Tel-Aviv University
Lester and Sally Entin Faculty of the Humanities
Department of Linguistics

Like Father, Like Son, Like Father: Investigating Developmental Stages in Child-Directed Speech

M.A Thesis submitted by

Stav Klein

Under the supervision of
Prof. Ruth Berman

October 2021

Abstract

Empiricist studies on language acquisition tend to focus on the effects of the parental input on the child and how this input can be used as proxy to the child's developmental stage. This study takes the opposite approach by focusing on the effects the developmental stage of the child have on the parental input.

Essentially, the study starts with suggesting developmental stages for the CDS data based on the same measure used to determine stages in CS. Then it examines the distribution of morpho-syntactic elements in each stage, with respect to the age-of-acquisition of these elements in children. Findings show that the suggested adult-stages are very sensitive to the current developmental stage of the child, and that the distributions of the morpho-syntactic elements within these stages can serve as precursors to which morpho-syntactic elements will be acquired (if acquired at all). This study also offers a usage-based account for the absolute minimum number of occurrences that must be provided by the adult in order to acquire an element.

Acknowledgment

"It takes a village..."

First and foremost I want to express my deepest appreciation and gratitude to my advisor, Prof. Ruth Berman, who sticks with me through the rotten and the bliss of this very long journey, supporting me every step of the way, always having a useful insight or advice and overall being the most patient and open-minded scholar I have ever met.

I also wish to thank Prof. Omri Abend from the Hebrew University of Jerusalem for sharing with me their child-directed speech corpus, which was unpublished at the time, and for Prof. Reut Tsarfaty for introducing us.

I was so fortunate to study in the Linguistics dept. I thank Ruti Zussman for her company and working her administrative magic, Prof. Aya Meltzer-Asher, Prof. Einat Shetreet and their lab members for their tremendous help with my thesis defense and Prof. Evan-Gary Cohen and all the participants in my colloquium whose helpful comments are incorporated in this work.

I want to thank all my friends, for still being my friends - Noa Geller, Alon Fishman, Nicole Katzir, Hezi Shabanov, Avital Zaruvimsky, Daniel Asherov, Tal Ness - I could not have done it without your moral support. A special thanks goes to Hila Davidovitch and Maya Ochayon for their exceptional encouragement, holding my hand and pushing me - literally and figuratively.

Finally, I wish to thank my family - my mother Rachel and my sisters Rozi and Dvir for their support and sincere attempts to understand what this work is about, and to my awesome partner Sivan and my two beautiful boys Guy and Yuval, my pride and joy, for giving me a reason to get out of bed every. single. morning. at 5am. This is for you.

Contents

Abstract ii

Acknowledgment iii

Contents iv

List of Figures vi

List of Tables viii

1 Introduction 1

2 Method 4
- 2.1 Research Questions . . . 4
- 2.2 Research Hypotheses . . . 5
- 2.3 Validation . . . 6
- 2.4 Measure of Complexity - MPU . . . 7

3 Data 9
- 3.1 Child Speech . . . 9
- 3.2 Child-directed Speech . . . 11
  - 3.2.1 Berman's Original Project . . . 11
  - 3.2.2 Hagar's CDS Re-analyzed . . . 12
- 3.3 The Data in CHILDES . . . 15

4 Defining Adult Stages 17

4.1 MPU Clustering 17

4.2 Pre-processing 18

4.3 Results 18

5 Developmental Evidence for Stages 21

5.1 Pre-processing 21

5.2 Results 22

5.2.1 Upper Bound 26

5.2.2 Lower Bound 27

5.3 Summary 28

6 Limitations of the Study 31

6.1 Case-study 31

6.2 Automated Analysis 32

6.3 Built-in Corpus 33

7 Discussion 35

A Supplementary Material - CS 39

A.1 Kaplan's Results - Exhaustive List of Correct Usage 39

A.2 Results for all morpho-syntactic elements 42

A.2.1 Age 1;9 - 2;0 (21-24 months) 42

A.2.2 Age 2;1 - 2;3 (25-27 months) 43

A.2.3 Age 2;4 - 2;6 (28-30 months) 44

A.2.4 Age 2;7 - 3;0 (31-36 months) 47

A.2.5 Acquired by 3;6 (42 months) 50

A.2.6 Not acquired by 3;6 54

Bibliography 56

List of Figures

3.1 Gradual and U-Shaped Learning 10
3.2 An example sentence analysis 13
3.3 Dependency relations between words in a sentence 15
3.4 Example sentence from the Hebrew CHILDES corpora . . . 16

4.1 MPU of Hagar's CDS according to Hagar's age 19
4.2 Child stages and adult stages projected on CDS MPU data . 19
4.3 Adult stages and child stages combined 20

5.1 Density of adult stages 23
5.2 Regular conjunction, example and distribution 24
5.3 Future forms to convey imperative meaning, example and distribution 26
5.4 Present tense verbs, example and distribution 27
5.5 Inflected ʕal, example and distribution 28
5.6 Distribution of occurrences of elements 29

A.1 40
A.1 Percentage of correct usage of elements by Kaplan 41
A.2 Present tense verbs, example and distribution 42
A.3 Past tense verbs, example and distribution 43
A.4 Imperative verbs, example and distribution 43
A.5 Future forms to convey imperative meaning, example and distribution 44
A.6 'In + the', example and distribution 44
A.7 Isolated 'with'(preposition), example and distribution 45

A.8 shel (possessive preposition), example and distribution . . . 45
A.9 et (accusative marker) + pronoun, example and distribution 45
A.10 I (1sc pronoun), example and distribution 46
A.11 He (3ms pronoun), example and distribution 46
A.12 Initial conjunction marker, example and distribution 46
A.13 Regular conjunction, example and distribution 47
A.14 'ki' for reasoning, example and distribution 47
A.15 with + pronoun, example and distribution 48
A.16 on/about in isolation, example and distribution 48
A.17 'In' in isolation, example and distribution 49
A.18 Plural nouns (masculine siffix), example and distribution . . 49
A.19 Future tense verbs, example and distribution 49
A.20 Plural nouns feminine suffix, example and distribution . . . 50
A.21 Future plural adjectives masculine suffix, example and distribution 50
A.22 Preposition 'betox' (inside), example and distribution 51
A.23 Preposition 'me' (from), example and distribution 51
A.24 Preposition 'kmo' (like), example and distribution 52
A.25 Preposition 'be' (in) + pronoun, example and distribution . 52
A.26 Pronoun 'hi' (3fs), example and distribution 52
A.27 Pronoun 'hem' (3mp), example and distribution 53
A.28 Complementizer 'kshe' (when) , example and distribution . . 53
A.29 Inflected ʕal, example and distribution 54
A.30 Plural adjectives (feminine suffix) , example and distribution 54
A.31 Preposition 'mitaxat' (under) , example and distribution . . 55

List of Tables

2.1 Summary of mean MPU for age group in Kaplan's study . . . 8

3.1 Summary of the data in the 'Berman Longitudinal' corpus in CHILDES . . . 12

5.1 Number of utterances per CDS stage . . . 22

5.2 Differences between late-acquisition and no-acquisition . . . 30

Chapter 1

Introduction

> *"Motherese is as likely an effect on the mother by the child as an effect on the child by the mother"* (Newport et al., 1977)

How children acquire language can easily be considered the holy grail of modern linguistics and stands in the core of the endless nature-vs-nurture debate. While this work does not address the nature-nurture debate explicitly, it does belong to the nurture side of it.

As the opening quote eloquently puts - Motherese (or Child-directed Speech (CDS), as it is called nowadays) is likely a two-way street. Yet, most of the studies on CDS done so far focus on the effects of the parent on the child. They study the properties of CDS as a whole (Ravid et al., 2016; Arnon, 2016) and how these properties affect and maybe facilitate language acquisition in children (Arnon, 2021; You et al., 2021; Hiller and Fernández, 2016). This study, on the other hand, focuses on the bi-directional influence that child-directed speech and child-speech have on each-other. The motivation behind this study is multi-fold; first, as mentioned, this line of work is understudied and is full of potential to shed new light on the parent-child interaction and the role of input from the environment on the acquisition process, beyond the pathological ways in which poor and inadequate input can harm proper language development (Tomasello, 1992; Bloom, 2002; Hollich et al., 2000; Nelson, 2009; Cristia et al., 2019). Second, it is well known and well established that CDS is very different in nature from adult-directed speech (see Phillips, 1970; Remick, 1971; Snow, 1972, among others), and there is also evidence that CDS is specifically designed for a learner (You et al., 2021). This calls for a different analysis of CDS, instead of studying CDS in the same methods used for the 'regular' adult-to-adult speech. The other side of that argument is that CDS is not only very different than adult-to-adult speech, but it is also similar in certain ways to child-speech. For example, both CS and CDS are known to change over time (Newport et al., 1977; Tal et al., 2021), and there is a strong correlation between the parental input frequencies and the children's language acquisition (Moerk, 1980; Tomasello, 1992; Hiller and Fernández, 2016). Taken together, these similarities and connections between CDS and CS call for a new bidirectional approach to CDS analysis.

The approach suggested in this study is a step in that direction. It starts with the influence of the child on the parent. Specifically it examines how adult speech, in the form of CDS, changes with respect to the developmental stage of the child at the same point in time. Then, in the other direction, it shows how these developmental stages in adults can serve as indicators for the age-of-acquisition of certain elements in children. The findings show that not only that CDS changes drastically during the acquisition process, but rather that CDS can (and perhaps should) be described as a series of developmental stages that correspond to the developmental stages thoroughly studied in children (as in Kaplan, 1983; Ben-David and Bat-El, 2016; Lustigman, 2015; Berman, 2004, and more).

This study adopts the general framework of social interactionism, which is empiricist by nature. It places a greater burden on the learning experiences of the child during the acquisition than on innate linguistic knowledge and whether such knowledge exists or not. Such focus on learning from experience requires identifying domain-general learning mechanisms, one of which is statistical learning, as demonstrated in this study. Although it was claimed by Chomsky that the input from the environment is erroneous and insufficient for grammar induction (Chomsky et al., 2006), it was shown in several cases that CDS is in fact very intelligible compared to adult-directed speech (Newport et al., 1977), and that children can use it to detect distributions and form categories around frequently occurring signals (Maye et al., 2002). They can also learn conditional probabilities (Aslin et al., 1998), transition probabilities (Pelucchi et al., 2009) and can do so in different domains, such as phonology, lexicon, morphology and syntax (McMurray and Hollich, 2009; Saffran, 2009; Arnon, 2015). From the CDS perspective, it was shown that words from different grammatical categories actually appear in significantly different frames (Mintz, 2003, 2006), which facilitates tracking their distribution and therefore facilitates learning. This study focuses on the changes in the distribution of morpho-syntactic elements in CDS during the course of acquisition, and shows how these changes are not random but rather very accurately fine-tuned to the morpho-syntactic development in children.

This work is organized as follows; Chapter 2 elaborately describes the research questions and the way they are being validated in this study. Chapter 3 reviews the CS and the CDS corpora used in this study, carefully addressing the frameworks in which they were built and their premises. Chapter 4 then draws the suggested borders of the adult stages and Chapter 5 inspects the developmental evidence that exists for each stage. Chapter 6 reviews in detail the different limitations on this type of research and Chapter 7 summarizes and discusses the implications and contributions that this study might have and suggests different lines of work for future research. This work is also accompanied by an appendix that includes all the calculations that were used for the results as well as all the charts. The code used to analyze the corpora is open-sourced and can be found in a GitHub repository under the address: `https://github.com/stavkl/CDS-thesis`

Chapter 2

Method

2.1 Research Questions

In chapter 1 it was established that CDS is similar to CS in several properties. First, CDS is different than regular adult speech to other adults (Phillips, 1970; Remick, 1971; Snow, 1972), and it goes without saying that CS is obviously different than adult speech as well. Second, like CS, CDS is also changing over time (Tal et al., 2021; Newport et al., 1977). Besides this similarity, it is well known that the input that children receive during the acquisition process is crucially related to proper language development (Bloom, 2002; Hollich et al., 2000; Nelson, 2009; Cristia et al., 2019) especially by means of generating and maintaining feedback loops between CDS and CS (Tomasello, 1992). The fact that CDS is similar and in other ways related to CS raises questions about the way CDS is studied and thought of, namely as a type of adult speech. Numerous studies acknowledge the effect of the input frequency on language learning (Ambridge et al., 2015; Dąbrowska et al., 2009; Tatsumi et al., 2018; Reali and Christiansen, 2007, to mention a few). However, despite the known similarities and relations between CDS and CS, all these studies treat CS (and generally the entire input that the child receive during their language acquisition process) as a single or unitary phenomenon, as an agent promoting or underlying for the language acquisition process and not as a developing system of its own, with unique and distinguished developmental stages like the ones found in CS research.

In this study I aim to investigate this understudied aspect of CDS. Specifically my research questions are:

- Do adults exhibit developmental stages in their CDS?
- How can such stages be determined?
- What is the relation between the adult stages and the developmental stages of children?

2.2 Research Hypotheses

First, I hypothesize that CDS can in fact be separated into meaningful and well defined developmental stages, at least when the children are very young (up to 3;6 years old). Second, that these developmental stages should be determined by a measure of grammatical complexity of the CDS, and not according to the developmental stages of the children, nor by choosing an arbitrary split according to the child's age. There's also no necessity for the adult developmental stages to be equal in size or length to one another, as they should reflect a meaningful change in the grammatical complexity of the adult. In this study the chosen measure of grammatical complexity is MPU (Dromi and Berman, 1982, Morphemes per Utterance), but other measures of complexity, like MLU (Brown, 1973, Mean length of Utterance), can be more appropriate when studying the developmental stages in languages with simpler morphology than in Hebrew (which is non-concatenative, highly ambiguous and context-sensitive). Last, I hypothesize that the adult stages correspond to the child stages, such that the adult is slightly ahead of the child. In other words, we expect to see the adult advancing to the next developmental stage, and then the child should move to their next stage, in turns, until the child reaches a certain level of maturity where they don't require adult mediation. The following section elaborates on how these hypothesis are validated in this study.

2.3 Validation

Ideally, in order to examine the relation between the developmental stages of the adult and those of the child in a case-study environment, we should obtain a dataset that contains both the child speech and the corresponding child-directed speech, both analyzed in a uniformed manner. The child-speech itself should be separable into developmental stages according to some schema, and those stages would later be compared against the suggested developmental stages of the corresponding adult. It goes without saying that in order to make better generalizations outside the case-study we should obtain multiple pairs of CS-CDS, however this kind of dataset does not exist even for one pair (see further elaboration under Chapter 3).

In this study, instead of a CS-CDS pair we have a schema of developmental stages in children that is based on a cross-sectional study on 40 Hebrew-acquiring children (Kaplan, 1983) that will serve as the CS part, and the transcriptions of the CDS by the caregivers of one child from a longitudinal study (Szubert et al., 2021), not related to the cross-sectional study as the CDS part. Keeping the CS part independent of the CDS part guarantees that the comparison between the suggested adult-stages and the children's stages remains unbiased. Nevertheless, findings from a cross-sectional study are not a trivial replacement for the corresponding CS which is missing. The support for the validity of that replacement is found in several studies (Brown, 1973; De Villiers and De Villiers, 1973; Moerk, 1980) that showed how the order of acquisition of morphological elements is invariant across children, and since this study focuses on the acquisition of morpho-syntactic elements as well (albeit in another language) we can safely assume that the missing CS data would have followed the same order of acquisition, and so we can extrapolate from the findings of the cross-sectional study to the findings we would have got had we had the corresponding CS.

The validation of the suggested adult stages will go as follows - first, I will introduce a previous work on the the developmental stages in children and the measure of grammatical complexity that was used to establish them (Kaplan, 1983). Then, using that same measure, the developmental stages for the CDS will be suggested, and compared against the child-stages. Afterwards, further justification for the adult stages will be provided in the following way - the CDS data will be analyzed for the morpho-syntactic elements that were investigated in the cross-sectional study by Kaplan. Per element, it will be shown how the adult usage of that element is distributed across the suggested developmental stages, and as was explained before, we expect to see a peak in the adult usage of an element, during or slightly before that element is acquired in children (again, according to Kaplan's schema). This study also provides a usage-based account for elements that were not acquired by children according to Kaplan's cross-sectional study, as well as to elements that were acquired very early and that their pattern of distribution in the CDS does not match the prediction described above.

2.4 Measure of Complexity - MPU

The chosen measure of grammatical complexity in this study is the MPU - Morphemes per Utterance (Dromi and Berman, 1982). This measure was suggested by Dromi and Berman as an alternative to Brown's measure which is the MLU - Mean Length of Utterance (Brown, 1973). While the MLU is very suitable for languages with a low rate of morphemes per word, like English, it gets dramatically skewed for languages with rich morphology like Hebrew, due to its concatenative morphology and the number of possible bound morphemes that can appear in a word. Consider, for example, a word like וכשהלכנו and its English counterpart - ‘and when we were walking’. The MLU for the Hebrew word, since it's a single word, is 1, while in English it is 5, but the Hebrew word conveys just as much information as the English phrase, and is much harder to process and acquire than other single word tokens like כלב */dog/*, which also has an MLU of 1. The MPU, on the other hand, will account for those differences, and will rate the complexity of וכשהלכנו as 5 and the complexity of כלב as 1, which faithfully reflects their differences in Hebrew.

Kaplan's cross-sectional study also used MPU to split the children's data into groups. It was found that the MPU is positively correlated with the children's age ($r = 0.86$), so the children are producing utterances that are more complex grammatically as they grow, which is an indicator of typical development. The age groups in Kaplan's study were designed to have an equal number of participants in each (as much as possible), in order to remove the biasing factor of group size on the results. Table 2.1 is a summary of the MPU and children's age in Kaplan's study . As can be seen from this table, age-group 5 is a control group that was added later to the cross-sectional study to verify the acquisition (or lack thereof) of other morpho-syntactic elements that were not acquired by the end of the original study (including children by the age of 3;0). This study will also accounts for the elements acquired relatively late or not acquired at all according to Kaplan, see 3 for more details.

Age Group	Youngest	Oldest	Mean MPU	No. of Participants
1	1;9	2;0	2.02	9
2	2;1	2;3	2.3	9
3	2;4	2;6	3.01	9
4	2;7	3;0	3.67	8
5	3;6		4.82	5

Table 2.1: Summary of mean MPU for age group in Kaplan's study

Chapter 3

Data

This study relies mostly on two previous works – a cross sectional study on 40 Hebrew acquiring children, carried out by Kaplan (1983), and a part of the Berman Longitudinal Corpus by Berman (Berman, 1990, 1996, 1997; Uziel-Karl, 2001; Armon-Lotem and Berman, 2003)

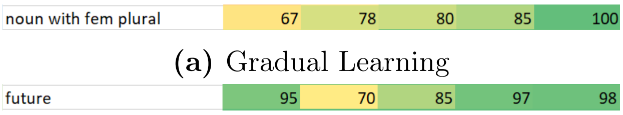
This chapter is organized as follows, first it extensively reviews the findings from Kaplan's study. Then, it elaborates on Berman's original corpus and how the re-analysis by Abend's group (Szubert et al., 2021) took place. This part also includes an introduction to the universal dependencies schema (Nivre et al., 2016), which was used in the re-analysis. The chapter ends with an overview on the current state of the public corpora available for language acquisition research (namely CHILDES (MacWhinney, 2000)) and discusses the limitations this platform imposes on this study.

3.1 Child Speech

Kaplan's study investigated the speech of 40 Hebrew-acquiring children from age 1;9 (21 months) to 3;6 (42 months). The study focused on a group of children from age 1;9 - 3;0, which consisted of 35 children, and later a group of another 5 children aged 3;6 was added as a control group (that is, it was assumed most if not all the morpho-syntactic elements would be acquired by that time). All the children are native speakers of Hebrew, monolingual and from the same social economic status.

The children were recorded in their homes or kindergarden, in two sessions that took place within a week. Overall, the children were recorded for 60-90 minutes. During the meeting the children were encouraged to speak as freely as possible through conversation, play and picture naming. The recordings were transcribed close to the time of recording, and there was a 90% agreement rate between transcribers.

Kaplan analyzed the transcriptions and calculated the mean MPU for each child, then, the utterances were examined for correct usage of morpho-syntactic elements such as subject-predicate agreement, noun-modifier agreement, inflections, bound morphemes. Kaplan kept a record of correct usages of the elements, and once the children reached at least 90% correct usage in context, from a certain point on, the element was considered acquired by children. So, for example, using future form to convey imperative meaning is an element that was acquired between ages 2;4 - 2;6 (28 - 30 months). Looking at the results for the correct usage of different elements in children's speech, it is clear that almost no element exhibits a U-shaped learning, but rather the learning is gradual, where for some elements mastery is achieved relatively early, and for others it is achieved late or not at all. A complete table of Kaplan's results of the stage of acquisition of elements can be found in Appendix A. An example for the differences between gradual and U-shaped development are demonstrated in Figure 3.1.



noun with fem plural	67	78	80	85	100

(a) Gradual Learning

future	95	70	85	97	98

(b) U-shaped Learning

Figure 3.1: Gradual and U-Shaped Learning

In Kaplan's study, only two elements (future verbs and plural feminine markers in adjectives) had a U-shaped learning curve, while the other 59 elements were mastered gradually. This study accounts for the gradual learning phenomena and demonstrates how the children's learning is tightly related to the element distribution in the adult input around the time of acquisition (or lack thereof).

3.2 Child-directed Speech

This section discusses CDS part of this study. It is split between the original longitudinal study (Berman, 1990, 1996, 1997; Uziel-Karl, 2001; Armon-Lotem and Berman, 2003) from the 90's and the latest re-analysis of the data (Szubert et al., 2021) according to the universal dependencies schema (Nivre et al., 2016). This section concludes with an overview of the current state of collaborative corpora of first language acquisition in Hebrew and the extent to which it can be used in this study and other studies alike.

3.2.1 Berman's Original Project

This Hebrew longitudinal data-base consists of naturalistic longitudinal data collected on a weekly basis from four Hebrew-speaking children, three girls (Hagar, Smadar, and Lior) and one boy (Leor). All four children are native speakers of Hebrew raised in monolingual, highly educated Hebrew-speaking homes in urban communities of central Israel. Smadar was the youngest of three girls, Hagar and Leor were only children at the time of recording, and Lior had a baby brother.

Each child was audio-recorded at his or her home for a total of around one hour per week, typically two or three times a week in different situations (mealtime, bath time, playing on their own or with siblings or parents and grandparents). Recordings were done over a period of one to three years (see Table 3.1 below). The contact person and main recorder for three of the children was the mother, and in one case (Leor's) the aunt – all four native speakers of Hebrew that had majored in linguistics at the university. Those doing the recording were also instructed to specify the exact situation in which recording took place at the outset and in the course of each session. Information about the situation in specific sessions is provided in each file's metadata.

This data-base has several features that make it well-suited to child language research. The interactions are natural since they were recorded in the homes, a setting familiar to the children, in the presence of a primary caregiver and / or other members of the family. The data were collected over several sessions each week and so allowed a variety of contexts for the children to express themselves. Rich contextual information was provided by the caregivers, and the latter were regularly available to the transcriber for consulting and clarifications. Finally, both the transcribers and the researchers involved in the project knew the children and their parents, and were familiar with the children's linguistic development beyond the data provided by the recorded sessions.

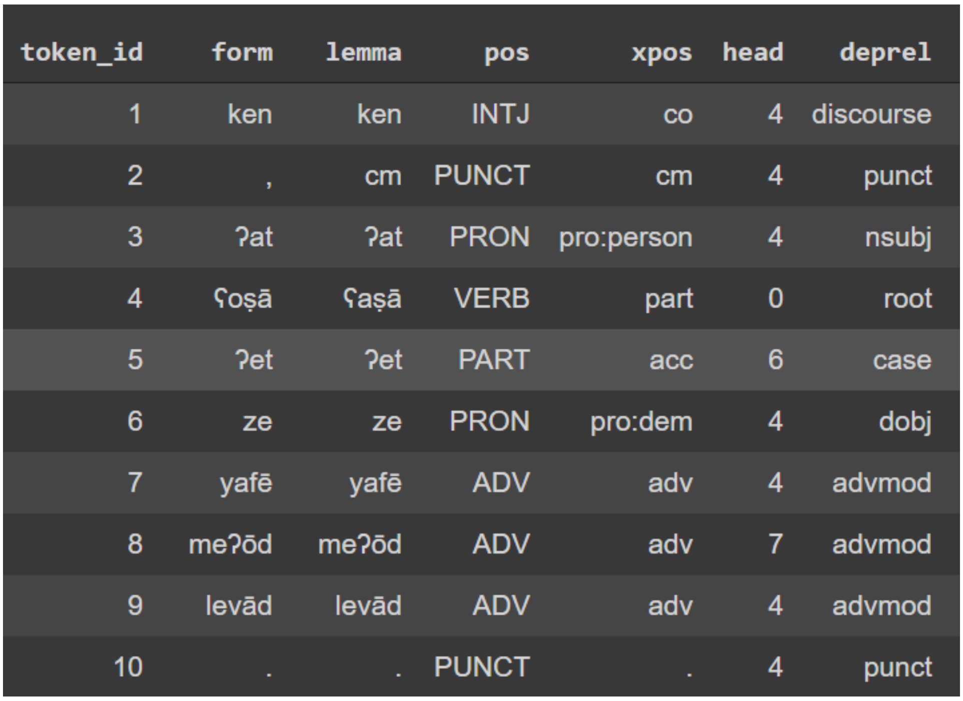
Child Name	Gender	Age Range	No. Child Utterances	No. CDS Utterances
Hagar	F	1;7 - 3;3	16,636	24,398
Leor	M	1;9 - 3;0	16,434	18,360
Lior	F	1;5 - 3;1	6,689	8,685
Smadar	F	1;4 - 2;4	3,753	3,427

Table 3.1: Summary of the data in the 'Berman Longitudinal' corpus in CHILDES

Table 3.1 gives details of the data-base that was uploaded to the CHILDES platform. Notice that the numbers here and the numbers available on the corpus page on the CHILDES website differ in the number of utterances of children and adults. This mismatch between the reported numbers on the webpage and the numbers in Table 3.1 is due to the fact that not all the recordings and transcriptions were eventually uploaded to the platform, but rather remained private. Private and semi-private corpora are one drawback of the joint Hebrew corpora in CHILDES that make it unsuitable for collaborative research on Hebrew acquisition. An elaborated discussion on the current state of the Hebrew CHILDES can be found under Section 3.3.

3.2.2 Hagar's CDS Re-analyzed

In 2019 Omri Abend and his research group from the Hebrew University of Jerusalem re-analyzed the CDS part of Hagar's sub-corpus from 'Berman Longitudinal'. Although their research was not yet published, it is clear from Table 3.1 that choosing Hagar's CDS was not coincidental, but rather intentional as she exhibits the richest available CDS in Hebrew. I am again thankful for their courtesy to share their re-analysis with me. The CDS data is transcribed with respect to the Hebrew orthography, which is an advantage compared to a phonetic or even a broad phonemic transcription, since in modern Hebrew there are several homophones, some of which participate in morpho-phonemic processes. Therefore, a broad phonemic transcription would introduce a great amount of ambiguity to the data and is also unnecessary for the purposes of this study that only accounts for morpho-syntactic phenomena. Below is an example sentence of the CDS corpus of Hagar. This sentence exemplifies how the transcription is orthographically-faithful, as well as the facts that the prosodic stress is marked for words with more than one syllable and that other prosodic markers that are expressed through punctuation are also specifically addressed. Overall, Hagar's corpus contains 24,171 utterances.



token_id	form	lemma	pos	xpos	head	deprel
1	ken	ken	INTJ	co	4	discourse
2	,	cm	PUNCT	cm	4	punct
3	ʔat	ʔat	PRON	pro:person	4	nsubj
4	ʕoṣā	ʕaṣā	VERB	part	0	root
5	ʔet	ʔet	PART	acc	6	case
6	ze	ze	PRON	pro:dem	4	dobj
7	yafē	yafē	ADV	adv	4	advmod
8	meʔōd	meʔōd	ADV	adv	7	advmod
9	levād	levād	ADV	adv	4	advmod
10	.	.	PUNCT	.	4	punct

"Yes, you are doing it very well on your own"

Figure 3.2: An example sentence analysis

Introduction to Universal Dependencies

This subsection gives a short overview of the Universal Dependencies schema (Nivre et al., 2016). The Universal Dependencies was established to make a uniform analysis (as much as possible) of syntactic relations between pairs of words, with an objective not to rely on empty elements, movements and intermediate representations as much as possible. This objective makes it possible for languages with very different syntax to share analytic tools and promotes the creation of multi-lingual tools, especially in the field of Natural Language Processing (NLP). The schema defines certain columns that the analysis table must contain. First, `token_id` is the serial number of each word or token in the utterance. The schema enables us to analyze the utterance word for word or to split certain words into their composing morphemes, such as the French article *du*, for example, which contracts the articles *de* and *le*, each of which is considered a token, even though they are not realized as separate units on the surface. This design enables researchers to keep a consistent analysis for all the *de* + definite article in French and for contracted forms in general and was therefore widely accepted.

The next columns are the `form` and the `lemma`. The `form` is the word or token as they appear on the surface (as much as possible) and the `lemma` is best described as 'what can be found in a dictionary' which is usually the 3rd person, singular, masculine, past tense inflection. The `pos` column specifies the part-of-speech tag for each token, out of a limited set of POS-tags, and the `xpos` column enables to make language-specific distinctions, such as the `pro:dem` label, that specifies that the word *ze* (`token_id = 6`) is a demonstrative, and not just a simple `PRON` (Pronoun).

The `head` and `deprel` columns specify to which each word is connected and in which type of relation. The types of relations between words also form a closed and relatively small list. A graphic demonstration of the relations between words and their heads can be found in Figure 3.3 below.

In figure 3.3 (that was generated by Qi et al. (2020)) we can see the POS-tags, such as `JJ` (adjective), `DT` (determiner), etc. above each words, and the labeled and directed edges, specifying that `fox` is the `nsubj` (nominal subject) of `jumped`, and so on. The word `jumped` is unique in that it is defined

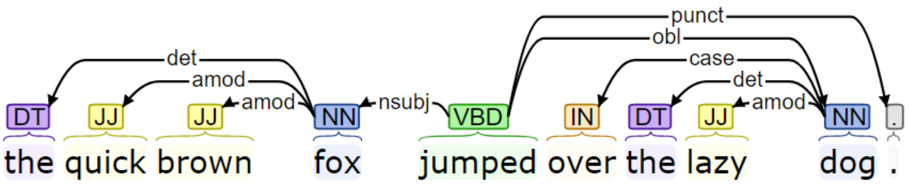


Figure 3.3: Dependency relations between words in a sentence

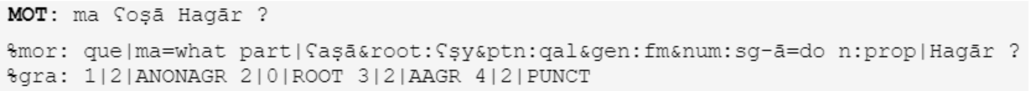
as the `root` of the sentence. There must be only one root in a sentence, even if it contains an embedded clause or a conjunction of sentences. The root itself is not dependent on any other token, therefore its `head` is always 0.

It should be stated that the UD convention also require an analysis for morphological features, like person, number, gender, etc, which this analysis lacks although this information can be extracted from the transcriptions and analyses available in CHILDES.

3.3 The Data in CHILDES

It was explained in Chapter 2 that in an ideal situation we would want to have a pair of the CS and its corresponding CDS, both analyzed in the same manner. On the one hand this type of data was said to be non existent, and on the other hand we know the CHILDES platform hosts the transcriptions and recordings of both CS and CDS, from the Berman Longitudinal corpus as well as others. This raises the questions how can we say the ideal pair does not exist, and why can't we support our claims with more data from what already exists in CHILDES? The answer is two-fold.

In 2013, all the data in the Hebrew section of CHILDES underwent re-analysis as part of Gretz et al. efforts to build tools for Hebrew transcription and morphological analysis (Albert et al., 2013; Gretz et al., 2015). Their re-analysis was later accepted to the CHILDES project and replaced the previous analysis. Here is an example of a sentence from Hagar's corpus after their re-analysis: As can be seen from the image, the syntactic relations in the %gra tier are `ANONAGR` (argument without agreement) and `AAGR`



```
MOT: ma ʕoṣā Hagār ?
%mor: que|ma=what part|ʕaṣā&root:ʕṣy&ptn:qal&gen:fm&num:sg-ā=do n:prop|Hagār ?
%gra: 1|2|ANONAGR 2|0|ROOT 3|2|AAGR 4|2|PUNCT
```

"What does Hagar do?"

Figure 3.4: Example sentence from the Hebrew CHILDES corpora

(argument with agreement) which only represent valency of predicates and whether their arguments agree with their predicates in gender and number. There is also no notion of a subject relation or any other relation from the UD schema, even though the data is analyzed as dependency relations. Unfortunately, reducing the syntactic relations described in the UD scheme to mere agreement between a predicate and its arguments (which is almost completely unattested in Hebrew except for the predicate's subject) ignores important syntactic roles and relations that are needed in order to automatically extract several morpho-syntactic elements. It is important to state that the UD schema allows to add additional information in a specified, so that the two analyses (the standard UD and Gretz et al. revised edition) can potentially live side by side and there is no need to exclude one or another.

Another drawback of the data in CHILDES was mentioned earlier in this chapter and that is the mismatch between the metadata that is reported (namely, the number of files and as a result the number of utterances for each child) and the actual data that was uploaded and is browsable and downloadable. This mismatch creates a false image of the available data and can affect the decision making process when deciding to use that data for new projects.

Chapter 4

Defining Adult Stages

The purpose of this chapter is to introduce the suggested adult stages of Hagar's CDS and to elaborate on the motives and research procedure that led to those stages. This procedure closely follows Kaplan's methodological choices in order to make Kaplan's data on CS and this study's data on CDS more compatible with each other. The findings suggest that the adult stages proposed here can in fact be determined according to the adult MPU clustering described below, and that the transition of the adult from one developmental stage to the next is interwoven in the transition of the children from one stage to the next, as hypothesized.

4.1 MPU Clustering

Splitting the data according to a measure of grammatical complexity is a methodological choice that is grounded in both aspects of this study – CS research and CDS research. In CS research and specifically in Hebrew Kaplan found a strong correlation between the children's MPU and their age ($r = 0.86$), and this correlation also corresponds to De Villiers and De Villiers (1973) study that found that MLU in English speaking children is a far better predictor of the age of acquisition than chronological age. From the CDS research perspective, it was attested by Newport et al. (1977) that the *parental* propositional complexity increases with the children's MLU. It was therefore decided to follow the same logic in this study and to split the

CDS data of Hagar according to the adult MPU. To make the CS and the CDS most comparable, and since Kaplan's children age range significantly overlaps with Hagar's age, it was also decided to follow Kaplan's choice of having 5 age groups and split Hagar's CDS into 5 developmental stages. The clustering itself is the output of applying the K-Means algorithm (Lloyd, 1982; Forgy, 1965; MacQueen, 1967) on Hagar's CDS with $k = 5$, for the reasons mentions above. The K-Means algorithm assigns each data-point to the cluster with the nearest mean, with the objective to keep the in-cluster variance to a minimum. Note that it is not required that the mean is a data-point of its own, it can be (and usually is) an abstract point in the area of the suggested cluster, and during the iterative process of the algorithm it may change its place several times, until the algorithm converges.

4.2 Pre-processing

Before calculating the adult MPU, some pre-processing needs to take place; First, unnecessary lines are removed, that is - lines of un-interpretable speech (where the `form` columns has the value `xxx`) and punctuation lines (where the `pos` column has the value `PUNCT`). Then, the length of every utterance is calculated to look for anomalies. In the case of Hagar's CDS there were only two utterances that were exceptionally long – one with 63 morphemes and the other with 48 morphemes. After removing the two longest utterances the next 5-longest utterances had between 32 and 37 morphemes each.

4.3 Results

Figure 4.1 shows the mean MPU of the CDS utterances according to Hagar's age in months. So, for example, when Hagar is 2 years old (24 months) the mean MPU of her CDS is 4.747, and so on. Figure 4.1 demonstrates how Hagar's CDS becomes more complex grammatically and more varied over time. We want to cluster these data points and define each cluster as a developmental stage for the adult - this procedure is the CDS adaptation of Kaplan's procedure to split the CS data into stages, only Kaplan's motives

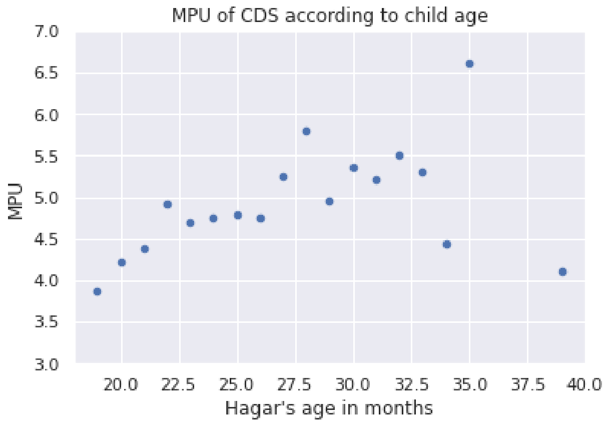
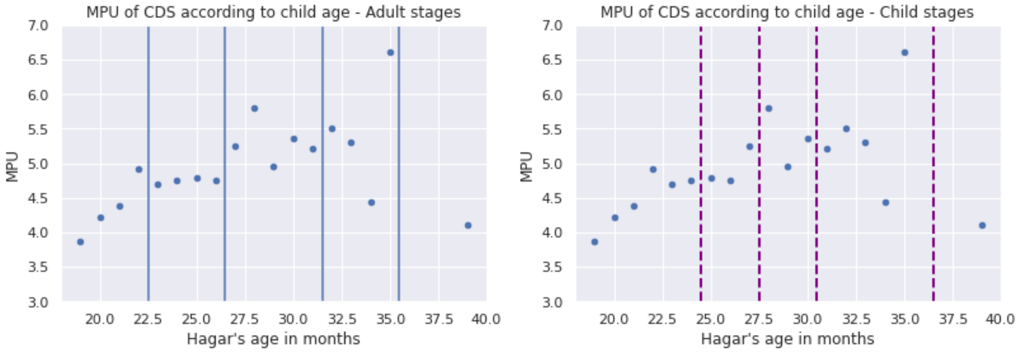


Figure 4.1: MPU of Hagar's CDS according to Hagar's age

were to split the children into evenly numbered groups (see Table 2.1) and the distinct mean MPU between the age groups was a byproduct of that split, whereas in this study the mean MPU of the CDS determines the split via a clustering algorithm called K-Means clustering (Lloyd, 1982; Forgy, 1965; MacQueen, 1967). This algorithm partitions n observations into k clusters. In this case I chose $k = 5$ clusters to fit the number of age-groups in Kaplan's study. The algorithm assigns each data point to the cluster with the nearest mean, with the objective to keep the in-cluster variance to a minimum.



(a) Adult Stages by K-Means Clustering **(b)** Child Stages by Kaplan

Figure 4.2: Child stages and adult stages projected on CDS MPU data

Figure 4.2 shows the results of two possible clustering methods. Figure 4.2a shows the results of the K-Means clustering described above. This clustering matches our intuition about the 'correct' split and is also mathematically sound. The clustering on figure 4.2b, on the other hand, represents the clustering we would have got had we used the developmental stages determined by Kaplan. In other words, it represents a scenario where we study CDS according to the developmental stages of children. The clustering derived from that decision is irrelevant to CDS research and strengthen the need for a designated analysis of CDS. Figure 4.2a strengthens this study hypotheses that CDS does in fact exhibit clear and distinguishable stages, and that those stages can be determined by the adult MPU. Nevertheless, the adult-stages and the child-stages are related. If we put the two charts on top of each other, we get figure 4.3 below:

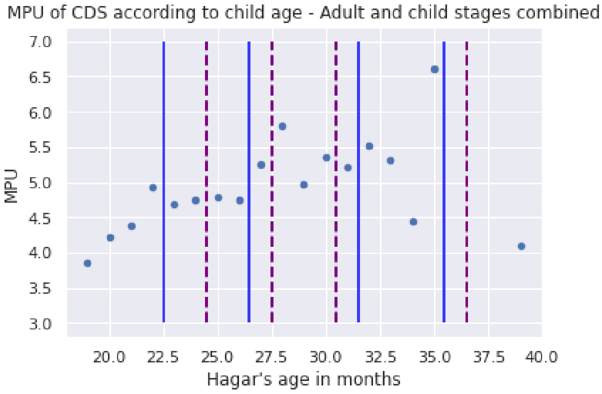


Figure 4.3: Adult stages and child stages combined

In figure 4.3, as in 4.2 the solid blue vertical lines indicate the adult stages and the dashed purple lines indicate the developmental stages determined by Kaplan. The adult stages seem to alternate almost perfectly with the children stages, thus strengthening the hypothesis that the adult is slightly ahead of the child in terms of the development of their CDS.

Chapter 5

Developmental Evidence for Stages

This chapter provides further evidence for the developmental stages in CDS. It does so by examining the distribution of usage of morpho-syntactic elements in the CDS and compare that distribution with the age of acquisition of such elements in children according to Kaplan's findings. We expect to see that the adult's usage of an element peaks slightly before or during the time this element is considered acquired. Acquisition in that sense is defined by Kaplan as achieving at least 90% accuracy in correct usage in contex.

The procedure of examining the distribution of an element in the CDS starts with extracting the relevant utterances (i.e. utterances that contain an occurrence of the element) - this step is not always straightforward as the data is annotated according to the UD standards and not every distinction that is meaningful for CDS research in Hebrew is part of that schema. To deal with that, some of the elements were extracted using heuristics or by elimination. Still, several elements could not be examined at all - see chpater 6 for an elaborated discussion.

5.1 Pre-processing

Before the distributional analysis can take place it's important to review the meta-data of the various stages determined in chapter 4. Table 5.1

summarizes the number of utterances in each of the suggested adult stages.

	Stage 1	Stage 2	Stage 3	Stage 4	Stage 5
Number of utterances	6568	7813	6362	2967	462

Table 5.1: Number of utterances per CDS stage

Stages 1-3 contain roughly the same number of utterances, stage 4 contains significantly less utterances but the difference is not an order of magnitude. Stage 5, on the other hand, is about an order of magnitude smaller than the others. This is due to the fact that this stage contains a single session that took place 3 months after the main part of the study. For stages 1-4, each data point in the chart (figure 4.1) represents the average MPU over at least 10 different sessions.

For this reason it was decided to exclude stage 5 from any further analysis. In that way we would not make an unreasonable and unfair expectation to see the distribution of an element peaking before (i.e. in stage 4 which is also significantly smaller) or during stage 5. Also, recall from chapter 2 that Kaplan's study involved children aging 3;6 (42 months, which falls under stage 5 in adult stages) as a control group. By this time, many elements that were not acquired before were already acquired in the control group, so there is also a question of what exactly is the age of acquisition for these elements. There are also several elements that were not acquired even by 3;6, which raises more questions regarding the nature of CDS in later stages of language acquisition (age 3;1 and above). These questions are beyond the scope of this study and are discussed in the future research section. Therefore, the primary focus henceforth would be on stages 1-4.

5.2 Results

Out of 59 elements that Kaplan examined in CS, only 30 can be automatically or heuristically extracted from the CDS data (see chapter 6 for a detailed discussion). This section starts with an overview of the different adult stages in terms of their grammatical complexity and the overall distribution of morpho-syntactic elements across different stages. Let us define two measures for evaluating the elements' size and the density of each stage.

We denoe μ as the average number of occurrences across all the elements in a stage, since there are 30 elements overall and all the elements have fully specified values for each stage, the average is the the sum of occurrences for all the elements, divided by 30. In other words, this average represents the size of the average element in a stage.

$$\mu_j = \frac{1}{30}\sum_{i=1}^{30} occurrences\ of\ i\text{-}th\ element$$

So, for example in stage 1 the sum of all the occurrences of the elements in 7229, therefore the average element in stage 1, denoted by $\mu_1 = \frac{7229}{30} = 240.97$. It goes without saying that a single utterance can contain more than one element, but we are interested in the size of the average element and what we can learn from this size with respect to the the total size of the stage. Now, we can define *Density* as the percentage of the average element size in the total size of the stage.

$$Density_j = \frac{\mu_j}{number\ of\ utterances\ in\ stage\ j} \cdot 100$$

The *Density* measure reflects the amount of grammatical complexity in a stage. Since this value is normalized by the size (i.e. number of utterances) of the stage, we can safely compare the grammatical complexity of the stages without the bias caused by their different sizes.

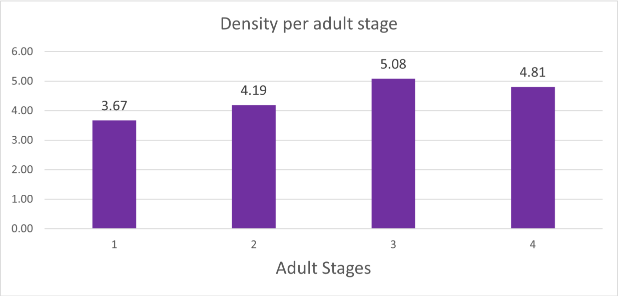
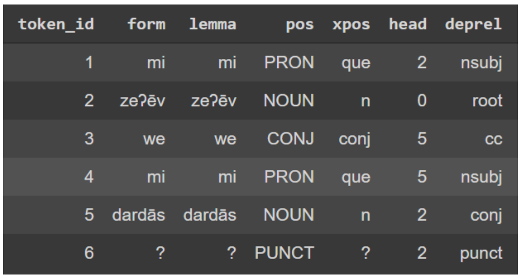


Figure 5.1: Density of adult stages

Figure 5.1 shows the growth in the density of the different stages. This growth is a result of the growth in the average element size, combined with the number of utterances in each stage. So, even though the number of utterances in stage 3, for example, is lower than in stages 1 and 2, there are more morpho-syntactic elements per utterance, and therefore the overall density of the stage is higher. An exhaustive list of all the morpho-syntactic elements that were inquired and their number of occurrences per stage is brought in Appendix A.

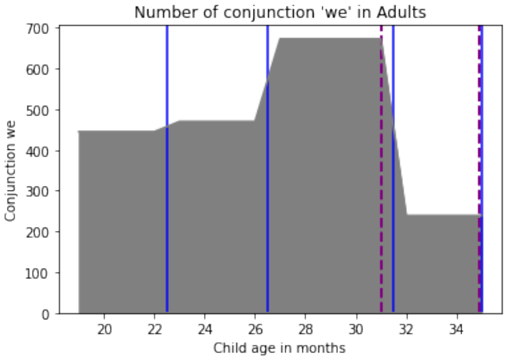
Let us now move to a demonstration of the correspondence between the distribution of elements in the CDS and their stage of acquisition by children according to Kaplan. Recall that we predict that morpho-syntactic elements would be acquired at the point where the adult uses them the most in their CDS. Elements that exhibit this pattern, that is - a peak in usage, during which (or followed by) the element is acquired are considered matching the prediction (henceforth will be referred to as 'matching distribution elements').

Examples for such elements are shown below alongside with exemplary usage



token_id	form	lemma	pos	xpos	head	deprel
1	mi	mi	PRON	que	2	nsubj
2	zeʔēv	zeʔēv	NOUN	n	0	root
3	we	we	CONJ	conj	5	cc
4	mi	mi	PRON	que	5	nsubj
5	dardās	dardās	NOUN	n	2	conj
6	?	?	PUNCT	?	2	punct

(a) *who's a wolf and who's a smurf?*



(b) Distribution of regular conjunction

Figure 5.2: Regular conjunction, example and distribution

Figure A.13 shows the distribution and acquisition of the conjunction element 'we' (ו, *'and'*), as well as a real example of usage taken from the data . This elements includes conjunction of names, NPs, predicates (that have the same subject) and utterances. It does not include 'demi'-conjunctions that appear at the beginning of the utterance and do not connect two elements.

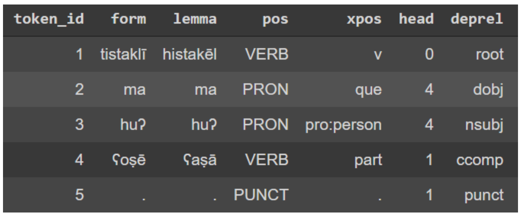
These conjunctions, termed here 'initial-we' are considered a separate element and are also acquired at a different stage according to Kaplan. The regular conjunction brought here is acquired by children between 2;7 and 3;0 (31-36 months), which is indicated in the figure by the dashed purple lines.

The distribution of the element across the different stages is represented by the grey area, and the adult stages are the solid blue lines. It is clear from the figure that the adult usage of regular conjunctions is increased during stage 3, and the acquisition of the element according to Kaplan's findings immediately follows that peak, as expected.

The regular conjunction is easily extracted automatically from the data, as can be seen in the exemplary utterance in A.13a, the conjunction morpheme is separated in its own line and its POS-tag is `CONJ` - this is a unique POS tag that can only be assigned to conjunction elements. Therefore, to automatically extract all the conjunction we simply need to extract all the lines that have a `form = we, pos = CONJ` where the `token_id` does not equal to 1 (to separate them from the 'initial-we' that is not used for conjunction).

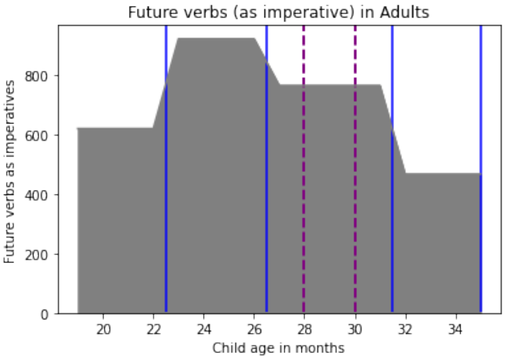
The following example belongs to a morpho-syntactic element that cannot be extracted automatically, and demonstrates how such elements can be extracted heuristically - using verbs in future form to convey imperative meaning (as described in Kalev (2017)).

Looking at the UD-analysis of the verb `tistaklī`, we see it is only marked as a verb, with `xpos = v`, which is used for all the verb forms except present tense ones, therefore heuristics must be used. First, a list of all the verb in the corpus was extracted. Notice that this corpus is orthographically faithful, so we can easily extract all the verbs that start with one of the four letters that can indicate future form (**איתנ**). Then, a manual inspection of those verbs (about 1000 verb types) takes place to keep only those verbs that start with (**איתנ**) and are in future tense. Finally, the list of future verbs is filtered to include only the verbs that start with /t/, assuming that the imperative meaning is addressed to 2nd person. A manual inspection of 100 verb tokens (randomly sampled) was conducted post-processing to make sure the usage



token_id	form	lemma	pos	xpos	head	deprel
1	tistaklī	histakēl	VERB	v	0	root
2	ma	ma	PRON	que	4	dobj
3	huʔ	huʔ	PRON	pro:person	4	nsubj
4	ʕoṣē	ʕaṣā	VERB	part	1	ccomp
5	.	.	PUNCT	.	1	punct

(a) *look what he's doing*



(b) Distribution of future forms as imperative

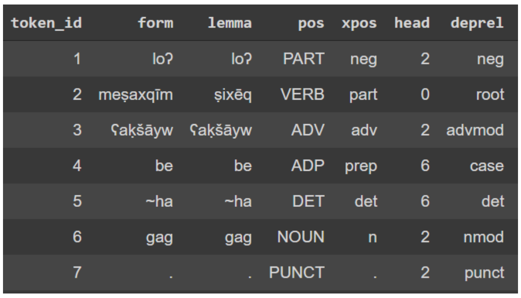
Figure 5.3: Future forms to convey imperative meaning, example and distribution

of those verbs was to convey imperative meaning, with roughly 2% error.

Figure A.5b shows how the adult usage of this element peaks during adult-stage 2 while the acquisition of this element by Kaplan happens during adult-stage 3, as initially hypothesized. The following sections elaborate on cases where the distribution combined with the stage of acquisition does not match the prediction, i.e. where the acquisition happens before the peak (the 'too many' case) and where the acquisition does not happen at all, regardless of the peak (the 'too few' case). The quantitative difference between elements that were eventually acquired and elements that were not acquired at all by the end of Kaplan's study is also discussed.

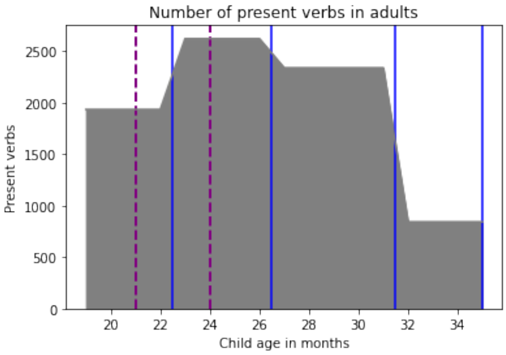
5.2.1 Upper Bound

Some elements, namely the present and past verb forms, are acquired before the adult usage peaks. They are considered acquired at the very first stage in children, 1;9 – 2;0 (21-24 months). Since Kaplan's study only begins at 21 months it is plausible that the children reach 90% accuracy in these elements even before 21 months, yet it is clear from figure A.2b that the adult usage peaks during stage 2 (and we know it is not just a matter of size of the stage because stage 3 contains less utterances than stage 1, yet in present tense verbs stage 3 surpasses stage 1).



token_id	form	lemma	pos	xpos	head	deprel
1	loʔ	loʔ	PART	neg	2	neg
2	meṣaxqīm	ṣixēq	VERB	part	0	root
3	ʕaḳšāyw	ʕaḳšāyw	ADV	adv	2	advmod
4	be	be	ADP	prep	6	case
5	~ha	~ha	DET	det	6	det
6	gag	gag	NOUN	n	2	nmod
7	.	.	PUNCT	.	2	punct

(a) *(we're) not playing on the roof right now*



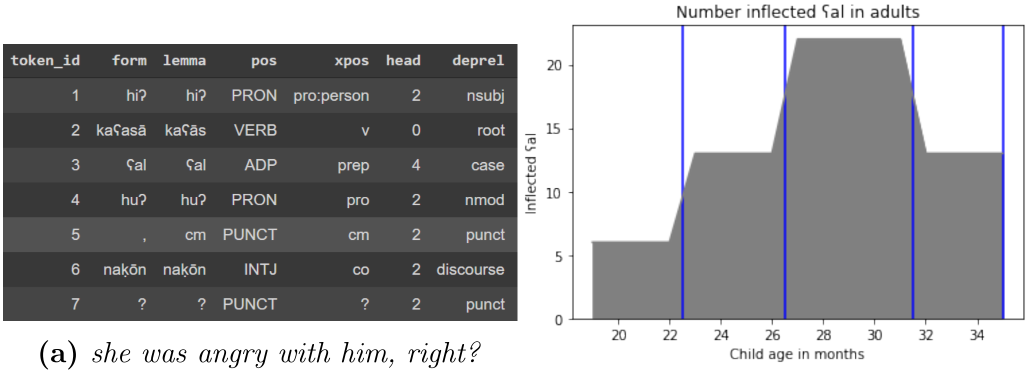
(b) Distribution of present tense verbs

Figure 5.4: Present tense verbs, example and distribution

What could be the reason for this early acquisition? Taking a closer look on the adult usage of present tense verbs we see a very extensive usage of this element from the very beginning. In stage 1 alone there are almost 2000 occurrences of present tense verbs, compared to several hundreds occurrences in stage 1 for matching-distribution elements. We must therefore conclude that the acquisition of an element relies not only on the distribution of the element's occurrences in the adult speech, but also on some sort of upper-bound above which the element will be acquired regardless of the overall distribution. In other words, getting "too many" examples of an element can facilitate acquisition, and it is beyond the scope of this study to discuss why and how the mental state of children and their ability to grasp the passage of time at 21 months make present and past verbs ideal candidates to surpass that upper bound.

5.2.2 Lower Bound

This section accounts for elements that were not acquired at all by age 3;6, but according to their usage pattern should have been acquired. Figure A.29 shows the usage of the preposition ʕal (*on, about*) when it is inflected (that is, followed by a pronoun. Since this corpus lists each morpheme in a separate line we need to use our knowledge on Hebrew to conclude that every ʕal that is followed by a pronoun is actually a single word and not two). Kaplan asserts that this element was not acquired by age 3;6 (42 months), even though its usage distribution suggests it should have been acquired by 32 months.



token_id	form	lemma	pos	xpos	head	deprel
1	hiʔ	hiʔ	PRON	pro:person	2	nsubj
2	kaʕasā	kaʕās	VERB	v	0	root
3	ʕal	ʕal	ADP	prep	4	case
4	huʔ	huʔ	PRON	pro	2	nmod
5	,	cm	PUNCT	cm	2	punct
6	naḳōn	naḳōn	INTJ	co	2	discourse
7	?	?	PUNCT	?	2	punct

(a) *she was angry with him, right?*

(b) Distribution of inflected ʕal

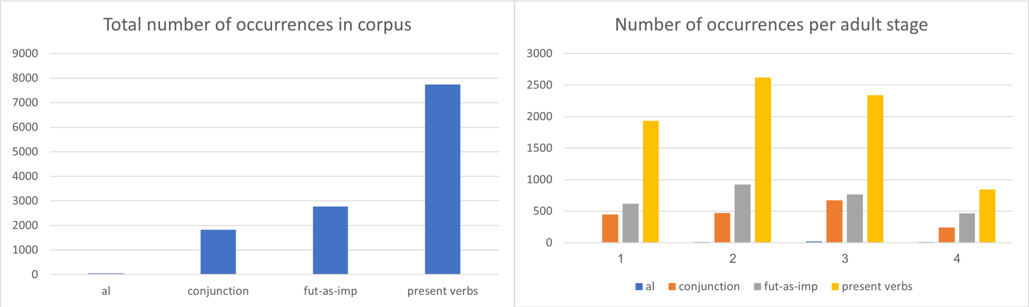
Figure 5.5: Inflected ʕal, example and distribution

Taking a closer look, we can see that there are only 54 occurrences of ʕal in the entire corpus, compared to hundreds of occurrences *per stage* for elements that were acquired. Therefore, we also have to take into account a lower-bound for the minimum number of occurrences required for an element to be acquired.

5.3 Summary

The findings of this study show that for most elements the distribution of the adult usage can serve as a good indicator for the stage of acquisition in children. In other words, we can expect an element to be acquired during or after the peak in the adult usage.

There are, however, several elements that do not follow this distributional pattern, i.e. are acquired before the peak or not acquired at all even when the distribution suggests a valid and reasonable time for acquisition. A close look at the absolute number of occurrences in each stage reveals that elements that were acquired before the peak had about an order of magnitude more occurrences than matching-distribution elements, and on the other hand elements that were not acquired at all had about an order of magnitude less occurrences than matching distribution elements. The proposed explanation for these exceptional elements is the existence of upper and lower bounds for acquisition, where elements that surpass the upper bound very early are expected to be acquired very early as well regardless of their distribution in adult speech, and elements that do not reach the lower bound will not be acquired even if their distribution can fit a certain age of acquisition.



(a) Total number of occurrences in corpus

(b) Number of occurrences per adult stage

Figure 5.6: Distribution of occurrences of elements

Figure 5.6 above reviews the total number of occurrences of the exemplary elements presented in this chapter. Figure 5.6a shows the total number of occurrences in the entire corpus - the middle columns represent matching-distribution elements (conjunction-we and future-indicating-imperative), the leftmost column represents an element below the lower bound (inflected ʕal) and the rightmost column represents an element above the upper bound (present tense verbs). Across the entire corpus, the difference in the proportions of the matching-distribution elements and the extreme elements is significant. Figure 5.6b shows a finer-grained distribution of the number of occurrences of elements per stage. This figure demonstrates the differences in the number of occurrences especially in the first stage, where the upper bound limit is (probably) already surpassed.

Another point of discussion is the difference between elements that were eventually acquired by 3;6 (42 months) and elements that were not acquired even by that time. Since stage 5 was initially excluded from the analysis, these elements were not part of the main results, yet it is interesting to investigate the difference between these groups.

	Late Acquisition	No Acquisition
Total number of elements	10	9
Number of extractable elements	9	3
Smallest element	70	54
Largest element	496	80
Avg. size of element	284	64.67
Median element size	316	60

Table 5.2: Differences between late-acquisition and no-acquisition

Table 5.2 shows how elements that were acquired (except one) are significantly larger than elements that were not, albeit most element from the no-acquisition group cannot be extracted automatically or heuristically. These results further support this study's main claim that a quantitative difference in the adult speech along with a lower bound may determine which elements will be acquired and which elements will not. It goes without saying that these elements will eventually be acquired in a later stage that is beyond the scope of this study.

Chapter 6

Limitations of the Study

This study was conducted using relatively limited resources, with non-trivial properties. Therefore, in order to generalize from its results several aspects should be taken into consideration. This chapter reviews these aspects and suggest ways that future research can use to overcome these limitations in similar studies.

6.1 Case-study

It should be stressed that the CDS part of this study is based on a single corpus of CDS. In general, a single corpus can be biased towards certain phrases or structures that are idiosyncratic to the participating adults. In a cross sectional studies such differences would have been canceled out by other idiosyncrasies and the corpora would be less biased overall. Another non trivial property of this corpus is that it was originally recorded in the 80's and re-analyzed in 2019, and therefore the usage of morpho-syntactic elements may be different than what we expect from modern Hebrew - for example, the usage of the word **עם** (*'with'*). It might be the case that in the 80's the inflected version of **עם** was **עמו** (*'with him'*), which is phonetically similar to the non-inflected form, so the relation between the two forms is more transparent, whereas nowadays **עם** exists only in isolation (not inflected) and the inflected form is **אתו**, and that relation is much more opaque. There is no way of knowing what was the pronunciation Hagar actually heard, as the corpus is split by the morpheme, so either way there are two separate lines, one with עם and the following with הוא. Also on that topic, the usage of certain morpho-syntactic elements in the 80's might have been due to their overall frequency in adult-to-adult speech, but such analysis on the token frequency in spoken Hebrew in the 80's does not exist. Further research should take into account different fashions and time differences and try to include corpora that is as homogeneous as possible.

The last non-trivial property of Hagar's CDS corpus is that is addressed to Hagar, who is a girl. This single property may skew the distribution of every morpho-syntactic element that is related to grammatical gender in Hebrew. Ashkenazi et al. (2015) found that girl's parents use more feminine verbs and boy's parents use more masculine verbs. The feminine form, however, is considered marked and therefore should be acquired after the unmarked form (this is also supported in Kaplan's findings). The differences between CDS that is addressed to a girl versus CDS that is addressed to a boy deserve a research of their own (beyond the results of Ashkenazi et al., 2015), and so are the assumptions on the order of acquisition in boys and in girls. This study's findings support the idea that there might be different orders of acquisition, including the acquisition of marked and unmarked forms, that are due to the distributions of these forms in the CDS (as we assume CDS-to-girl would be much richer in marked forms than CDS-to-boy).

6.2 Automated Analysis

This study uses a corpus of about 24K utterances. This amount of utterances cannot be searched manually, and therefore the extraction of morpho-syntactic elements relies only on automated and heuristic-based methods. However, out of 59 elements in Kaplan's research, only 30 can be extracted automatically or by heuristics. This is due to several factors. First, the entire category of agreement related elements is undetectable in adults, and this is because adults produce utterances with the correct agreement most of the time (there are mistakes, of course, but they are not systematic uses that can count towards a distribution of their own). Second, many elements cannot be extracted automatically because they are defined through context. For example the element "using a pronoun as a subject when describing a picture", which is an extreme, but also more standard phenomena like the different meanings of dative - location, beneficiary, etc. These elements can only be understood by a manual examination of the entire context in which the utterance takes place. Such context does not exist in the data (only in the metadata on the CHILDES platform), and cannot be automatically understood since the possible contexts are varied, not coded in a systematic way, and even if such coding existed it is not clear how each context-related element would benefit from it. Future research can build a coding system for the different contexts with the morpho-syntactic elements in mind, and that might enable to investigate the acquisition of context-related elements as well.

6.3 Built-in Corpus

The last type of limitation comes from the fact that this corpus is transcribed in Latin script (and special IPA characters (Association et al., 1999)) and analyzed (by Abend's group (Szubert et al., 2021)) based on the UD-scheme, each imposes its own limitations. Starting with the latter, the UD-based scheme used for this corpus is missing several components that the general UD scheme defined, the most important of which is the morphological features component. This component defines the morphological features for each token, based on the morphological distinctions that each language makes. In Hebrew, the UD scheme defines that tokens should be specified for gender, number, definiteness, person, Binyan (morphological template), tense and voice (each of them is specified when relevant). The lack of morphological annotation makes it very difficult to find morpho-syntactic elements that are related to those morphological features, for example the element 'plural feminine adjectives', 'plural demonstratives in agreement with a noun' and others.

A possible workaround would be to analyze the utterances through a modern Hebrew analysis tool such as YAP (More et al., 2019), AlephBERT

(Seker et al., 2021) or UDPipe (Straka and Straková, 2017), that outputs the morphological features to some degree of accuracy. However, using these tools requires the input would be in Hebrew script.

In the case of the Latin script, although it is orthographically faithful to the Hebrew script, there isn't a one-to-one mapping between them, that is, it is not trivial to automatically transform the Latin script into Hebrew script, due to non-deterministic writing conventions in Hebrew, such as when to include vowels (Matres lectionis). This decision requires attention and examination and native Hebrew speakers can disagree on the writings of words in that aspect (and in others). It is therefore very difficult to automatically transform from one script to another without knowing a-priori what the conventional writing of a word is, because modern models rely on conventional writing to identify words, and using different writings may cause them to misdiagnose a word and lose accuracy.

Future research on the morphology and syntax of CDS can spare the effort of transcribing recordings according to IPA conventions and use modern Hebrew script, which is much faster and more accurate to transcribe, alongside with modern tools, to receive fast and fairly accurate morpho-syntactic analyses.

Chapter 7

Discussion

The following discussion summarizes the main findings of this study, raises questions regarding the relationship between morpho-syntactic acquisition and distribution and suggests ideas for future research.

This study is grounded in the theoretical approach of social-interactionism that emphasizes the learning experiences in the course of language acquisition. To support such theoretical framework domain-general mechanisms must be taken into account, and specifically this study focuses on statistical learning as one such mechanism. In the light of this theory, it was this study's aim to investigate CDS from a perspective that hasn't been investigated before. Most studies on CDS treat it as a unitary phenomenon and analyze it as one piece, however this study shows that CDS is more similar in nature to CS, and therefore should be analyzed in a similar way, essentially as a series of developmental stages, where each adult developmental stage has its own unique properties.

To meet this goal two corpora of CDS and CS were used. First it was established that the CDS corpus can be split into developmental stages based on a measure that is also used for analyzing CS - the MPU. This split yielded adult stages that correspond almost perfectly to the child stages, indicating that the adult is slightly more advanced than the child (the alternation shows that first the adult transitions to the next stage and then the child). The second step involved a thorough analysis of the distribution of morpho-syntactic elements in the CDS, and comparing these distributions with the known ages of acquisition of the same elements. The findings show that for most elements there is a peak in the adult usage of an element slightly before or during the period in which it is acquired. This indicates, as hypothesized, that CDS is not a unitary phenomenon, but rather a fine-tuned developmental process with great sensitivity to the child's abilities and knowledge.

The findings also show that there are upper and lower bounds for acquisition, meaning that the absolute number of occurrences of an element in CDS can strongly affect the age of acquisition. It was shown that for elements with order of magnitude more occurrences the age of acquisition is rather early, even though those elements only peak at a later stage (i.e. these elements exceed the upper bound, so they are acquired before this study's predictions), and also the mirror image of that phenomenon - elements with too few occurrences were not acquired in the CS data (until age 3;6). These thresholds offer a usage-based account for the order of acquisition of morpho-syntactic elements.

This study provides another justification for Newport et al. results that the language environment of the child becomes increasingly more complex in correspondence with the child's language skills, as well as to the results of Moerk that the frequency of the input from the environment is highly related to the frequency of production and thus to the age of acquisition (as it is calculated as a percentage of the correct usage). In another study, Ashkenazi et al. found that the complexity of verb paradigms also grows with the child's age, thus aiding the verb acquisition process.

This study has two main contribution. First it demonstrates the bi-directional relationship and influence that CDS and CS have on each-other. A part of that bi-directionality was also attested in Irvin et al. (2016), that found mutual influence on syntactic elements between mothers and children. Second and more important contribution is the new approach to CDS analysis as a series of developmental stages. This raises several questions regarding the nature of CDS as a sub-system of the adult mature and change-resistant system - is the adult language really change-resistant? If not, what drives the change? There is also a question about the duration in which these two systems co-exist and if at some point the adults "steps down" since the child is now able to make their own examples. Of course, the most interesting question in that aspect is how to incorporate all the environmental input, including the child-generated content and its feedback, into a comprehensive model. Such model will be able to shed new light on the role of input in the process of acquisition as well as on the mutual influence between parents and children.

Future Research

Several directions for future research can be used to replicate and enhance the results of this study. First and most obvious is to use an actual parent-child dyad instead of extrapolating from cross-sectional data. Even better would be to use multiple such dyads, if this sort of data can be provided. Another understudied aspect is the role of parental input beyond the age of 3;6, that can suggest what happens to the adult system once the child can make their own examples.

The split into developmental stages, being this study's biggest contribution, can be further studied and generalized over a large number of CDS corpora. It may be interesting to see whether there are differences in this aspect between parents of typically developing children and parents of children with speech and communication disorders.

Gender differences are also interesting, Ashkenazi et al. (2015) found that there are significant differences between CDS that is addressed to boys and CDS that is addressed to girls, but their primary focus was on verb paradigms. This idea can be extended to include all morpho-syntactic elements that are influenced by grammatical gender and provide a unified analysis of the differences in acquisition of these elements between boys and girls as well as the differences in the distribution of these elements in CDS (addressed to either boys or girls).

On the technical side, modern tools for dependency analysis can be used to re-analyze the entire corpus. Such tools can also eliminate a large part of the complexity of transcribing speech, as they work directly on the Hebrew script, on which the inter-annotator agreement is significantly higher than on the appropriate IPA transcription (as evident in the current state of the Hebrew corpora in the CHILDES project). The output of these tools is a dependency analysis between words (or morphemes) in an utterance, and this analysis can be expanded further to include dependencies on more complex phenomena such as co-reference, raising verbs, etc. and can also provide insights on the developmental path of other relatively less studied aspects of language like discourse.

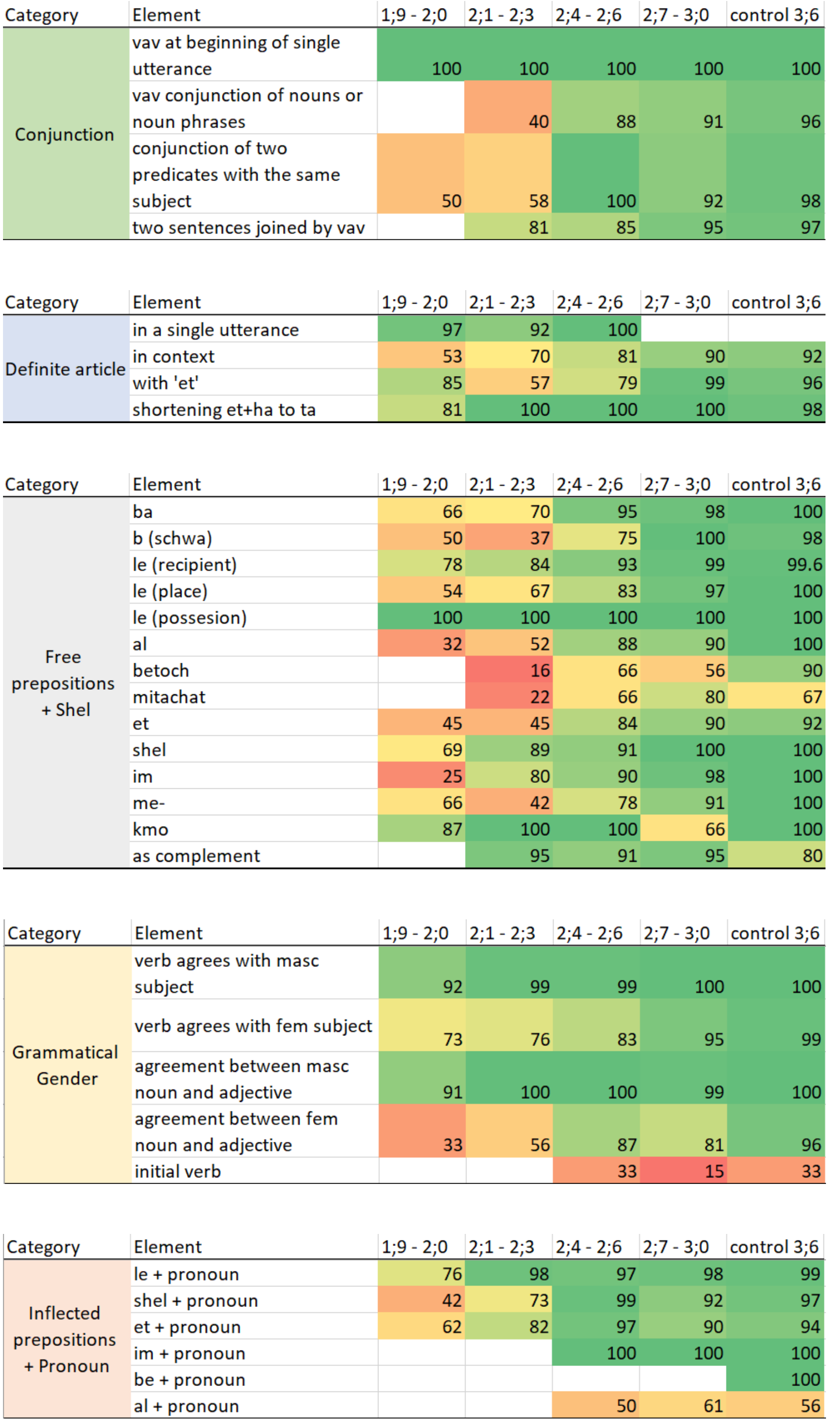
Finally, incorporating child-generated examples and the parental feedback that follows (or lack thereof) into the predictive model and then examining the distribution of the joint data. Taken together the directions for future research suggested here can shed new light on the mutual influence between CDS and CS and the general role of social interaction in the process of language acquisition .

Appendix A

Supplementary Material - CS

A.1 Kaplan's Results - Exhaustive List of Correct Usage

Figure A.1 below is an comprehension of all the results tables from Kaplan's study. As was stated in Chapter 3, it is clear from the tables that the morpho-syntactic element do not exhibit a U-shaped learning curve, but rather develop linearly. The tables are conditionally formatted as a heat map, with 70 as the middle value (coded with yellow). It is important to note that these are raw data from Kaplan's study and were not manipulated in this study, the only intervention is the coloring.



Category	Element	1;9 - 2;0	2;1 - 2;3	2;4 - 2;6	2;7 - 3;0	control 3;6
Conjunction	vav at beginning of single utterance	100	100	100	100	100
	vav conjunction of nouns or noun phrases		40	88	91	96
	conjunction of two predicates with the same subject	50	58	100	92	98
	two sentences joined by vav		81	85	95	97

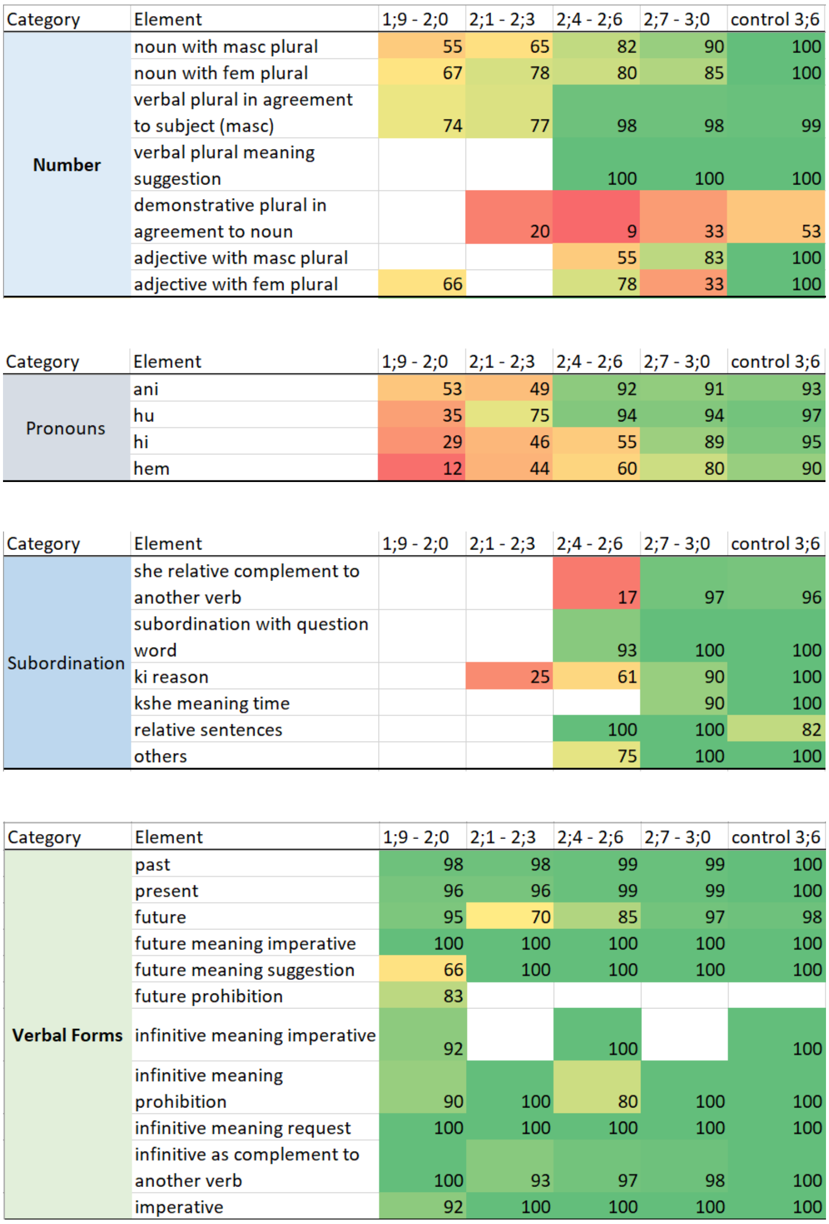
Category	Element	1;9 - 2;0	2;1 - 2;3	2;4 - 2;6	2;7 - 3;0	control 3;6
Definite article	in a single utterance	97	92	100		
	in context	53	70	81	90	92
	with 'et'	85	57	79	99	96
	shortening et+ha to ta	81	100	100	100	98

Category	Element	1;9 - 2;0	2;1 - 2;3	2;4 - 2;6	2;7 - 3;0	control 3;6
Free prepositions + Shel	ba	66	70	95	98	100
	b (schwa)	50	37	75	100	98
	le (recipient)	78	84	93	99	99.6
	le (place)	54	67	83	97	100
	le (possesion)	100	100	100	100	100
	al	32	52	88	90	100
	betoch		16	66	56	90
	mitachat		22	66	80	67
	et	45	45	84	90	92
	shel	69	89	91	100	100
	im	25	80	90	98	100
	me-	66	42	78	91	100
	kmo	87	100	100	66	100
	as complement		95	91	95	80

Category	Element	1;9 - 2;0	2;1 - 2;3	2;4 - 2;6	2;7 - 3;0	control 3;6
Grammatical Gender	verb agrees with masc subject	92	99	99	100	100
	verb agrees with fem subject	73	76	83	95	99
	agreement between masc noun and adjective	91	100	100	99	100
	agreement between fem noun and adjective	33	56	87	81	96
	initial verb			33	15	33

Category	Element	1;9 - 2;0	2;1 - 2;3	2;4 - 2;6	2;7 - 3;0	control 3;6
Inflected prepositions + Pronoun	le + pronoun	76	98	97	98	99
	shel + pronoun	42	73	99	92	97
	et + pronoun	62	82	97	90	94
	im + pronoun			100	100	100
	be + pronoun					100
	al + pronoun			50	61	56

Figure A.1



Category	Element	1;9 - 2;0	2;1 - 2;3	2;4 - 2;6	2;7 - 3;0	control 3;6
Number	noun with masc plural	55	65	82	90	100
	noun with fem plural	67	78	80	85	100
	verbal plural in agreement to subject (masc)	74	77	98	98	99
	verbal plural meaning suggestion			100	100	100
	demonstrative plural in agreement to noun		20	9	33	53
	adjective with masc plural			55	83	100
	adjective with fem plural	66		78	33	100

Category	Element	1;9 - 2;0	2;1 - 2;3	2;4 - 2;6	2;7 - 3;0	control 3;6
Pronouns	ani	53	49	92	91	93
	hu	35	75	94	94	97
	hi	29	46	55	89	95
	hem	12	44	60	80	90

Category	Element	1;9 - 2;0	2;1 - 2;3	2;4 - 2;6	2;7 - 3;0	control 3;6
Subordination	she relative complement to another verb			17	97	96
	subordination with question word			93	100	100
	ki reason		25	61	90	100
	kshe meaning time				90	100
	relative sentences			100	100	82
	others			75	100	100

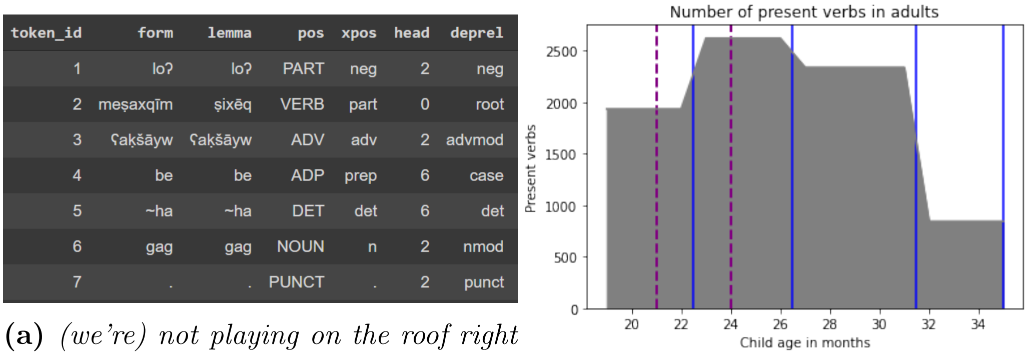
Category	Element	1;9 - 2;0	2;1 - 2;3	2;4 - 2;6	2;7 - 3;0	control 3;6
Verbal Forms	past	98	98	99	99	100
	present	96	96	99	99	100
	future	95	70	85	97	98
	future meaning imperative	100	100	100	100	100
	future meaning suggestion	66	100	100	100	100
	future prohibition	83				
	infinitive meaning imperative	92		100		100
	infinitive meaning prohibition	90	100	80	100	100
	infinitive meaning request	100	100	100	100	100
	infinitive as complement to another verb	100	93	97	98	100
	imperative	92	100	100	100	100

Figure A.1: Percentage of correct usage of elements by Kaplan

A.2 Results for all morpho-syntactic elements

The following figures contain all the morpho-syntactic element that were examined in this study, including the ones presented in Chapter 5. For each element the right-hand side of the figure displays the distribution of this element in CDS (grey area) and the age of acquisition in children (dashed purple lines). The left-hand side contains an example sentence from the CDS corpus that uses this element. The elements are organized by their age of acquisition in children. Note that for elements that were eventually acquired by the age of 3;6 the purple dashed lines are not marked since the exact period is unknown, and the same holds for elements that were not acquired at all by the end of Kaplan's study.

A.2.1 Age 1;9 - 2;0 (21-24 months)

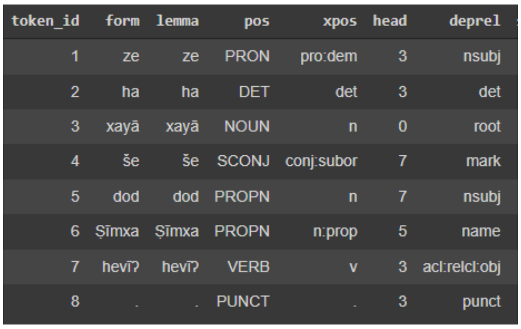


token_id	form	lemma	pos	xpos	head	deprel
1	loʔ	loʔ	PART	neg	2	neg
2	meṣaxqīm	ṣixēq	VERB	part	0	root
3	ʕakšāyw	ʕakšāyw	ADV	adv	2	advmod
4	be	be	ADP	prep	6	case
5	~ha	~ha	DET	det	6	det
6	gag	gag	NOUN	n	2	nmod
7	.	.	PUNCT	.	2	punct

(a) *(we're) not playing on the roof right now*

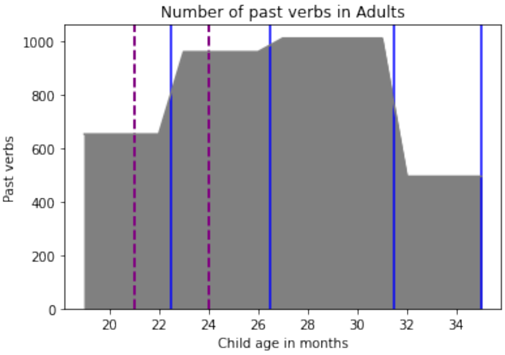
(b) Distribution of present tense verbs

Figure A.2: Present tense verbs, example and distribution



token_id	form	lemma	pos	xpos	head	deprel
1	ze	ze	PRON	pro:dem	3	nsubj
2	ha	ha	DET	det	3	det
3	xayā	xayā	NOUN	n	0	root
4	še	še	SCONJ	conj:subor	7	mark
5	dod	dod	PROPN	n	7	nsubj
6	Ṣīmxa	Ṣīmxa	PROPN	n:prop	5	name
7	hevīʔ	hevīʔ	VERB	v	3	acl:relcl:obj
8	.	.	PUNCT	.	3	punct

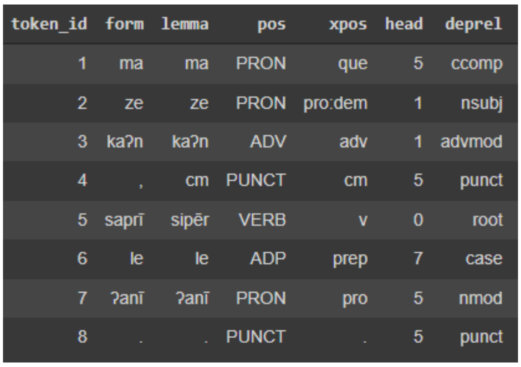
(a) *That's the animal that uncle Simcha brought.*



(b) Distribution of past tense verbs

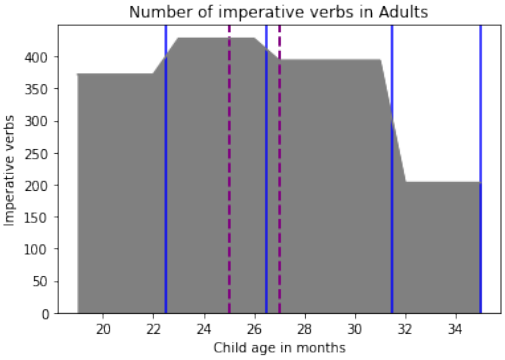
Figure A.3: Past tense verbs, example and distribution

A.2.2 Age 2;1 - 2;3 (25-27 months)



token_id	form	lemma	pos	xpos	head	deprel
1	ma	ma	PRON	que	5	ccomp
2	ze	ze	PRON	pro:dem	1	nsubj
3	kaʔn	kaʔn	ADV	adv	1	advmod
4	,	cm	PUNCT	cm	5	punct
5	saprī	sipēr	VERB	v	0	root
6	le	le	ADP	prep	7	case
7	ʔanī	ʔanī	PRON	pro	5	nmod
8	.	.	PUNCT	.	5	punct

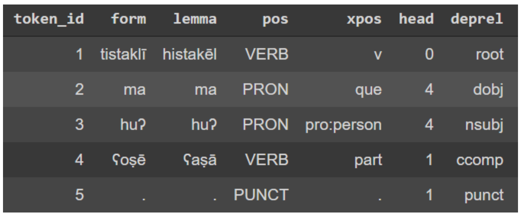
(a) *What is that here, tell me.*



(b) Distribution of imperative verbs

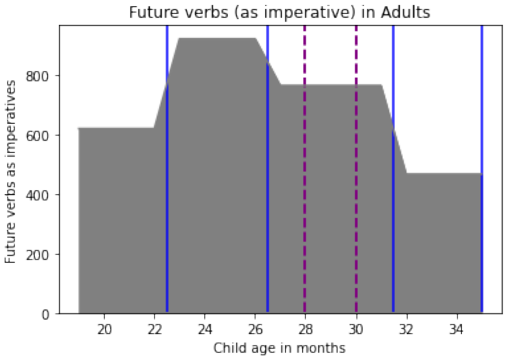
Figure A.4: Imperative verbs, example and distribution

A.2.3 Age 2;4 - 2;6 (28-30 months)



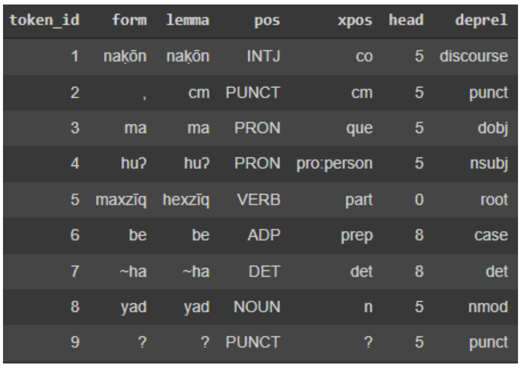
token_id	form	lemma	pos	xpos	head	deprel
1	tistaklī	histakēl	VERB	v	0	root
2	ma	ma	PRON	que	4	dobj
3	huʔ	huʔ	PRON	pro:person	4	nsubj
4	ʕoṣē	ʕaṣā	VERB	part	1	ccomp
5	.	.	PUNCT	.	1	punct

(a) *look what he's doing*



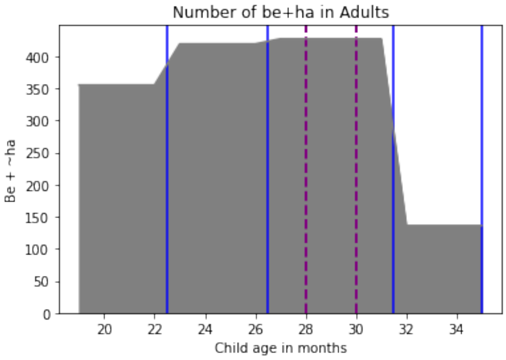
(b) Distribution of future forms as imperative

Figure A.5: Future forms to convey imperative meaning, example and distribution



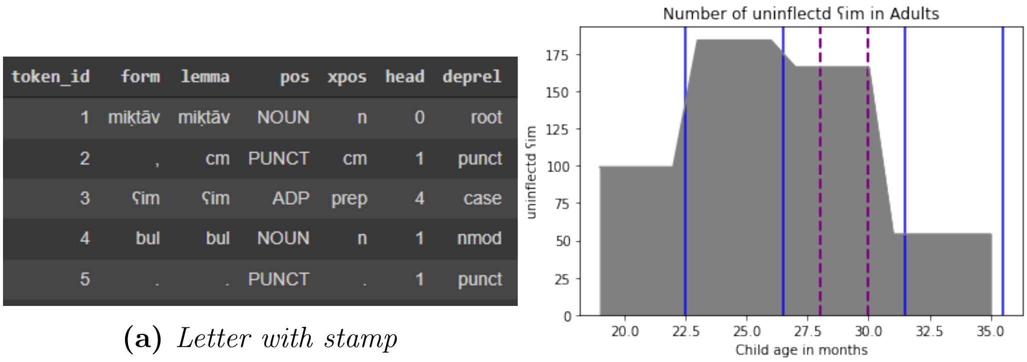
token_id	form	lemma	pos	xpos	head	deprel
1	naḳōn	naḳōn	INTJ	co	5	discourse
2	,	cm	PUNCT	cm	5	punct
3	ma	ma	PRON	que	5	dobj
4	huʔ	huʔ	PRON	pro:person	5	nsubj
5	maxzīq	hexzīq	VERB	part	0	root
6	be	be	ADP	prep	8	case
7	~ha	~ha	DET	det	8	det
8	yad	yad	NOUN	n	5	nmod
9	?	?	PUNCT	?	5	punct

(a) *Right, what is he holding in the hand?*



(b) Distribution of be + ha

Figure A.6: 'In + the', example and distribution

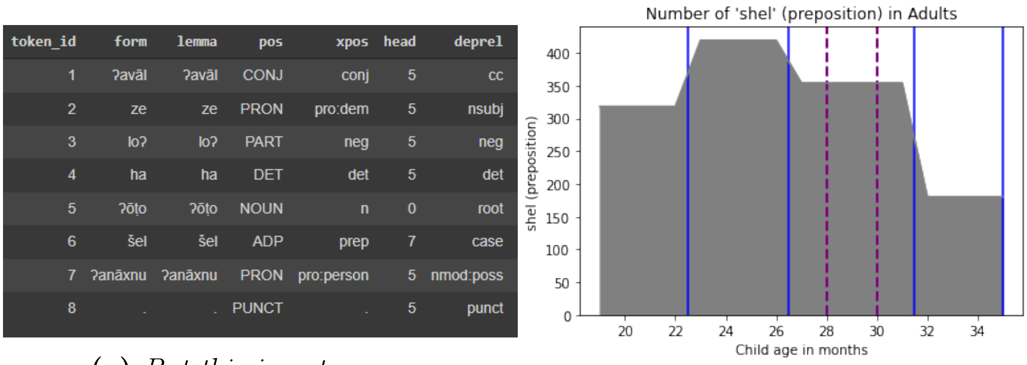


token_id	form	lemma	pos	xpos	head	deprel
1	miḳtāv	miḳtāv	NOUN	n	0	root
2	,	cm	PUNCT	cm	1	punct
3	ʕim	ʕim	ADP	prep	4	case
4	bul	bul	NOUN	n	1	nmod
5	.	.	PUNCT	.	1	punct

(a) *Letter with stamp*

(b) Distribution of isolated im

Figure A.7: Isolated 'with'(preposition), example and distribution

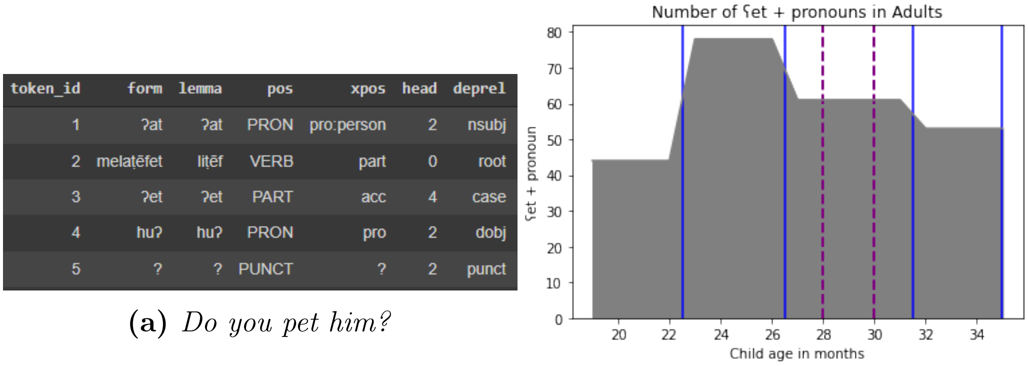


token_id	form	lemma	pos	xpos	head	deprel
1	ʔavāl	ʔavāl	CONJ	conj	5	cc
2	ze	ze	PRON	pro:dem	5	nsubj
3	loʔ	loʔ	PART	neg	5	neg
4	ha	ha	DET	det	5	det
5	ʔōṭo	ʔōṭo	NOUN	n	0	root
6	šel	šel	ADP	prep	7	case
7	ʔanāxnu	ʔanāxnu	PRON	pro:person	5	nmod:poss
8	.	.	PUNCT	.	5	punct

(a) *But this is not our car*

(b) Distribution of shel

Figure A.8: shel (possessive preposition), example and distribution

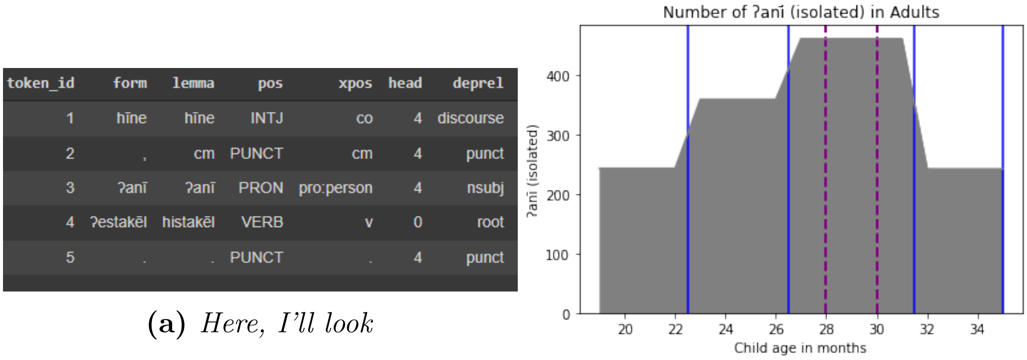


token_id	form	lemma	pos	xpos	head	deprel
1	ʔat	ʔat	PRON	pro:person	2	nsubj
2	melaṭēfet	liṭēf	VERB	part	0	root
3	ʔet	ʔet	PART	acc	4	case
4	huʔ	huʔ	PRON	pro	2	dobj
5	?	?	PUNCT	?	2	punct

(a) *Do you pet him?*

(b) Distribution of et + pronoun

Figure A.9: et (accusative marker) + pronoun, example and distribution

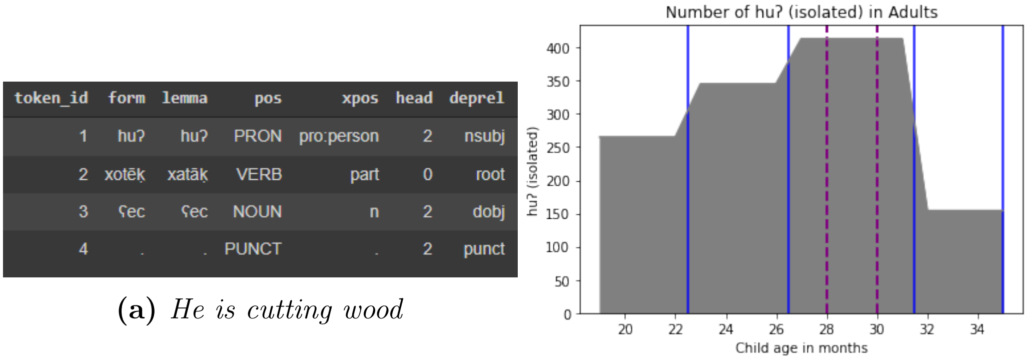


token_id	form	lemma	pos	xpos	head	deprel
1	hīne	hīne	INTJ	co	4	discourse
2	,	cm	PUNCT	cm	4	punct
3	ʔanī	ʔanī	PRON	pro:person	4	nsubj
4	ʔestakēl	histakēl	VERB	v	0	root
5	.	.	PUNCT	.	4	punct

(a) *Here, I'll look*

(b) Distribution of pronoun 'I'

Figure A.10: I (1sc pronoun), example and distribution

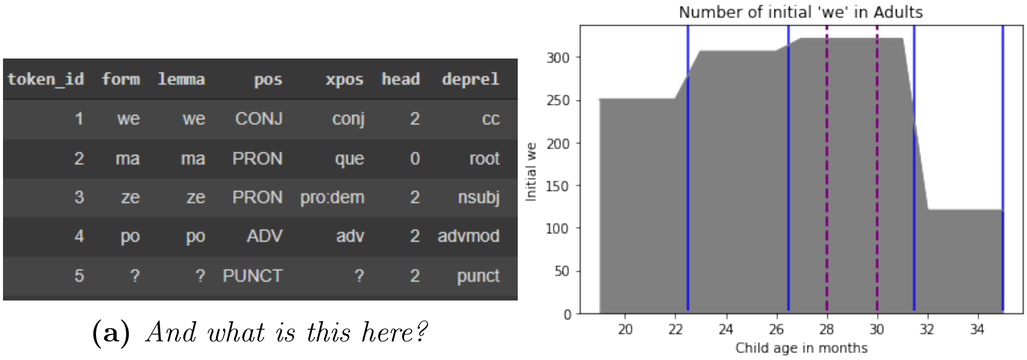


token_id	form	lemma	pos	xpos	head	deprel
1	huʔ	huʔ	PRON	pro:person	2	nsubj
2	xotēķ	xatāķ	VERB	part	0	root
3	ʕec	ʕec	NOUN	n	2	dobj
4	.	.	PUNCT	.	2	punct

(a) *He is cutting wood*

(b) Distribution of pronoun 'He'

Figure A.11: He (3ms pronoun), example and distribution



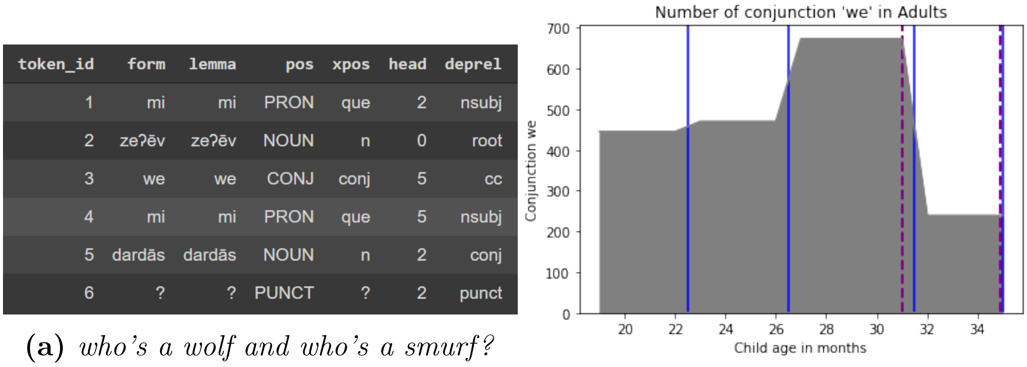
token_id	form	lemma	pos	xpos	head	deprel
1	we	we	CONJ	conj	2	cc
2	ma	ma	PRON	que	0	root
3	ze	ze	PRON	pro:dem	2	nsubj
4	po	po	ADV	adv	2	advmod
5	?	?	PUNCT	?	2	punct

(a) *And what is this here?*

(b) Distribution of initial 'we'

Figure A.12: Initial conjunction marker, example and distribution

A.2.4 Age 2;7 - 3;0 (31-36 months)

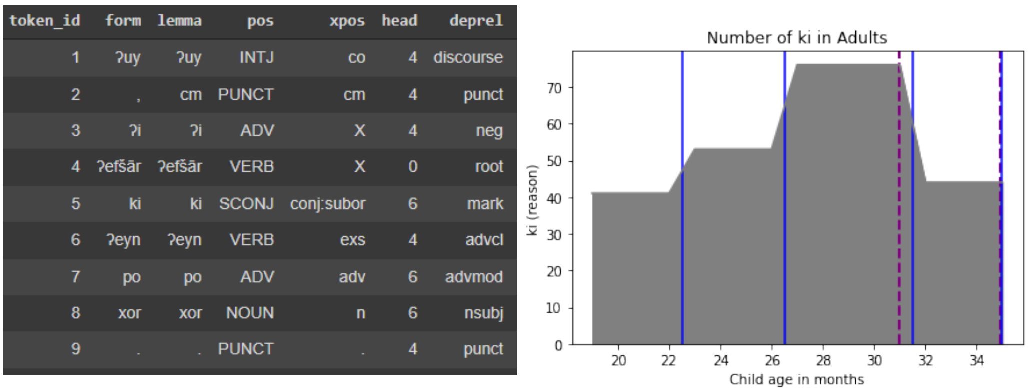


token_id	form	lemma	pos	xpos	head	deprel
1	mi	mi	PRON	que	2	nsubj
2	zeʔēv	zeʔēv	NOUN	n	0	root
3	we	we	CONJ	conj	5	cc
4	mi	mi	PRON	que	5	nsubj
5	dardās	dardās	NOUN	n	2	conj
6	?	?	PUNCT	?	2	punct

(a) *who's a wolf and who's a smurf?*

(b) Distribution of regular conjunction

Figure A.13: Regular conjunction, example and distribution

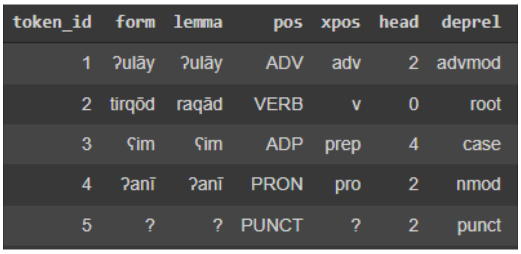


token_id	form	lemma	pos	xpos	head	deprel
1	ʔuy	ʔuy	INTJ	co	4	discourse
2	,	cm	PUNCT	cm	4	punct
3	ʔi	ʔi	ADV	X	4	neg
4	ʔefšār	ʔefšār	VERB	X	0	root
5	ki	ki	SCONJ	conj:subor	6	mark
6	ʔeyn	ʔeyn	VERB	exs	4	advcl
7	po	po	ADV	adv	6	advmod
8	xor	xor	NOUN	n	6	nsubj
9	.	.	PUNCT	.	4	punct

(a) *Oh, it's impossible because there is no hole.*

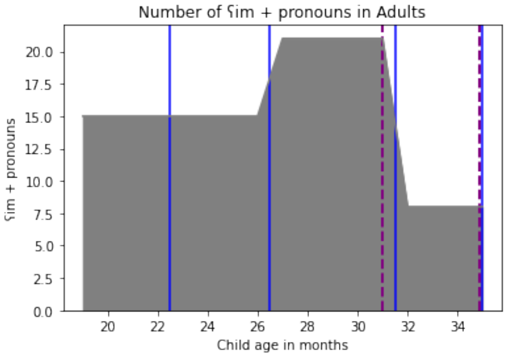
(b) Distribution of word 'ki'

Figure A.14: 'ki' for reasoning, example and distribution



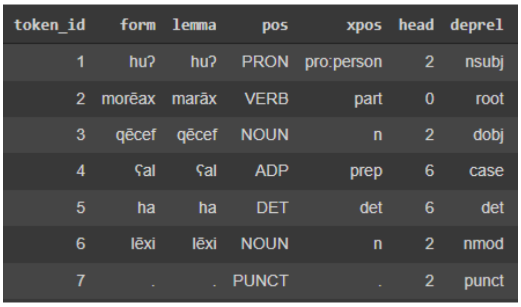
token_id	form	lemma	pos	xpos	head	deprel
1	ʔulāy	ʔulāy	ADV	adv	2	advmod
2	tirqōd	raqād	VERB	v	0	root
3	ʕim	ʕim	ADP	prep	4	case
4	ʔanī	ʔanī	PRON	pro	2	nmod
5	?	?	PUNCT	?	2	punct

(a) *Maybe dance with me?*



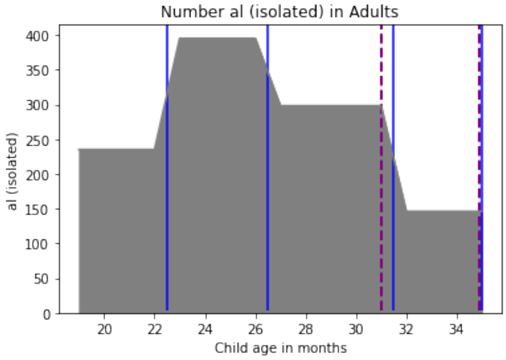
(b) Distribution of preposition 'im' + pronoun

Figure A.15: with + pronoun, example and distribution



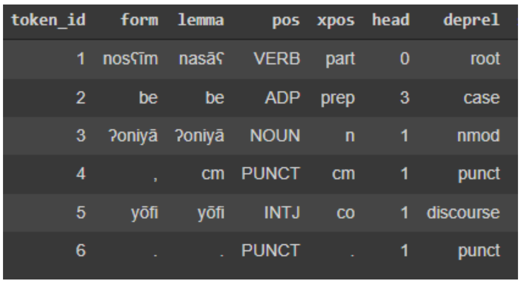
token_id	form	lemma	pos	xpos	head	deprel
1	huʔ	huʔ	PRON	pro:person	2	nsubj
2	morēax	marāx	VERB	part	0	root
3	qēcef	qēcef	NOUN	n	2	dobj
4	ʕal	ʕal	ADP	prep	6	case
5	ha	ha	DET	det	6	det
6	lēxi	lēxi	NOUN	n	2	nmod
7	.	.	PUNCT	.	2	punct

(a) *He rubs cream on the cheek*



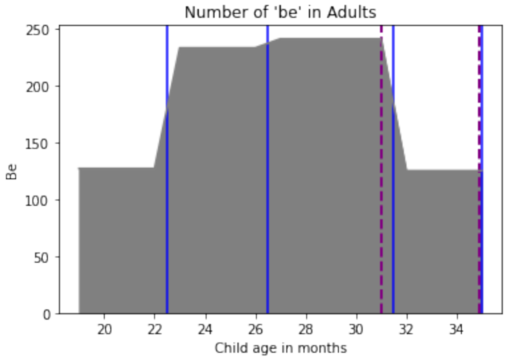
(b) Distribution of preposition 'al' in isolation

Figure A.16: on/about in isolation, example and distribution



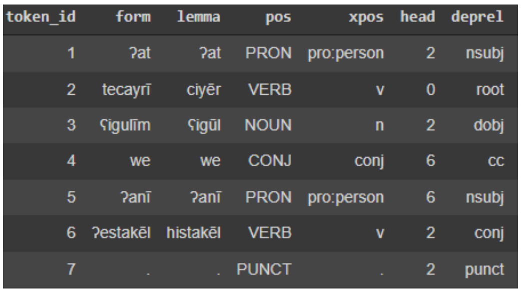
token_id	form	lemma	pos	xpos	head	deprel
1	nosʕīm	nasāʕ	VERB	part	0	root
2	be	be	ADP	prep	3	case
3	ʔoniyā	ʔoniyā	NOUN	n	1	nmod
4	,	cm	PUNCT	cm	1	punct
5	yōfi	yōfi	INTJ	co	1	discourse
6	.	.	PUNCT	.	1	punct

(a) *Travelling in a boat, good*



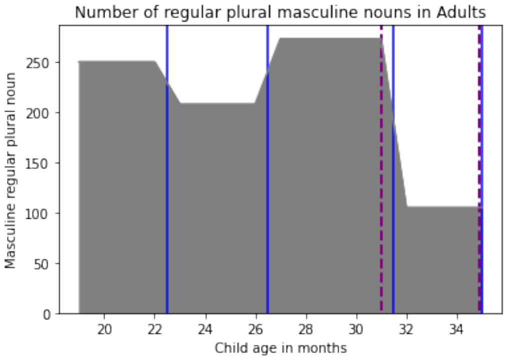
(b) Distribution of preposition ‘be’ in isolation

Figure A.17: ‘In’ in isolation, example and distribution



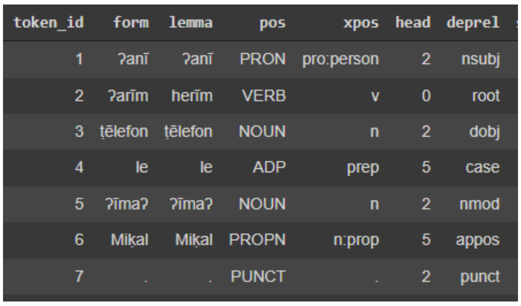
token_id	form	lemma	pos	xpos	head	deprel
1	ʔat	ʔat	PRON	pro:person	2	nsubj
2	tecayrī	ciyēr	VERB	v	0	root
3	ʕigulīm	ʕigūl	NOUN	n	2	dobj
4	we	we	CONJ	conj	6	cc
5	ʔanī	ʔanī	PRON	pro:person	6	nsubj
6	ʔestakēl	histakēl	VERB	v	2	conj
7	.	.	PUNCT	.	2	punct

(a) *You’ll draw circles and I’ll watch*



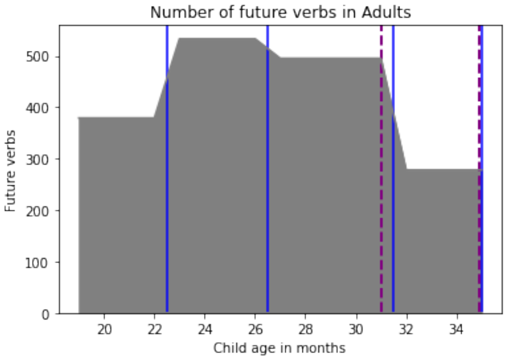
(b) Distribution of plural nouns (masculine suffix)

Figure A.18: Plural nouns (masculine siffix), example and distribution



token_id	form	lemma	pos	xpos	head	deprel
1	ʔanī	ʔanī	PRON	pro:person	2	nsubj
2	ʔarīm	herīm	VERB	v	0	root
3	ṭēlefon	ṭēlefon	NOUN	n	2	dobj
4	le	le	ADP	prep	5	case
5	ʔīmaʔ	ʔīmaʔ	NOUN	n	2	nmod
6	Mikal	Mikal	PROPN	n:prop	5	appos
7	.	.	PUNCT	.	2	punct

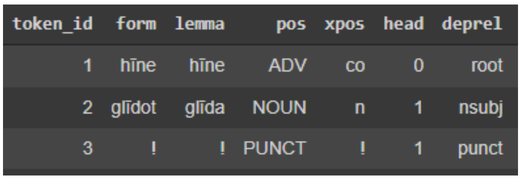
(a) *I’ll call Mother Michal*



(b) Distribution of future tense verbs

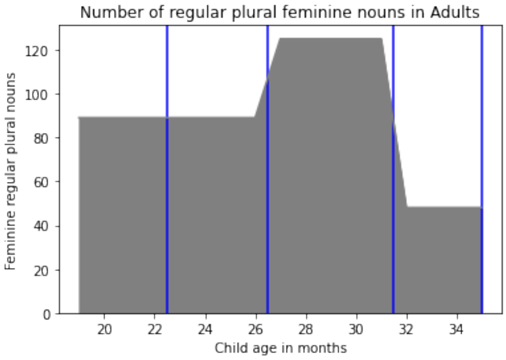
Figure A.19: Future tense verbs, example and distribution

A.2.5 Acquired by 3;6 (42 months)



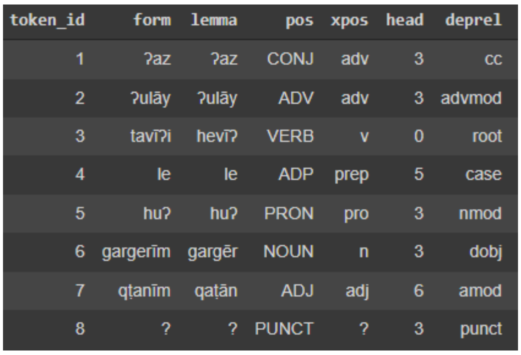
token_id	form	lemma	pos	xpos	head	deprel
1	hīne	hīne	ADV	co	0	root
2	glīdot	glīda	NOUN	n	1	nsubj
3	!	!	PUNCT	!	1	punct

(a) *Here's ice cream*



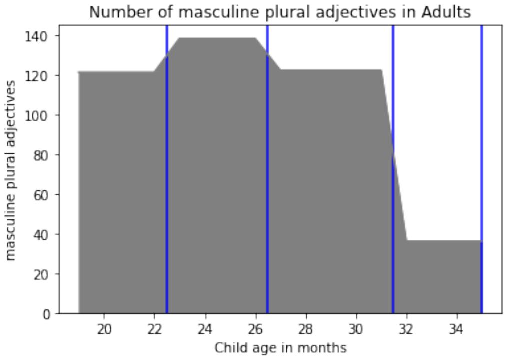
(b) Distribution of plural nouns with feminine suffix

Figure A.20: Plural nouns feminine suffix, example and distribution



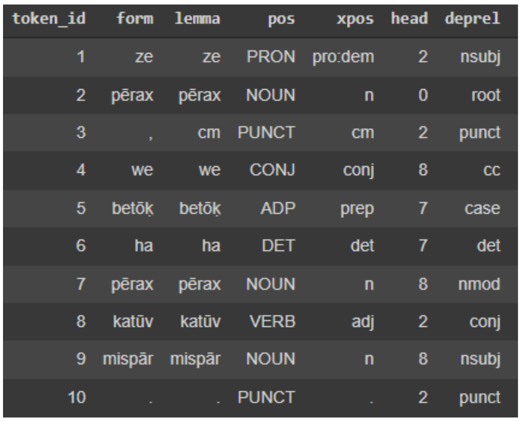
token_id	form	lemma	pos	xpos	head	deprel
1	ʔaz	ʔaz	CONJ	adv	3	cc
2	ʔulāy	ʔulāy	ADV	adv	3	advmod
3	tavīʔi	hevīʔ	VERB	v	0	root
4	le	le	ADP	prep	5	case
5	huʔ	huʔ	PRON	pro	3	nmod
6	gargerīm	gargēr	NOUN	n	3	dobj
7	qṭanīm	qaṭān	ADJ	adj	6	amod
8	?	?	PUNCT	?	3	punct

(a) *So maybe bring him little crumbs?*



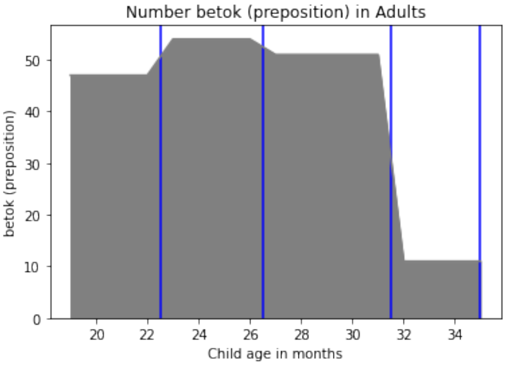
(b) Distribution of plural masculine adjectives

Figure A.21: Future plural adjectives masculine suffix, example and distribution



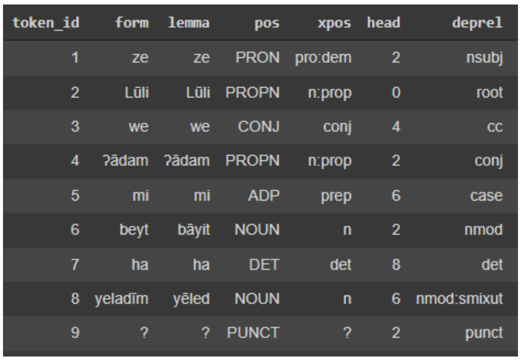
token_id	form	lemma	pos	xpos	head	deprel
1	ze	ze	PRON	pro:dem	2	nsubj
2	pērax	pērax	NOUN	n	0	root
3	,	cm	PUNCT	cm	2	punct
4	we	we	CONJ	conj	8	cc
5	betōḳ	betōḳ	ADP	prep	7	case
6	ha	ha	DET	det	7	det
7	pērax	pērax	NOUN	n	8	nmod
8	katūv	katūv	VERB	adj	2	conj
9	mispār	mispār	NOUN	n	8	nsubj
10	.	.	PUNCT	.	2	punct

(a) *That's a flower, and inside the flower there's a number*



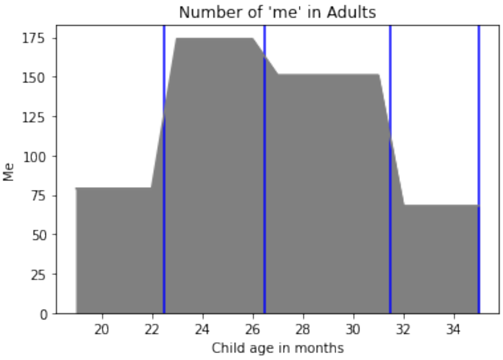
(b) Distribution of preposition 'betox'

Figure A.22: Preposition 'betox' (inside), example and distribution



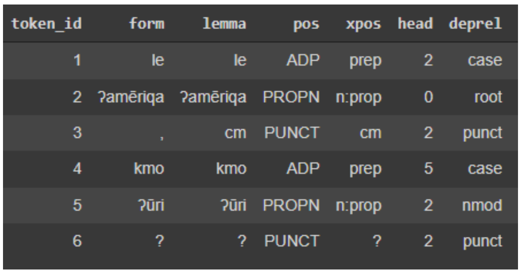
token_id	form	lemma	pos	xpos	head	deprel
1	ze	ze	PRON	pro:dem	2	nsubj
2	Lūli	Lūli	PROPN	n:prop	0	root
3	we	we	CONJ	conj	4	cc
4	ʔādam	ʔādam	PROPN	n:prop	2	conj
5	mi	mi	ADP	prep	6	case
6	beyt	bāyit	NOUN	n	2	nmod
7	ha	ha	DET	det	8	det
8	yeladīm	yēled	NOUN	n	6	nmod:smixut
9	?	?	PUNCT	?	2	punct

(a) *Are those Luli and Adam from the children's house?*



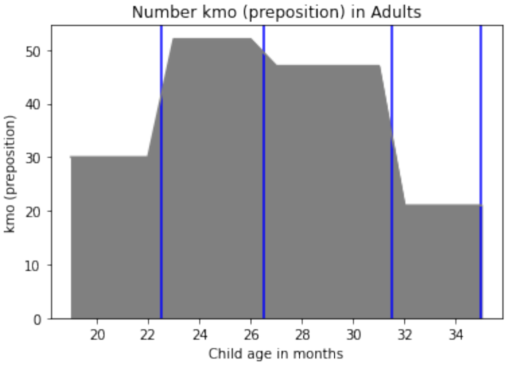
(b) Distribution of preposition 'me'

Figure A.23: Preposition 'me' (from), example and distribution



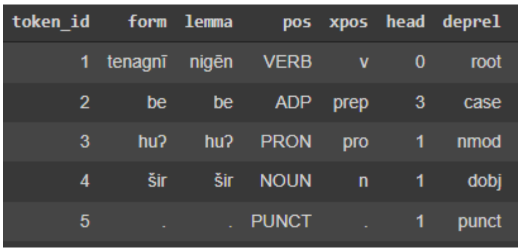
token_id	form	lemma	pos	xpos	head	deprel
1	le	le	ADP	prep	2	case
2	ʔamēriqa	ʔamēriqa	PROPN	n:prop	0	root
3	,	cm	PUNCT	cm	2	punct
4	kmo	kmo	ADP	prep	5	case
5	ʔūri	ʔūri	PROPN	n:prop	2	nmod
6	?	?	PUNCT	?	2	punct

(a) *To America, like Uri?*



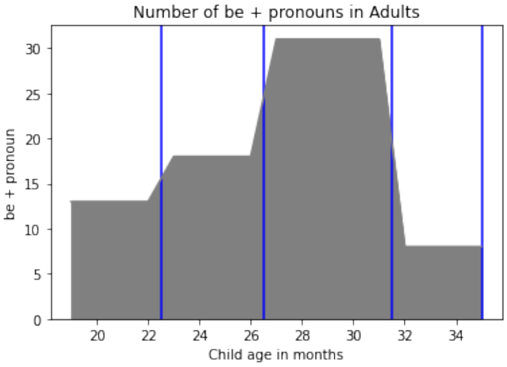
(b) Distribution of preposition ‘kmo’

Figure A.24: Preposition ‘kmo’ (like), example and distribution



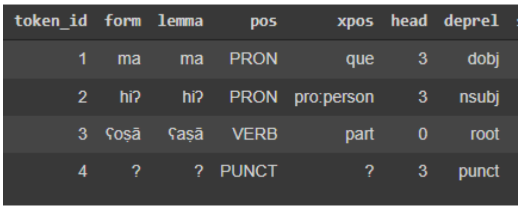
token_id	form	lemma	pos	xpos	head	deprel
1	tenagnī	nigēn	VERB	v	0	root
2	be	be	ADP	prep	3	case
3	huʔ	huʔ	PRON	pro	1	nmod
4	šir	šir	NOUN	n	1	dobj
5	.	.	PUNCT	.	1	punct

(a) *Play a song with it*



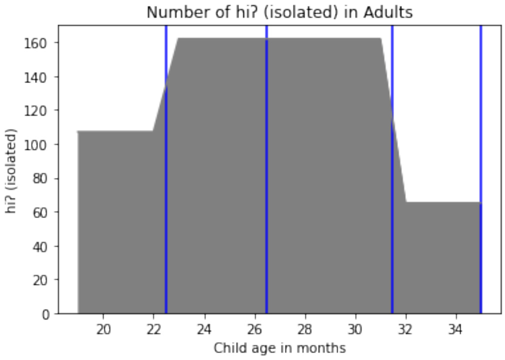
(b) Distribution of preposition ‘be’ + pronoun

Figure A.25: Preposition ‘be’ (in) + pronoun, example and distribution



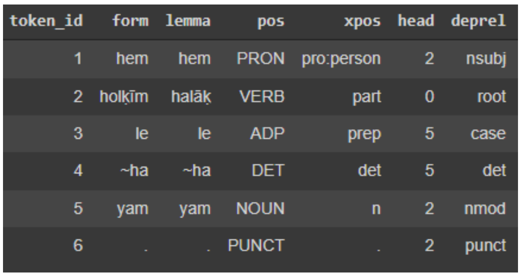
token_id	form	lemma	pos	xpos	head	deprel
1	ma	ma	PRON	que	3	dobj
2	hiʔ	hiʔ	PRON	pro:person	3	nsubj
3	ʕoṣā	ʕaṣā	VERB	part	0	root
4	?	?	PUNCT	?	3	punct

(a) *What is she doing?*



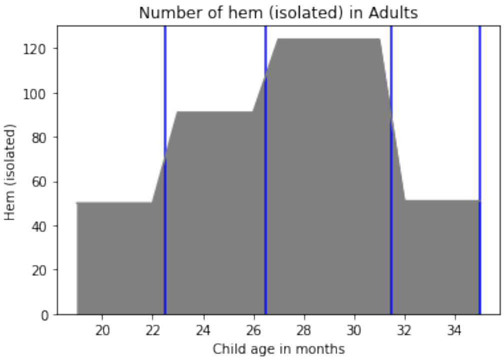
(b) Distribution of pronoun ‘hi’ (she)

Figure A.26: Pronoun ‘hi’ (3fs), example and distribution



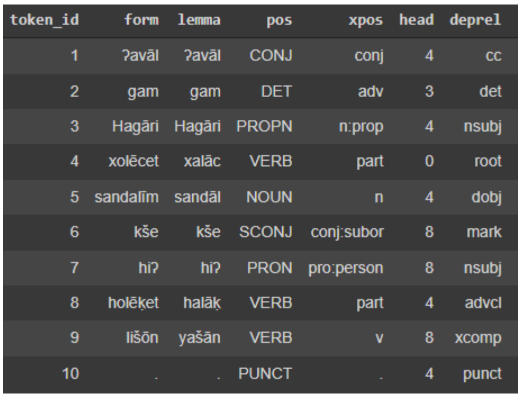
token_id	form	lemma	pos	xpos	head	deprel
1	hem	hem	PRON	pro:person	2	nsubj
2	holkīm	halāķ	VERB	part	0	root
3	le	le	ADP	prep	5	case
4	~ha	~ha	DET	det	5	det
5	yam	yam	NOUN	n	2	nmod
6	.	.	PUNCT	.	2	punct

(a) *They are going to the beach*



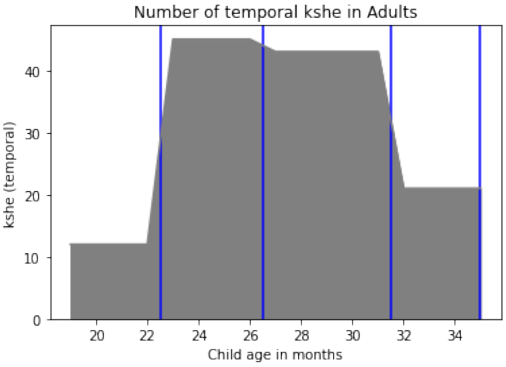
(b) Distribution of preposition ‘hem’ (they)

Figure A.27: Pronoun ‘hem’ (3mp), example and distribution



token_id	form	lemma	pos	xpos	head	deprel
1	ʔavāl	ʔavāl	CONJ	conj	4	cc
2	gam	gam	DET	adv	3	det
3	Hagāri	Hagāri	PROPN	n:prop	4	nsubj
4	xolēcet	xalāc	VERB	part	0	root
5	sandalīm	sandāl	NOUN	n	4	dobj
6	kše	kše	SCONJ	conj:subor	8	mark
7	hiʔ	hiʔ	PRON	pro:person	8	nsubj
8	holēķet	halāķ	VERB	part	4	advcl
9	lišōn	yašān	VERB	v	8	xcomp
10	.	.	PUNCT	.	4	punct

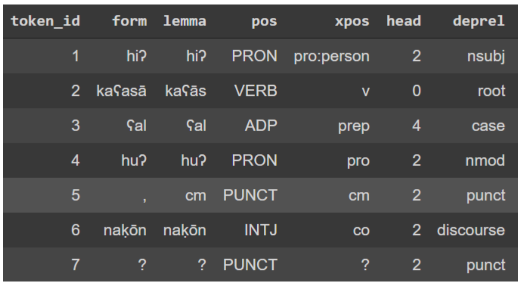
(a) *But Hagari also takes her sandals off when she goes to sleep*



(b) Distribution of temporal ‘kshe’ (when)

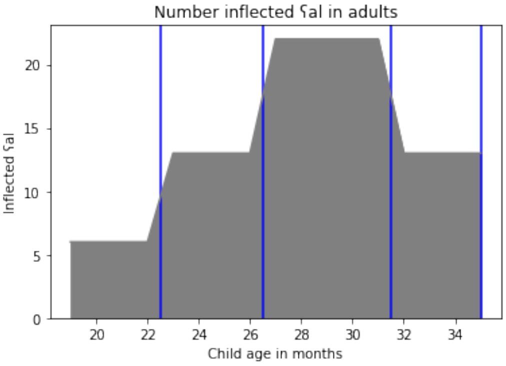
Figure A.28: Complementizer ‘kshe’ (when) , example and distribution

A.2.6 Not acquired by 3;6



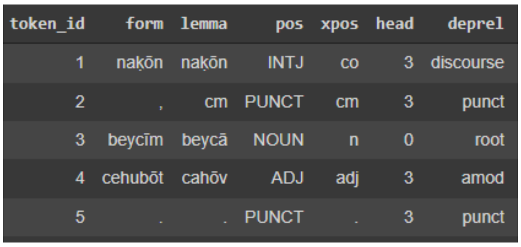
token_id	form	lemma	pos	xpos	head	deprel
1	hiʔ	hiʔ	PRON	pro:person	2	nsubj
2	kaʕasā	kaʕās	VERB	v	0	root
3	ʕal	ʕal	ADP	prep	4	case
4	huʔ	huʔ	PRON	pro	2	nmod
5	,	cm	PUNCT	cm	2	punct
6	naḳōn	naḳōn	INTJ	co	2	discourse
7	?	?	PUNCT	?	2	punct

(a) *she was angry with him, right?*



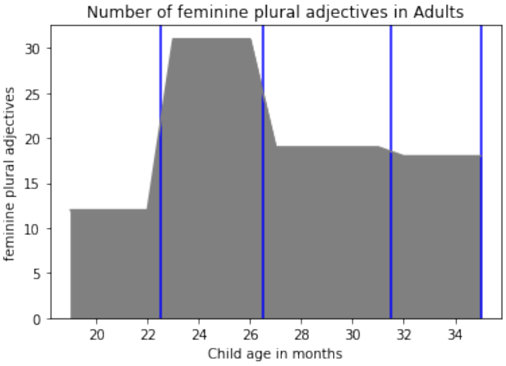
(b) Distribution of inflected ʕal

Figure A.29: Inflected ʕal, example and distribution



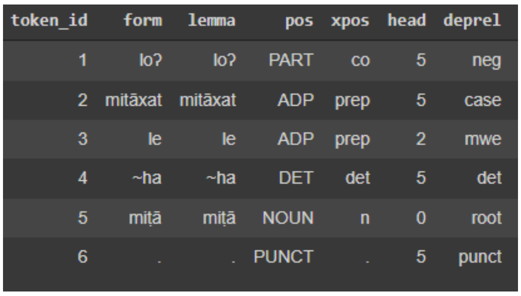
token_id	form	lemma	pos	xpos	head	deprel
1	naḳōn	naḳōn	INTJ	co	3	discourse
2	,	cm	PUNCT	cm	3	punct
3	beycīm	beycā	NOUN	n	0	root
4	cehubōt	cahōv	ADJ	adj	3	amod
5	.	.	PUNCT	.	3	punct

(a) *Right, yellow eggs.*



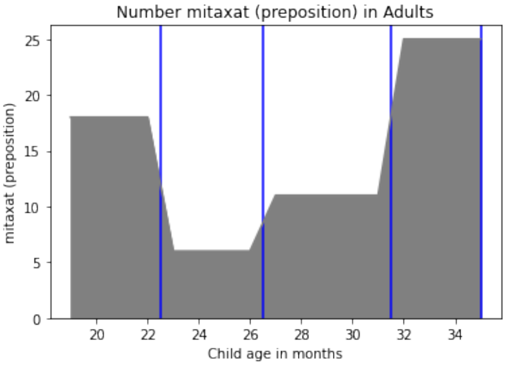
(b) Distribution of plural feminine adjectives

Figure A.30: Plural adjectives (feminine suffix) , example and distribution



token_id	form	lemma	pos	xpos	head	deprel
1	lo?	lo?	PART	co	5	neg
2	mitāxat	mitāxat	ADP	prep	5	case
3	le	le	ADP	prep	2	mwe
4	~ha	~ha	DET	det	5	det
5	miṭā	miṭā	NOUN	n	0	root
6	.	.	PUNCT	.	5	punct

(a) *Not under the bed.*



(b) Distribution of preposition ‘mitaxat’ (under)

Figure A.31: Preposition ‘mitaxat’ (under) , example and distribution

Bibliography

Albert, A., MacWhinney, B., Nir, B., and Wintner, S. (2013). The hebrew childes corpus: transcription and morphological analysis. *Language resources and evaluation*, 47(4):973–1005.

Ambridge, B., Kidd, E., Rowland, C. F., and Theakston, A. L. (2015). The ubiquity of frequency effects in first language acquisition. *Journal of child language*, 42(2):239–273.

Armon-Lotem, S. and Berman, R. A. (2003). The emergence of grammar: Early verbs and beyond. *Journal of Child Language*, 30(4):845.

Arnon, I. (2015). What can frequency effects tell us about the building blocks and mechanisms of language learning? *Journal of Child Language*, 42(2):274–277.

Arnon, I. (2016). The nature of cds in hebrew: Frequent frames in a morphologically rich language. In *Acquisition and Development of Hebrew*, pages 201–224. John Benjamins.

Arnon, I. (2021). The starting big approach to language learning. *Journal of Child Language*, pages 1–22.

Ashkenazi, O., Gillis, S., and Ravid, D. (2015). Input-output relations in the early acquisition of hebrew verbs. *Unpublished doctoral dissertation, Tel Aviv University.*

Aslin, R. N., Saffran, J. R., and Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological science*, 9(4):321–324.

Association, I. P., Staff, I. P. A., et al. (1999). *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge University Press.

Ben-David, A. and Bat-El, O. (2016). Paths and stages in acquisition of the phonological word in hebrew. *Acquisition and development of Hebrew: From infancy to adolescence*, 19:39–68.

Berman, R. A. (1990). On acquiring an (s) vo language: Subjectless sentences in children's hebrew. *Linguistics*.

Berman, R. A. (1996). Form and function in developing linguistic and narrative abilities: The case of 'and'. *Social interaction, social context, and language: Essays in honor of Susan Ervin-Tripp*, page 343.

Berman, R. A. (1997). Theory and research in the acquisition of hebrew as a first language. *Studies in the psychology of language*, pages 37–69.

Berman, R. A. (2004). Between emergence and mastery. *Language development across childhood and adolescence*, pages 9–34.

Bloom, P. (2002). *How children learn the meanings of words*. MIT press.

Brown, R. (1973). *A first language: The early stages*. George Allen & Unwin Ltd.

Chomsky, N. et al. (2006). On cognitive structures and their development: A reply to piaget. *Philosophy of mind: Classical problems/contemporary issues*, 751.

Cristia, A., Dupoux, E., Gurven, M., and Stieglitz, J. (2019). Child-directed speech is infrequent in a forager-farmer population: A time allocation study. *Child development*, 90(3):759–773.

Dąbrowska, E., Rowland, C., and Theakston, A. (2009). The acquisition of questions with long-distance dependencies.

De Villiers, J. G. and De Villiers, P. A. (1973). A cross-sectional study of the acquisition of grammatical morphemes in child speech. *Journal of psycholinguistic research*, 2(3):267–278.

Dromi, E. and Berman, R. A. (1982). A morphemic measure of early language development: data from modern hebrew. *Journal of Child Language*, 9(2):403–424.

Forgy, E. W. (1965). Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *biometrics*, 21:768–769.

Gretz, S., Itai, A., MacWhinney, B., Nir, B., and Wintner, S. (2015). Parsing hebrew childes transcripts. *Language Resources and Evaluation*, 49(1):107–145.

Gretz, S., Itai, A., and Wintner, S. (2013). *Syntactic Annotation of the Hebrew CHILDES Corpora*. PhD thesis, Computer Science Department, Technion.

Hiller, S. and Fernández, R. (2016). A data-driven investigation of corrective feedback on subject omission errors in first language acquisition. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 105–114.

Hollich, G. J., Hirsh-Pasek, K., and Golinkoff, R. M. (2000). *Breaking the Language Barrier: An Emergentist Coalition Model for the Origins of Word Learning*. University of Chicago Press.

Irvin, J., Spokoyny, D., and del Prado Martin, F. M. (2016). Dynamical systems modeling of the child-mother dyad: Causality between child-directed language complexity and language development. In *CogSci*.

Kalev, D. (2017). *Modern Times: New Aspectual and Modal Constructions in Contemporary Hebrew*. PhD thesis, PhD thesis. Tel-Aviv University [in Hebrew], Tel-Aviv.

Kaplan, D. (1983). *Order of Acquisition of Morph-Syntactic Categories among Hebrew Speaking 2 to 3 year olds*. PhD thesis, Master thesis. Tel-Aviv University [in Hebrew].

Lloyd, S. (1982). Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137.

Lustigman, L. (2015). Non-finiteness in early hebrew verbs. In *The Acquisition of Hebrew Phonology and Morphology*, pages 211–229. Brill.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA.

MacWhinney, B. (2000). The childes project: Tools for analyzing talk: Volume i: Transcription format and programs, volume ii: The database.

Maye, J., Werker, J. F., and Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, 82(3):B101–B111.

McMurray, B. and Hollich, G. (2009). Core computational principles of language acquisition: Can statistical learning do the job? introduction to special section. *Developmental Science*, 12(3):365–368.

Mintz, T. H. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition*, 90(1):91–117.

Mintz, T. H. (2006). Finding the verbs: Distributional cues to categories available to young learners. *Action meets word: How children learn verbs*, 31:63.

Moerk, E. L. (1980). Relationships between parental input frequencies and children's language acquisition: A reanalysis of brown's data. *Journal of child language*, 7(1):105–118.

More, A., Seker, A., Basmova, V., and Tsarfaty, R. (2019). Joint transition-based models for morpho-syntactic parsing: Parsing strategies for mrls and a case study from modern hebrew. *Transactions of the Association for Computational Linguistics*, 7:33–48.

Nelson, K. (2009). *Young minds in social worlds: Experience, meaning, and memory.* Harvard University Press.

Newport, E., Gleitman, H., and Gleitman, L. (1977). Mother, i'd rather do it myself: Some effects and non-effects of maternal speech style. *Talking to children: Language input and interaction*, pages 109–150.

Nivre, J., De Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., and Silveira, N. (2016). Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666.

Pelucchi, B., Hay, J. F., and Saffran, J. R. (2009). Learning in reverse: Eight-month-old infants track backward transitional probabilities. *Cognition*, 113(2):244–247.

Phillips, J. R. (1970). *Formal characteristics of speech which mothers address to their young children.* The Johns Hopkins University.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., and Manning, C. D. (2020). Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Compu-*

tational Linguistics: System Demonstrations.

Ravid, D., Ashkenazi, O., Levie, R., Zadok, G. B., Grunwald, T., Bratslavsky, R., and Gillis, S. (2016). Foundations of the early root category. *Acquisition and development of Hebrew: From infancy to adolescence*, 19:95–134.

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

Remick, H. (1971). The maternal environment of language acquisition. *Unpublished doctoral dissertation) University of California, Davis.*

Saffran, J. (2009). Acquiring grammatical patterns. *Infant pathways to language*, pages 31–48.

Seker, A., Bandel, E., Bareket, D., Brusilovsky, I., Greenfeld, R. S., and Tsarfaty, R. (2021). Alephbert: A hebrew large pre-trained language model to start-off your hebrew nlp application with. *arXiv preprint arXiv:2104.04052.*

Snow, C. E. (1972). Mothers' speech to children learning language. *Child development*, pages 549–565.

Straka, M. and Straková, J. (2017). Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99.

Szubert, I., Abend, O., Schneider, N., Gibbon, S., Goldwater, S., and Steedman, M. (2021). Cross-linguistically consistent semantic and syntactic annotation of child-directed speech.

Tal, S., Grossman, E., and Arnon, I. (2021). Infant-directed speech becomes less redundant as infants grow: implications for language learning.

Tatsumi, T., Ambridge, B., and Pine, J. M. (2018). Disentangling effects of input frequency and morphophonological complexity on children's acquisition of verb inflection: An elicited production study of japanese. *Cognitive Science*, 42:555–577.

Tomasello, M. (1992). *First verbs: A case study of early grammatical development.* Cambridge University Press.

Uziel-Karl, S. (2001). *A multidimensional perspective on the acquisition of verb argument structure*. PhD thesis, Tel-Aviv University Israel.

You, G., Bickel, B., Daum, M. M., and Stoll, S. (2021). Child-directed speech is optimized for syntax-free semantic inference. *Scientific Reports*, 11(1):1–11.

תקציר

מחקרים בגישה האמפיריציסטית לרכישת שפה נוטים להתמקד בהשפעת התשומה ההורית על הפקת הדיבור אצל הילד, וכיצד אפשר לראות בתשומה הזו איזשהו ייצוג לשלב ההתפתחותי של הילד. מחקר זה בוחן את התשומה ההורית מהכיוון ההפוך, כלומר בהסתכלות על ההשפעה של ההפקה הילדית על תשומת ההורה.

בפועל, מחקר זה מתחיל בהצעת שלבים התפתחותיים בקורפוס הדיבור המבוגר, בעזרת מדד שבאמצעותו נקבעים שלבים התפתחותיים אצל ילדים. לאחר מכן, נבדקת ההתפלגות של אלמנטים מורפו-תחביריים אצל המבוגר בשלבים השונים, ביחס לגיל הרכישה של האלמנטים האלה אצל ילדים. התוצאות מראות כי השלבים ההתפתחותיים המוצעים עבור דיבור המבוגר, ובפרט ההתפלגות של מופעי האלמנטים בכל שלב, רגישים מאוד לגיל הרכישה של אלמנטים אלו אצל הילד, ואף יכולים לשמש סמן-מקדים לאילו אלמנטים מורפו-תחביריים יירכשו (ומתי, אם בכלל). מחקר זה גם מציע הסבר מבוסס-שימוש למספר המינימלי של מופעים של אלמנט בדיבור המבוגר שנדרש כדי שהאלמנט יירכש ע״י הילד.

אוניברסיטת תל-אביב

הפקולטה למדעי הרוח ע״ש לסטר וסאלי אנטין

החוג לבלשנות

מעשה אבות סימן לבנים, ולהיפך –
מציאת שלבים התפתחותיים בדיבור המבוגר אל הילד

חיבור זה מוגש כעבודת גמר לקראת תואר

״מוסמך אוניברסיטה״ (MA) באוניברסיטת ת״א

על-ידי

סתיו קליין

בהנחיית

פרופ׳ רות ברמן

אוקטובר 2021