 := s[type=inf]\np : \x. !simp !prt !jump x;
jump iv1 := s[type=q]\aux[type=temp]\np : \x\y. !simp y !jump x ;
```

(a) Basic form

```
falling iving := s[type=inf]\np\aux[type=be] : \x\y. !cont x !fall y;
falling iving := s[type=q]\aux[type=be]\np : \x\y. !cont y !fall x;
```

(b) -ing form

```
fell iv2 := s[type=inf]\np : \x. !simp !pst !fall x;
```

(c) Past form

```
chirps ivs := s[type=inf]\np[agr=3,count=sg] : \x. !simp !prt !chirp x;
```

(d) 3rd person singular agreed form

```
gone iv3-away := s[type=inf]\np\aux[type=have]/"away" : \o\x\y. !prft x
(!go _ o) y ;
gone iv3-away := s[type=q]\aux[type=have]\np/"away" : \o\x\y. !prft y
(!go _ o) x ;
```

(e) Past participle form (the only one in the corpus)

Figure A.1: Example entries for intransitive verbs in various forms

```
get tv1> := s[type=inf]\np\aux[type=temp]/np : \x\y\z. !simp y !get x z;
get tv1< := s[type=inf]/np\np\aux[type=temp] : \x\y\z. !simp x !get z y;
get tv1> := s[type=inf]\np/np : \x\y. !simp !prt !get x y;
get tv1< := s[type=inf]/np\np : \x\y. !simp !prt !get y x;
get tv1> := s[type=q]\aux[type=temp]\np/np : \x\y\z. !simp z !get x y;
get tv1< := s[type=q]\aux[type=temp]/np\np : \x\y\z. !simp z !get y x;
get tv1 := s[type=q]\np[type=qw]\aux[type=temp]\np : \x\y\z. !simp y
!get z x;
```

(a) Basic form

```
building tving> := s[type=inf]\np\aux[type=be]/np : \x\y\z. !cont y !build
x z;
building tving< := s[type=inf]/np\np\aux[type=be] : \x\y\z. !cont x !build
z y;
building tving> := s[type=q]\aux[type=be]\np/np : \x\y\z. !cont z !build x
y;
building tving< := s[type=q]\aux[type=be]/np\np : \x\y\z. !cont z !build y
x;
building tving := s[type=q]\np[type=qw]\aux[type=be]\np : \x\y\z. !cont
y !build z x;
```

(b) -ing form

```
caught tv2> := s[type=inf]\np/np : \x\y. !simp !pst !catch x y;
caught tv2< := s[type=inf]/np\np : \x\y. !simp !pst !catch y x;
```

(c) Past form

```
improves tvs> := s[type=inf]\np[agr=3,count=sg]/np : \x\y. !simp !prt
!improve x y;
improves tvs< := s[type=inf]/np\np[agr=3,count=sg] : \x\y. !simp !prt
!improve y x;
```

(d) 3rd person singular agreed form

```
eaten tv3 := s[type=inf]\np\aux[type=be] : \x\y. !simp x !eat y !- ;
eaten tv3> := s[type=inf]\np\aux[type=have]/np : \x\y\z. !prft y !eat x z ;
eaten tv3< := s[type=inf]/np\np\aux[type=have] : \x\y\z. !prft x !eat z y;
eaten tv3 := s[type=q]\aux[type=be]\np : \x\y. !simp y !eat x !- ;
eaten tv3> := s[type=q]\aux[type=have]\np/np : \x\y\z. !prft z !eat x y ;
eaten tv3< := s[type=q]/np\aux[type=have]\np : \x\y\z. !prft y !eat z x ;
eaten tv3 := s[type=q]\np[type=qw]\aux[type=have]\np : \x\y\z. !prft y
!eat z x ;
```

(e) Past participle form

Figure A.2: Example entries for transitive verbs in various forms

```
take dv1> := s[type=inf]\np\aux[type=temp]/np/np : \x\y\z\a.!simp z !take y x a;
take dv1< := s[type=inf]/np/np\np\aux[type=temp] : \x\y\z\a.!simp x !take a z y;
take dv1> := s[type=inf]\np/np/np : \x\y\z.!simp !prt !take y x z;
take dv1< := s[type=inf]/np/np\np : \x\y\z.!simp !prt !take z y x ;
take dv1> := s[type=q]\aux[type=temp]\np/np/np : \x\y\z\a. !simp a !take y x z;
take dv1< := s[type=q]/np/np\aux[type=temp]\np : \x\y\z\a.!simp y !take a z x ;
take dv1 := s[type=q]\np[type=qw]\aux[type=temp]\np/np : \x\y\z\a. !simp z !take x a y;
```

(a) Basic form

```
putting dving> := s[type=inf]\np\aux[type=be]/np/np : \x\y\z\a.!cont z !put y x a;
putting dving< := s[type=inf]/np/np\np\aux[type=be] : \x\y\z\a.!cont x !put a z y ;
putting dving> := s[type=q]\aux[type=be]\np/np/np : \x\y\z\a.!cont a !put y x z;
putting dving< := s[type=q]/np/np\aux[type=be]\np : \x\y\z\a. !cont y !put a z x ;
putting dving := s[type=q]\np[type=qw]\aux[type=be]\np/np : \x\y\z\a. !cont z !put a x y;
```

(b) -ing form

```
gave dv2> := s[type=inf]\np/np/np : \x\y\z.!simp !pst !give x y z;
gave dv2< := s[type=inf]/np/np\np : \x\y\z.!simp !pst !give y z x;
```

(c) Past form

(d) There was no 3rd person singular agreed form for any ditransitive verb in the corpus.

```
put dv3 := s[type=inf]\np\aux[type=be]/np : \x\y\z. !simp y !put x z !- ;
put dv3> := s[type=inf]\np\aux[type=have]/np/np : \a\x\y\z.!prft y !put x a z ;
put dv3< := s[type=inf]/np/np\np\aux[type=have] : \x\y\z\a. !prft x !put a z y;
put dv3 := s[type=q]\aux[type=be]\np/np : \x\y\z. !simp z !put x y !- ;
put dv3> := s[type=q]\aux[type=have]\np/np/np : \a\x\y\z. !prft z !put x a y ;
put dv3< := s[type=q]/np\aux[type=have]\np : \x\y\z. !prft y !put z x ;
put dv3 := s[type=q]\np[type=qw]\aux[type=be]\np : \x\y\z. !simp y !put z x !- ;
put dv3 := s[type=q]\np[type=qw]\aux[type=have]\np/np : \a\x\y\z.!prft y !put a z x ;
```

(e) Past participle form

Figure A.3: Example entries for ditransitive verbs in various forms

Note that as auxiliary verbs are used with the basic form (*temp* auxiliaries), -ing form (*be* auxiliaries) and the past participle form (*have* auxiliaries), they have entries to handle auxiliary verbs. The past participle form also constructs passive sentences when used with *be* auxiliary (only a few sentences in the corpus). Other versions do not have entries with auxiliary verbs.

```
- pun := s[type=inf]\*s[type=inf] : \x.x;
- pun := s[type=imp]\*(s\np) : \x.!simp !imp (x !you);
- pun := s[type=inf]\*(s\np\aux[type=temp]) : \x. x !imp !you;
- pun := s[type=inf]\*(s\np\aux[type=be]) : \x. x !prt !you;
- pun := s[type=inf]\*np : \x.!simp !prt !eq x !it;
- pun := s[type=inf]\*n : \x.!simp !prt !eq x !a !it;
- pun := s[type=inf]\*(np/np) : \x.!simp !prt x !it;
```

(a) Lexical entries for the period(.)

```
? pun := s[type=q]\*s[type=inf] : \x.!q x;
? pun := s[type=q]\*s[type=q] : \x.!q x;
? pun := s[type=q]\*(s\np\aux) : \x.!q (x !prt !you);
? pun := s[type=q]\*np : \x.!q (!simp !prt !eq x !it);
? pun := s[type=q]\*n : \x.!q (!simp !prt !eq x !a !it);
? pun := s[type=q]\(s\np) : \x. !q (x !you) ;
```

(b) Lexical entries for the question mark(?)

Figure A.4: Lexical entries for punctuation

Appendix B

LOGICAL FORMS

Table B.1: Example training pairs for informative sentences

Preprocessed corpus entry	Logical form
THAT IS THE GIRL -	((("SIMP" "PRT") "EQ") ("THE" "GIRL")) "THAT")
WE WILL HAVE MILK FOR LUNCH -	("AIM" ((("SIMP" "FTR") "HAVE") "MILK") "WE")) ("FOR" "LUNCH"))
WE HAD BREAKFAST -	((("SIMP" "PST") "HAVE") "BREAKFAST") "WE")
YOU GET ONE -	((("SIMP" "PRT") "GET") "ONE") "YOU")
I SEE THAT BUTTON -	((("SIMP" "PRT") "SEE") ("THAT" "BUTTON")) "I")
MAMA IS FIXING IT -	((("CONT" "PRT") "FIX") "IT") "MAMA")
EVE WILL READ LASSIE -	((("SIMP" "FTR") "READ") "LASSIE") "EVE")
THEY ARE SPLASHING -	(("CONT" "PRT") "SPLASH") "THEY")
YOU HAVE A NICE NAP -	((("SIMP" "PRT") "HAVE") ("A" ("NICE" "NAP"))) "YOU")
I HAVE GOT YOU NOW -	(("TIME" ((("PRFT" "PRT") "GET") "YOU") "I")) "NOW")
YOU ARE GOING BACKWARDS -	(("LOC" (("CONT" "PRT") "GO") "YOU")) "BACKWARDS")
THERE IT GOES -	(("LOC" (("SIMP" "PRT") "GO") "IT")) "THERE")
PAPA MIGHT GIVE YOU A CRACKER -	(((("SIMP" "MIGHT") "GIVE") ("A" "CRACKER")) "YOU") "PAPA")
YOU ARE GONNA FALL -	(("SIMP" ("PRT" "GNG")) "FALL") "YOU")
IT IS A FORK -	((("SIMP" "PRT") "EQ") ("A" "FORK")) "IT")

Table B.2: Example training pairs for questions

Preprocessed corpus entry	Logical form
WHAT DID YOU DO ?	("Q" ((("SIMP" "PST") "DO") ("Q" "THING")) "YOU"))
WHERE IS CROMER ?	("Q" ((("SIMP" "PRT") "EQ") ("Q" "PLACE")) "CROMER"))
DID FRASER USE THE SUGAR ?	("Q" ((("SIMP" "PST") "USE") ("THE" "SUGAR")) "FRASER"))
YOU SPILLED IT ?	("Q" ((("SIMP" "PST") "SPILL") "IT") "YOU"))
WILL EVE READ FRASER LASSIE ?	("Q" (((("SIMP" "FTR") "READ") "LASSIE") "FRASER") "EVE"))
WHAT ELSE HAVE YOU BEEN DOING ?	("Q" ((("CONT" ("PRFT" "PRT")) "DO") (("Q" "THING") "ELSE")) "YOU"))

Table B.3: Example training pairs for non-sentential utterances

Preprocessed corpus entry	Logical form
YOUR TRUCK ?	("Q" ((("SIMP" "PRT") "EQ") ("YOUR" "TRUCK")) "IT"))
CHOCOLATE-ICECREAM -	((("SIMP" "PRT") "EQ") "CHOCOLATE-ICECREAM") "IT")
THE BOWL -	((("SIMP" "PRT") "EQ") ("THE" "BOWL")) "IT")

Table B.4: Example training pairs for imperative sentences

Preprocessed corpus entry	Logical form
STRAIGHTEN THE RUG -	((("SIMP" "IMP") "STRAIGHTEN") ("THE" "RUG")) "YOU")
CHEW IT UP -	((("SIMP" "IMP") (("CHEW" _) "UP")) "IT") "YOU")
PUT HER IN THE BASEMENT -	(((("SIMP" "IMP") "PUT") ("IN" ("THE" "BASEMENT"))) "SHE") "YOU")
EAT THEM WITH YOUR SPOON -	(("INST" ((("SIMP" "IMP") "EAT") "THEY") "YOU")) ("WITH" ("YOUR" "SPOON")))

Appendix C

CORRECTED/MODIFIED ENTRIES IN THE CORPUS

Some entries in the corpus are corrected or modified in order to avoid ad-hoc lexical entries or to simplify parsing. Some punctuation mistakes are corrected or a hyphen (-) is added at the end as a period (.) in case there was no punctuation.

Table C.1: Modified entries in the corpus

Original	**Modified/Corrected**
you having juice ?	you are having juice ?
coffee you are not having coffee -	you are not having coffee -
change your record would you ?	would you change your record ?
would you bring Mama your cup -	would you bring Mama your cup ?
catch the ball catch the ball -	catch the ball - catch the ball -
would you shut the door -	would you shut the door ?
would you step back -	would you step back ?
you drinking your milk ?	you are drinking your milk ?
that is a girl ? -	that is a girl ?
you you tell me about it ?	you tell me about it ?
is that Racketyboom -	is that Racketyboom ?
what -	what ?
where is the butterfly -	where is the butterfly ?
it is	it is -
Eve	Eve -
is not that funny -	is not that funny ?
Eve	Eve -
Eve	Eve -
shall we change your diaper -	shall we change your diaper ?
we will we will have a letter -	we will have a letter -
you put back stool back -	you put the stool back -
would you move into the room -	would you move into the room ?
would you bring the napkin -	would you bring the napkin ?
Eve	Eve -
what do you want me to do -	what do you want me to do ?
what -	what ?
cows go mooo mooo-	cows go mooo -

Continued on next page

Table C.1 – continued from previous page

Original	Modified/Corrected
would you put it back -	would you put it back ?
a truck going -	a truck is going -
smell flower -	smell the flower -
do you want a napkin too -	do you want a napkin too ?
dog barking ?	dog is barking ?
your name is what -	your name is what ?
what do you want me to do -	what do you want me to do ?
he has pipe -	he has a pipe -
what is is fraser doing ?	what is fraser doing ?
and Fraser they have not used it yet -	and they have not used it yet -
what -	what ?
you you put you in the wastebasket -	you put you in the wastebasket -
what -	what ?
what is it -	what is it ?
do not you peepee -	do not peepee -
what did you do my ?	what did you do ?
are you standing on the board -	are you standing on the board ?
do not you peepee -	do not peepee -
that is who -	that is who ?
is who -	is who ?
i said good day eve -	i said good-day-eve -
Eve do you have some glasses -	do you have some glasses ?
you be careful -	be careful -
old friends meeting once again -	old friends are meeting once again -
who is that -	who is that ?
whose daughters -	whose daughters ?
do not you make it -	do not make it -
what do you want Sarah ?	what do you want ?
what -	what ?
he working -	he is working -
you was gone away -	you were gone away -
she is Sarah listening to the story -	she is listening to the story -
say girl Eve	say girl -
Eve Eve -	Eve -
or do you have clean feet -	or do you have clean feet ?
they eating lunch -	they are eating lunch -
what was it what did you do ?	what was it ? what did you do ?
what what what ?	what ? what ? what ?
you sharing it with Sarah -	you are sharing it with Sarah -
what are you doing -	what are you doing ?
you get pencil -	you get a pencil -

Continued on next page

Table C.1 – continued from previous page

Original	Modified/Corrected
what am i almost finished?	what ? am I almost finished ?
thumb thumb ?	thumb - thumb ?
do not you shoot me -	do not shoot me -
no not England -	no - not england -
what were doing ?	what were you doing ?
what is the wise idea -	what is the wise idea ?
milk egg salt -	milk - egg - salt -
what is the wise idea -	what is the wise idea ?
where is the pitcher -	where is the pitcher ?
el-vl-el Eve -	el-vl-el - Eve -
what is the baby doing -	what is the baby doing ?
is not that awful -	is not that awful ?
down we go -	we go down -

Table C.2: Lexicalized compound nouns in the corpus

Original	**Unified**
grape juice	grape-juice
cheese sandwich	cheese-sandwich
noodle soup	noodle-soup
icecube trays	icecube-trays
apple trees	apple-trees
boullion cubes	bouillon-cubes
baseball coin	baseball-coin
paper bag	paper-bag
paper clip	paper-clip
rockabye baby	rockabye-baby
al bl cl	al-bl-cl
icecream cone	icecream-cone
bowel movement	bowel-movement
box top	box-top
cake plate	cake-plate
birthday sandwich	birthday-sandwich
birthday cake	birthday-cake
vitamin time	vitamin-time
mud pies	mud-pies
rice soup	rice-soup
chicken bone	chicken-bone
bath mat	bath-mat
vegetable soup	vegetable-soup
bus stop	bus-stop
camera spool	camera-spool
chocolate cookie	chocolate-cookie
baby sister	baby-sister
lobster salad	lobster-salad
dingdong dell	dingdong-dell
nursery rhyme	nursery-rhyme
cookie press	cookie-press
tomato sandwich	tomato-sandwich
peanutbutter sandwich	peanutbutter-sandwich
sock slippers	sock-slippers
el vl el	el-vl-el