Corrective Feedback in First Language Acquisition

MSc Thesis *(Afstudeerscriptie)*

written by

Sarah Hiller
(born July 21st, 1988 in Giessen, Germany)

under the supervision of **Dr Raquel Fernández**, and submitted to the Board of Examiners in partial fulfillment of the requirements for the degree of

MSc in Logic

at the *Universiteit van Amsterdam*.

Date of the public defense:	**Members of the Thesis Committee:**
January 15, 2016	Dr Raquel Fernández
	Dr Stella Frank
	Dr Floris Roelofsen (chair)
	Dr Willem Zuidema

Institute for Logic, Language and Computation

Declaration of Originality

I declare that the text and the work presented in this document is original and that no sources other than those mentioned in the text and its references have been used in creating it.

Sarah Hiller

Abstract

Children learn their first language in interaction with proficient users. This naturally exposes them to *positive input*, i.e. grammatically correct utterances in context. It is still unclear however, whether they also receive *negative input*. That is, responses informing them about the inadequacy of grammatically erroneous utterances.

In the present study we investigate parental reformulations, or *corrective feedback*, as a possible candidate for a conversational pattern conveying this information. Reformulations occur as a response to a wide variety of child errors. They indicate an error while simultaneously presenting its corrected form. We investigate whether these types of responses are indeed helpful for language acquisition.

To this end, a large scale empirical analysis is employed. All relevant transcripts available from the part of CHILDES database in English language are used (MacWhinney, 2000a). Candidate child-adult utterance pairs in a subset of files are manually annotated for the presence of corrective feedback and for the corrected errors. These manually annotated exchanges serve as the training set for an automatic classifier aimed at distinguishing corrective feedback from non-corrective feedback instances. The predictive accuracy scores, however, show that the phenomenon is too diverse to be captured globally with our approach. Hence the investigated phenomenon is restrained to responses following a *subject omission error* in the child utterance. We develop automatic extraction methods for both child utterances containing this error and responses correcting it in a reformulation.

The effect of corrective feedback on language learning is investigated by testing whether a higher rate of corrective feedback coincides with a greater decrease of the amount of error made at a later moment compared to at the starting age. A correlation analysis gives a first pointer in this direction. A subsequent linear regression analysis confirms that corrective feedback increases explanatory force of the model beyond what other features achieve, after a lag of at least 9 months between start and end age, with a peak after a difference in time of around 14 months.

Acknowledgements

First and foremost I want to thank my supervisor, Raquel Fernández. I learned a lot from working with you. It was always a great help and inspiration to go to one of our frequent meetings. Thank you for this, as well as for your patience with me.

Next, I would like to thank the members of my thesis committee, Floris Roelofsen, Jelle Zuidema and Stella Frank, for the valuable remarks made during my defense which were included into this revision or were helpful indicators for future work.

Thanks also go out to my friends here in Amsterdam and elsewhere, you know who you are, who have helped with this by studying together, distracting me to get my head free, making sure I had enough coffee and vegetable intake, and accepting my instable mental state throughout the final stage of writing.

And finally, I want to thank my family for the continuous unquestioned support, both spiritual and financial. I would not have been able to get this far without you, and it means a lot to me.

Für Ute und Christoph.

Contents

Chapter 1

Introduction

1.1 Motivation

In the present study we will be concerned with investigating parental reformulations as candidates for negative input during language acquisition. Children learn language in interaction with proficient users around them. There is no doubt that they are exposed to *positive input*: grammatically correct utterances in context. It is still an open question, however, whether they also receive *negative input*. That is, whether they are informed about the inadequacy of grammatically erroneous utterances. Several theoretical considerations indicate that this information should be available (for example Gold (1967) and Saxton (2010)). The most intuitively convincing argument in this respect is the fact that despite making errors at a certain stage during the learning process, children will proceed to the use of a correct grammar. In the learning process they might apply the regular past tense construction to irregular verbs, such as *go*, to give the false past tense *goed*. But they also accept the correct adult version *went*. Without any information that *goed* is incorrect, even with an abundance of evidence showing the correctness of *went*, how do they understand that only the latter form is indeed applicable?

Brown and Hanlon (1970) showed that a generally common pattern of negative input in other domains, explicit disapprovals, is not used to inform children about grammatical mistakes. Does this mean that no negative feedback is available? Another pattern has been suggested as a candidate for transmitting this information: reformulations (for example Chouinard and Clark (2003) or Sax-

ton et al. (2005)). Parents pick up their children's erroneous utterances in the following turn and present the corrected form, such as in the following example taken from 2-year old Lara in the CHILDES database (Rowland and Fletcher, 2006; MacWhinney, 2000a).

(1) *CHI: I climb up daddy .
*DAD: you did climb over daddy .

The fathers response is highly affirmative, but also presents the appropriate preposition to be used in this case. To the adult observer this clearly looks like an implicit correction. In addition to informing about the presence of an error, it also gives the precise location and the correct form. But is it picked up by the child as such?

Analysing reformulations as a candidate for negative input during language acquisition is not a new idea. The goal of the current investigation is to extend previous work by analysing comparably large amounts of data. In particular, the main contributions are the development and subsequent application of methods for investigating this phenomenon empirically on a large scale.

1.2 Research Questions

We have presented why *corrective feedback* seems like a promising candidate for a pattern that provides negative input concerning grammaticality to the language learning child. But the fact that an adult observer is able to infer a correction from an exchange such as the one presented in Example (1) does not immediately imply that language learning children perform the same inference. We will therefore analyse the effect that corrective feedback has. Is it actually taken up in the way one would think: do children use the provided information as a correction (possibly unconsciously)? Do reformulations help them in learning a language and retreating from error?

We are interested in investigating these questions empirically, using comparably large amounts of data for a linguistic inquiry. Thus we want to develop methods which enable this. First of all, we have to ask what precisely constitutes corrective feedback and how we can subdivide this complex phenomenon

in a meaningful way. After that, we want to establish which information needs to be contained in or possibly added to the available data to enable us to draw relevant conclusions. Next we ask which features in the data are representative of the investigated phenomenon; how can we manage to automatically classify an exchange as containing corrective feedback, given the information present in the transcripts? And finally, using all the answers to the previous questions, we want to find out whether corrective feedback is indeed instructive for language learning children.

1.3 Outlook

The thesis is organised as follows. In Chapter 2 we will present the theoretical debate of nature versus nurture as the framework into which the current research is embedded. One specific argument in favor of the nativist viewpoint is the *no negative feedback hypothesis*. Despite negative input being necessary for language acquisition, children do not receive it in form of explicit disapprovals. We will investigate another response type which was proposed as a means for communicating negative input: corrective feedback. This pattern will be defined in detail in Chapter 3. Additionally, a taxonomy of the phenomenon together with observable examples will be presented. Next, as we are concerned with a large-scale data driven analysis, the data selection will be specified in Chapter 4. The selected data will then be endowed with further necessary information, such as structural analyses but also manual annotation informing about corrective feedback. In Chapter 5 we will describe the attempted procedure to find an automatic extraction algorithm for general instances of corrective feedback. This proves to be infeasible however, most likely due to the diverse nature of the phenomenon and complex interplay between utterances at hand. Thus the object of investigation is limited to a subclass of instances of corrective feedback: as a response to subject omission errors, the most common kind of child error. Automatic extraction algorithms for the error as well as for reformulations correcting this error will be devised in Chapter 5. In Chapter 6 the effect of corrective feedback after subject omission errors on learning the correct inclusion of subjects will be investigated. A correlation analysis of the amount of corrective feedback given at a certain time and the improvement in frequency of

subject omission errors until a later time shows a positive relation, most strongly after a lag of about one year. This finding is confirmed in a linear regression analysis. Additionally, we show that corrective feedback increases predictive strength of the linear regression model above what can be achieved by other predictors, after a difference between start and end age of at least 9 months. Finally, Chapter 7 will present a summary of our findings as well as an outlook concerning possible future work in this direction.

Chapter 2

Theoretical Background

In the current chapter we will present the theoretical debate in which our current investigation is embedded. It is widely agreed that children learn their first language in interaction with proficient users (e.g. Saxton (2010) or Chouinard and Clark (2003)). Discussion takes place, however, concerning the precise nature of this learning mechanism. More specifically, one point of disagreement is in how far language is a specific quality that differs from other cognitive abilities. According to the nativist point of view the necessary foundations for language acquisition are specifically rooted in our cognitive framework; according to the behaviourist point of view the same general cognitive principles are employed for learning language as they are for other learning mechanisms. In Section 2.1 we will first of all present these two viewpoints in some more detail. One important argument in favor of the nativist viewpoint, the *poverty of the stimulus hypothesis*, is discussed in Section 2.2. An empirically verifiable refinement of this argument, the *no negative feedback hypothesis*, is shown in Section 2.3. It is this last refinement which we will be concerned with in the current investigation. Finally, in Section 2.4 we summarise the main points of the chapter.

2.1 Ideas from the Nature – Nurture Debate

The debate taking place concerning how language acquisition is rooted in the cognitive framework can be summarised as the differentiation of nature versus nurture. According to the nativist point of view (the *nature* part of the division), a *Universal Grammar* is innate. That is, an important structural

part of the grammar to be learned is present at birth (for example Chomsky (1965)). Non-nativists (the *nurture* side of the division) counter this idea of an innate structure. Saxton (2010), on the non-nativist side, phrases the division as follows:

> Nativists argue that the grammar acquiring capacity is dedicated exclusively to language. Non-nativists, on the other hand, consider grammar to be simply one of many mental achievements acquired by general-purpose cognitive learning mechanisms.

In the following we will have a closer look at the precise nature of the *Universal Grammar* which is under discussion here. First of all, the observed wide variety in grammar between and within languages must fit within the framework. Hence the *Universal Grammar* is not presumed to contain specific grammatical rules. Rather, it is taken to contain broad features. Parameters for these are set according to exposition, which consequently limits the space for possible exact rules. One example would be the recognition of the parameter *head first* (as opposed to *head last*) when exposed to English.

At first this idea might seem rather obscure: how do children recognise which parameter setting is triggered by a certain input? An interesting experiment is worth mentioning in this respect (see the introduction for Chomsky (1980) in Piattelli-Palmarini (1980)). Newborn kittens were put into a visual environment such that they would see only horizontal or only vertical lines. This lead to the development of some and the degeneration of other neurons, depending on the orientation of the lines. Granted, the described experiment concerns the cognitive development of kittens under varying visual input, but clearly children cannot be investigated in a similar form and no other animal uses language. The important information to take from this exposition is that it might make the idea of certain structural elements being hard-wired into the brain and triggered by exposure more intuitively convincing.

Nativism is mostly motivated with reference to the contrast between the perceived speed and ease with which children learn complex grammatical rules as opposed to the sparsity of the input (see for example Saxton (2010) for a discussion). The second part of this argument will be examined more closely in Section 2.2.

2.2 The Poverty of the Stimulus Hypothesis

An important argument in favor of the nativist viewpoint is the *poverty of the stimulus hypothesis*. Namely, the quality of the parental input is simply seen as too poor to enable acquisition of the correct grammar for the child, unless an important structural part of the grammar is already present before the learning begins. This argument does seem intuitively convincing if we look at the many disfluencies, repetitions, non-sentential utterances, etc. which can be observed in spoken language. Compared to the actual rules for grammaticality of sentences such utterances are largely substandard. For example, Maclay and Osgood (1959) quantified the repeats and false starts observable in talks given at a linguistics conference. They found these irregularities to be ubiquitous. Now one should think that linguists do not form a subgroup of the population which makes particularly erroneous use of language. Hence the way a common child's caregivers talk can reasonably be assumed to be just as ungrammatical. Clearly, if external input contains insufficient information to achieve the proficiency later on exhibited by the children, then the necessary structural framework for language acquisition must be innate. For example Chomsky (1965), as the most prominent proponent of the naturalist approach, uses this line of argument:

> A consideration of the character of the grammar that is acquired, the degenerate quality and narrowly limited extent of the available data, the striking uniformity of the resulting grammars, and their independence of intelligence, motivation, and emotional state, over wide ranges of variation, leave little hope that much of the structure of the language can be learned by an organism initially uninformed as to its general character.

To justify a behaviourist viewpoint these considerations clearly need to be countered. One notable argument in this direction is that the input which language acquiring children receive is very different from what is observable when several proficient language users converse with each other. Saxton (2010) presents a summary of studies concerning a certain way of talking which he calls *Child Directed Speech* or CDS in short. Speech directed to children is different from speech directed to adults. While it does not matter for our purposes why this

difference occurs - whether parents intend to simplify their speech for teaching purposes or whether they unconsciously do so because the child reacts better to this form of input - the first important statement is that the following observations apply across languages, cultures and social status.

CDS is way more restricted than speech between adults. Clearly the topics of discussion relevant to the child are not as diverse as they can be for adults. More importantly for the current case, also the phonology, morphology and syntax are simplified. To be precise, CDS exhibits exaggerated intonation, less disfluencies and is slower than adult directed speech. In morphologically complex languages, such as Russian, it contains a simplified morphology which is complexified according to the level of the language learner. Syntactically, sentences are very short and - surprisingly - well formed. That is, as long as correct modules such as noun phrases are counted as syntactically well formed. Lastly, as a connection to semantics, the subject of a sentence is even more predominantly also the agent in CDS than in adult directed speech.[1]

Overall this synopsis gives a good impression of the ways that speech directed to language acquiring children is at the same time structurally easier and grammatically more correct than speech between adult interlocutors. Hence it may be that the input which children receive is not quite as degenerate as one might think when listening to linguists giving talks.

In addition to the discussed quality of the input, children also simply receive a very high quantity of input. Moerk (1983) counted the number of different syntactical constructions in one hour time slots of the Adam and Eve files in the Brown corpus (Brown and Hanlon, 1970) to estimate the number of times children are presented with these constructions over the course of a year. Overall, he counted an average of around 260 adult utterances per hour of interaction. With an average length of utterances between 3 and 5 words – averaged between both parents and children, making the number for only adults even higher – this gives at least 800 to 1300 words of input for the child per hour. As for the frequency of syntactic constructions, he for example counted an average of 135 S-V-(O) sentences per hour, which he extends to a cautious estimate of about 40,000 per month. He came to the conclusion that, taking into account the lim-

[1] See the references provided in Saxton (2010) for more detail.

ited semantic variation, children will extremely often be presented with similar sentences subject to different minor variations. He claims that this makes it very easy for them to extract the functions of the variable parts of those sentences:

> That many phrases will thereby be repeated, with just a sufficient number of minor alterations to make sentence constituents and sentence structure more obvious, can confidently be concluded.

This constitutes another argument against the poverty of the stimulus hypothesis.

2.3 The No Negative Feedback Hypothesis

As convincing as either of the above considerations may be, they do give rather speculative arguments on the (in)sufficiency of parental input for the child's language acquisition. A refinement of the *poverty of the stimulus* hypothesis gives a more solid indication in this respect. Given certain preliminary assumptions a clearly specified type of input is proven to be essential for language acquisition in the absence of an innate structural grammar. These proofs will be discussed in Section 2.3.1. Consequently, as the type of input discussed is so restricted it can be investigated whether it actually does occur. First examinations seemed to show that it does not. We will present these in Section 2.3.2. Finally then, if this type of input is necessary but unavailable to the child, it can be concluded that the obtained stimulus alone is indeed too poor to enable unsupported language acquisition. This refinement is called the *no negative feedback hypothesis*, after the type of input under investigation.

Now, what is negative feedback? When learning a language children first of all receive *positive input*, that is, any kind of correct utterance directed to them or overheard by them. *Negative input* on the contrary reveals the inadequacy of a certain utterance. The term *negative feedback* was chosen to indicate negative input which refers back to a previously uttered child statement.[2] Negative

[2]Note that this differentiation is only of practical interest. Theoretically, the same information can be extracted by a language learner from an informant who lists all possible sentences and specifies whether they are correct or incorrect as from an informant who responds to inquiries whether sentences are grammatical or not. To get from the first to the second case,

feedback can be presented in many different ways, such as through explicit disapproval but also for example through clarification questions. It is not always evident from this type of feedback which part of the utterance - grammar, pronunciation, pragmatical implication, etc. - was inadequate. But now let us have a look at the proofs for the necessity of grammatical negative input during language acquisition.

2.3.1 Negative input is crucial

It has been argued from different perspectives that positive input alone is not sufficient for learning a language and that negative input is crucial. We will now present two justifications for this claim, the first of which is rather intuitive whereas the second constitutes a formal proof.

Before mastering a language children go through a phase where they use it incorrectly. They consequently retreat from these errors and continue to use the correct adult grammar. Saxton et al. (2005) focus on this development. They claim that during the erroneous phase the child's grammar is a superset of the correct adult one. While the adult grammar allows for only one form, the correct one, the child grammar allows for at least two different forms, the correct and the erroneous one. We can illustrate this with the incorrect application of the regular plural construction to irregular nouns, such as *man*. Whilst learning the child might admit both *mans* and *men*, or even *mens*. At a later stage only *men* is accepted. Now the crucial argument is that at the intermediate stage further presentations of correct sentences cannot shrink the child's grammar. As Gold (1967) puts it:

> The problem with text [i.e. positive input alone] is that, if you guess too large a language, the text will never tell you that you are wrong.

Therefore negative feedback is required.

A formal proof showing the necessity of negative input for a formalised notion of language learning was presented by Gold (1967). A brief summary of

if the learner wants to get the response whether a sentence is grammatical she simply needs to wait until it occurs in the list. And to get from the second to the first case, the learner can follow a predescribed enumeration of all possible sentences and query them one after the other.(cf. Theorem I.3 in Gold (1967))

the ideas in this proof is given here, more detail is presented in Appendix A. Gold considers a model of language learnability for Turing machines. This model applies to classes of languages. A class of languages is regarded as learnable if it is *identifiable in the limit*. That is, if given any set of non-repetitive input derived from a language in the class, an algorithm can be devised, which recognises the corresponding language after a finite number of steps. Two kinds of input are considered: solely positive instances, or positive and negative information. Gold shows that the class taken to represent natural languages cannot be learned from positive input alone. Hence either the search space must be restricted to a smaller class - by a *Universal Grammar* - or negative input is necessary. To extend this to natural language acquisition we need to assume that the presented formalised notion of learning is a good representation.

2.3.2 No explicit negative feedback is given

Now that the proofs for the necessity of negative input for language acquisition have been presented we can go on to look at the empirical findings supposed to show its unavailability. One main study was taken to supply this evidence. Brown and Hanlon (1970) investigated why it is that children go from an ungrammatical use of certain constructions to a grammatical use of them.[3] They searched for positive or negative reinforcers in the adult input. The definition of a positive reinforcer is somewhat circularly based on its effect. Thus with the observable improvement of child grammaticality it is impossible to exclude that *some* sort of reinforcer was involved in this development. To be able to still make an empirical investigation they reverted to looking at a behaviour which is well known to serve as a general reinforcer: approval. With regard to this specific type of reinforcement, they find that:

> In neither case [differentiating only correctness/incorrecteness of the preceding utterance or also distinguishing degrees of incorrectness] is there even a shred of evidence that approval and disapproval are contingent on syntactic correctness. [...] They are rather linked to

[3] Actually the main topic of their research was a different one, this question solely emerged as a side problem. However, it is these findings we are interested in here.

the truth value of the proposition, which the adult fits to the child's generally incomplete and often deformed sentence.

For long years this was taken as substantial evidence that children do not receive any negative input for their grammatical errors (see Chouinard and Clark (2003) or Saxton (2010) for a discussion).

2.4 Summary of the Chapter

In the present chapter we have presented the theoretical debate in light of which our current investigation is placed. In Section 2.1 a brief introduction into the nature - nurture controversy concerning language learning was given. Nativists view a certain part of the grammar as innate, features of grammatical rules are learned through triggers which set parameters in the native *Universal Grammar*. Contrastingly, behaviourists consider language learning to be accomplished by the same cognitive processes by which general learning about the surrounding world takes place. In Section 2.2 we went into some more detail concerning one argument presented by nativists: that the input presented to the language learning child is too impoverished to allow for the observed acquisition of grammar without a certain innate structure. A refinement of this argument is the *no negative feedback hypothesis*, presented in Section 2.3. Several considerations show that negative input is necessary for language learning in the absence of an innate structural grammar. However, this negative input is not given in the form of explicit parental disapproval successive to a child's grammatical error. It seems as though we are facing a paradox: while negative input is proven to be necessary for language acquisition, children do not receive it. But they *do* learn language. How is this possible? In Chapter 3 we will first of all discuss several ways of overcoming this paradox and finally select one of the given ideas for detailed investigation.

Chapter 3

Corrective Feedback

In Chapter 2 we saw that information about the ungrammaticality of utterances is necessary for language acquisition, given certain preliminary assumptions. However, this information is not given in the form of explicit parental disapproval. This can lead to several conclusions.

One possible inference is that unassisted language learning is indeed impossible and therefore the structure of the grammar to be learned must be already present in the child's brain before learning starts. This would be the line of argument taken by naturalists. While it seems like a very elegant way out of the paradox, when looking closer it becomes clear that assuming an innate grammatical structure does not actually solve the presented issues. Negative input is needed in language acquisition for the retreat from error, for *unlearning*, as Saxton et al. (2005) name it. It is without doubt that children do make errors at certain stages before they improve their grammar and finally apply correct rules. But it is precisely in this respect that innateness of the structure of the grammar to be learned does not provide sufficient justification for the learning. Either the wrong rules are possible according to the soft bounds of the universal grammar, thus enabling the child to pass through the erroneous phase. Or the wrong rules are not possible according to the universal grammar, and it is this fact that is responsible for the child's retreat from the erroneous phase. It is not immediately clear how both could be possible together.

Luckily however, postulating an innate structural grammar is not the only way out of the paradox. We will present a short survey of other suggestions in Section 3.1, before continuing to investigate one particular proposition: that

children receive negative input in the form of parental reformulations. This idea has been explored before (Saxton, 2010; Chouinard and Clark, 2003), the novelty in the current analysis is that it is based on comparably large amounts of data. In Section 3.2 we specify the features of the phenomenon under investigation, before differentiating it into subclasses and giving examples in Section 3.3. In Section 3.4 we present a survey of two different accounts that explain why and how corrective reformulations might aid in language acquisition. Finally, in Section 3.5 we briefly summarise the chapter.

3.1 Avoiding the Negative Input Paradox

Recall the paradox we are discussing: we saw in section 2.3 that negative input is necessary for language acquisition, but not available in form of explicit parental disapproval. This *necessity* was justified using two different proofs. Hence, there are at least as many ways out of the paradox as there are preconditions in these proofs.

Scholz (2004), for example, discusses some ways in which the apparent contradiction due to Gold's proof - see Section 2.3.1 and Appendix A - can be avoided by disabling its premises. Crucially, Gold's proof applies to a formalised notion of learnability, identification in the limit, and learning takes place via hypothesis formation and testing. Also, what is learned is a complete generative grammar for the language: a tester for the language, defined via a turing machine. It can be contradicted for all of these points that they apply to children learning their first natural language.[1]

Next, also Saxton's argument does not need to be universally accepted: we could contradict his assumption that the child's erroneous grammar is a superset of the adult grammar and needs to be reduced via negative input. Some other representation could be imagined to overcome this point.

And finally, we can also leave the presented proofs and their assumptions in place and instead counter the consideration that no negative input is available to

[1] Contradicting that the class of natural languages is a superset of the class of languages containing all finite languages and at least one infinite one constitutes a limitation of the search space, which precisely boils down to postulating an innate limitation of the possible grammar.

the language learning child. This is the path we will take here. Recall that what Brown and Hanlon (1970) showed was simply that *explicit* parental approval and disapproval are not contingent on grammaticality of the preceding child utterance. But negative input could be presented differently, and possibly give even more information than simply the statement that the preceding utterance was inappropriate.

It has been widely observed that parents pick up their children's erroneous utterances in a following statement and repeat them correctly (e.g. Saxton et al. (2005)), such as in the following example from 2-year-old Lara in the CHILDES database (Rowland and Fletcher, 2006; MacWhinney, 2000a).

(1) *CHI: I climb up daddy .
*DAD: you did climb over daddy .

The father's response to the child's grammatically spurious utterance is highly affirmative, but at the same time presents the corrected form of the previous statement. To an adult observer this seems like a correction, with the information about ungrammaticality implicitly embedded in the structure of the dialogue. Additionally, this type of feedback seems to also give the child the necessary information on how to correctly phrase the preceding sentence. Brown and Hanlon (1970) themselves already noticed that:

> Repeats of ill-formed utterances usually contained corrections and so could be instructive.

This remark has unfortunately long been lost in the shadow of their other findings.

It is this type of recast that will be considered as a candidate for negative feedback on child grammaticality in the following investigation. We name this schema *corrective feedback* to stress the points that first of all, the parental recast refers back to a previous child utterance, hence *feedback*, and second, an appropriate form of the corresponding utterance is presented, the statement is *corrective*. Intuitively, it does look like these forms could be picked up as corrections and thus help the child in acquiring the language. Chouinard and Clark (2003) investigate children's immediate responses to reformulations and find that acknowledgements and repeats are very frequent. This indicates that

children attend to their parent's corrections. Saxton et al. (2005) show that reformulations have a positive effect on the correct use of the corresponding grammatical structure 12 weeks later. We will extend this work by investigating the same phenomenon using much more data.

3.2 Definition of Corrective Feedback

For the following analysis we will need a precise definition of the phenomenon we are investigating: *corrective feedback*.

A common reaction to a child's grammatical error is for the adult interlocutor to pick up the sentence and present a correction of the erroneous form embedded into the next utterance. The following exchange, from 2-year-old Trevor in the Demetras corpus in CHILDES Demetras (1989), is an example for this:

(2) *CHI: I waked evwybody [: everybody] up .
*FAT: you woke everybody up .

The following general properties of corrective feedback are all visible in this example.

Definition 1 (Corrective Feedback) *1. The child makes a mistake. There is a basis for possible negative feedback.*

2. The words in the parent and child utterance overlap. The correction is anchored to the erroneous form through at least one exactly matching word.

3. The adult utterance is different from the child utterance. There must be space for an adjustment.

4. This alteration constitutes a correction. That is, the parent's utterance contrasts with the child's one by presenting a change from an erroneous to a grammatical form.

These four properties identify corrective feedback. What is not discussed in this definition are any deeper implications of the structure of the described exchange, such as intention of the adult, reception by the child etc. This exchange pattern is boldly named corrective feedback, because it superficially does look

like negative feedback in form of a correction. The implications concerning the intentions of the caregivers and the reception by the child invoked through the use of this terminology still need to be tested.

3.3 Types of Corrective Feedback

This section presents a taxonomy of types of corrective feedback. Several implementations are possible for a more finegrained classification of child-adult exchanges containing corrective feedback. Mainly, the exchanges can either be discriminated via the kind of error in the child utterance, or via the kind of correction employed by the adult. Which of these to choose clearly depends on the relations we want to be able to deduce from instances of the investigated phenomenon. For example, Sokolov (1993) investigates the fine-tuning hypothesis and differentiates child-adult exchanges via the way in which the parent utterance diverges from the child utterance. As the focus of his investigation lies in the way parents react to their children's utterances this division is appropriate. However, as Chouinard and Clark (2003) point out, classifying exchanges according to the parental reply-type can lead to corrections of errors and non-corrective continuations of the conversation being in the same class. For example, adult expansions of a child utterance can be corrective, but they can also simply progress to a new topic. Chouinard and Clark (2003) and Saxton et al. (2005) look at the effect of parental responses on language learning. They partition exchanges according to the type of error observable in the child utterance. Using this division allows for subsequent testing of whether a certain response to a given error influences the child's comprehension of this structure as erroneous. The focus of the current analysis thus suggests the second type of partitioning.

There are two degrees of distinguishing features in children's grammatical errors. First of all one can differentiate the linguistic level at which the error occurs. It can for example be located at the subject of a sentence, or at the irregular past form of a verb. The next degree of differentiation is the type of error observed at this level. Very often with children, this will be omission. This distinction is important here, and we will continue using the expressions *level* and *type* in precisely this meaning.

The first decision to be taken to obtain a taxonomy of corrective feedback instances concerns the set of linguistic error*levels* which will be explicitly distinguished. Note that this will only be a selection of all imaginable possibilities. Several attributes play into this selection, such as observability in a few sample conversations, feasibility of automatic extraction, etc. To have a starting point, we will have a look at the literature mentioned above.

Chouinard and Clark (2003) classify the location of children's errors into four different categories: phonology, morphology, lexicon and syntax. Depending on the way a conversation is transcribed, extracting phonological errors may be impossible. The CHAT transcription format used in the CHILDES database (MacWhinney, 2000a), which will serve as the basis for our empirical analysis, allows for transcription of mispronunciations with additional information specifying the intended word. This is represented as in the following example by 2-year old Trevor in the Demetras corpus (Demetras, 1989).

(3) *CHI: let's play dis [: this] .

However, this feature is not consistently employed in all transcripts. Therefore phonological errors had to be disregarded. A similar reasoning led to disregarding lexical errors: one can only expect to be able to automatically extract them when they do get corrected. This counters the idea of comparing how often an error is countered with corrective feedback to the improvement of the corresponding construction in the child's grammar. Overall, that leaves syntactical and morphological errors to be differentiated.

From a different perspective, Saxton et al. (2005) distinguish 13 more fine-grained grammatical functions of words at which errors can occur. Being grammatical these error locations all fall into the categories of morphology or syntax differentiated by Chouinard and Clark (2003). This overlaps with the above restrictions we made due to applicability considerations. Thus overall, combining the differentiation of Chouinard and Clark (2003) and Saxton et al. (2005) with practical considerations concerning the extractability of given error types we end up with the 13 error locations also specified by Saxton et al. (2005). They are all listed below, with one addition: *main verb*. This linguistic level does not occur in Saxton's list as a possible error location but does occur very often in

child speech, thus it was added here. It was merged with the category *copula omission* to prevent ambiguity.

Next we have to look at the possible *types* of errors. As Saxton et al. (2005) already show, most child errors are omissions. This should not be surprising. It would seem rather difficult to consistently construct sentences which contain, to keep it simple, a subject, main verb, determiner and object using expressions with an MLU of 2. But other kinds of error are also possible and observable: children can include the required structural element but use a wrong realization, such as a regular past tense form instead of the irregular one. We will label these *substitution* errors. Alternatively, children include words where none are necessary, which we label *addition* errors. Finally again, this list is not exhaustive and one can well imagine *other* kinds of errors. Especially, errors in word order do not fit anywhere in the above distinction, but do occur.

It is time to present our classification of errors with examples of corrective feedback taken from the corpora in CHILDES (MacWhinney, 2000a). For every error *location* one or two examples are given, with the *type* of error being specified separately. Any information originally transcribed in addition to the text - presence of an error, temporal overlap, etc. - was deleted for better readability.

(4) **Syntax**

a. **Subject - omission**
*CHI: can't get that out .
*DAD: you can get these out , look .

b. **Main Verb / copula - omission**
*CHI: what you doing round here ?
*MOT: what are you doing round here ?

c. **Object - omission**
*CHI: I'm squashing .
*MOT: you're squashing the poor squirrel ?

Noun morphology

a. **Possessive -'s - omission**
*CHI: hold Mummy fork .
*MOT: you want to hold Mummy's fork .

b. **Regular plural -s - addition**

*CHI: cars is driving home .

*MOT: the car will drive them home .

c. **Irregular plural - substitution**

*CHI: two foot came out .

*MOT: two feet .

Verb morphology

a. **3rd person singular -s - omission**

*CHI: machine squash it all the way home .

*MOT: what machine squashes it all the way home ?

b. **Regular past -ed - omission**

*CHI: I think I nick somebody .

*MOT: you think you've nicked somebody ?

c. **Irregular past - substitution**

*CHI: he falled out and bumped his head .

*MOT: he fell out and bumped his head .

Unbound morphology

a. **Determiner - omission**

*CHI: hat .

*MOT: a hat .

b. **Preposition**

(i) **omission**

*CHI: everybody sit down the train .

*MOT: everybody sit down on the train .

(ii) **substitution**

*CHI: I climb up daddy.

*DAD: you did climb over daddy.

c. **Auxiliary verb - omission**

*CHI: we done that one .

*MOT: we've done that one, haven't we ?

d. **Present progressive (auxiliary) - omission**

*CHI: looking the dustbin wagon .

*MOT: is she looking for the dustbin wagon ?

3.4 Ideas on Causes and Functioning

Discussion takes place concerning the issue of why parents reformulate their children's erroneous utterances and how these reformulations are supposed to make children recognise the error in their own utterance.

For Chouinard and Clark (2003) corrective feedback is given to inquire on the intended meaning of the preceding child utterance, which was obscured by the error. They refer to the Gricean Maxim of Manner (Grice, 1975), which requests cooperative speakers amongst other things to avoid ambiguity. With the interpretation of ungrammatical forms not necessarily being well-defined their occurrence sparks uncertainty in interpretation:

> But young children often use erroneous forms [...]. These violations of the Maxim of Manner can obscure children's meaning, so adults may need to check up on just what they intended to convey.

One way of achieving clarification is through repeating what one understood to be the meaning of the preceding utterance. In case of correct understanding this precisely constitutes corrective feedback, given that the parent employs the correct adult grammar. Through the contrast arising from the different ways of expressing the same *meaning* children are considered to be able to infer that their own utterance contained an error.

Saxton et al. (2005) criticise the idea that parents are incessantly unsure about what their children are trying to tell them. They state (highlights as in the original):

> A further problem with Chouinard & Clark's approach lies in their focus on parents as constant monitors of children's meaning. Undoubtedly, there are occasions when parents are not sure precisely what meaning their child intends to express. Arguably, however, for the vast majority of GRAMMATICAL errors, confusions of this kind are rare. For example, it is very doubtful that a parent would need to check up on the child's intended meaning when the latter says: *I drawed a lovely picture for you.* We would argue that the meaning of *drawed* is entirely transparent. It is only the linguistic FORM that the adult might take issue with.

Instead of presenting a different representation for the same meaning parents giving corrective feedback are taken to present a different form for the same *grammatical function*. Again, a notion of contrast is employed to explain why the child recognises its own utterance as erroneous when presented with a following correct adult sentence. However, as was described above, the contrast which Saxton et al. focus on lies in the different representations for the same grammatical function, instead of the same semantic function.

It is also implicitly clear that the corrective force of this feedback is viewed as intended by the parents. For example, they state that:

> In creating a contrast in usage between the two alternatives within the discourse, it is predicted that the ungrammaticality of [the erroneous form] is revealed to the child.

In general, different ideas have been proposed as to why corrective feedback can be observed in conversations between adults and children and how it is assumed to help the child in acquiring its first language and retreating from error. For the present question it is of no great importance which intention the caregiver follows when presenting this correction. As for the possible ways of extracting the negative input from a reformulation, both accounts agree that a contrast is created between the erroneous and the correct form. The two different modes of establishing this contrast will be kept in mind and investigated later on.

3.5 Summary of the Chapter

We started out this chapter by making explicit the paradox that arises from the *no negative input hypothesis*. Namely, language learning is achieved by children despite the supposed lack of necessary negative input. Several approaches for explaining this fact were discussed. One of these is the suggestion that children do receive negative input, in the form of parental reformulations of erroneous utterances. It is this concept which we are investigating here. We defined what constitutes a child-adult utterance pair containing such a reformulation, or *corrective feedback* as it is named here: an erroneous child utterance, partial overlap between the child and parent utterances, and a correction in the parent utter-

ance. Subsequently we gave a taxonomy of the observed phenomenon according to the level and types of errors observable in the child utterance and saw in examples that corrective feedback occurs for all of them. Finally we discussed ideas presented in the literature concerning how children extract the corrective force from reformulations. These consistently rely on a contrast established between the child and parent utterance. In Chapter 4 we will present the data used for the empirical investigation and develop the necessary preliminary processing.

Chapter 4

Data Selection and Annotation

Now that it is clear why corrective feedback in first language acquisition presents an interesting subject of investigation, how the features for identifying it are defined here, and what the specific questions are that will be investigated in this respect it is time to proceed to the empirical analysis. We are interested in an investigation of comparably large amounts of data, so all relevant transcripts from the CHILDES database (MacWhinney, 2000a) will be used. In the following Section 4.1 first of all the selection criteria for what will be considered *relevant* are developed. A certain set of additional information, next to the transcription of utterances, will be useful in the later analysis. The procedures to obtain this information are described in Section 4.2. As the amount of transcribed conversations is too large to allow for manual qualitative investigation we needed to develop a method for automatic extraction of child-adult utterance pairs which contain corrective feedback. To have a training set for an automatic classifier a subset of all transcripts was manually annotated. The annotation scheme is established in Section 4.3. Certain properties of corrective feedback were extracted from the annotated files and are presented in Section 4.4. Finally, Section 4.5 briefly sums up the chapter.

4.1 Data Selection

To obtain data for the empirical investigation the part of the CHILDES database which covers the English language was used as a basis (MacWhinney, 2000a). However, as the CHILDES database is very large and diverse, not all available

corpora, folders and files could be used in the end. Many constraints needed to be employed. We will now describe this initial selection process.

For the current study normal children's natural language development over a certain period of time in conversation with adult interlocutors needed to be represented. Thus first of all, studies with a wholly different focus were excluded. These contain conversations of children with hearing impairments, non-spontaneous speech (such as identification of pictures), conversations between children without adults, conversations between adults without children and folders which were not grouped according to the investigated child but for example according to the situation.

For those studies which did observe normally developing children longitudinally in conversation with adults still a greater amount of coherence in the data needed to be assured. First of all, some transcripts in otherwise useful corpora contained input by only one speaker and were thus excluded. Next, as *grammatical* errors in child speech play an important role in the current study, we wanted to make sure that the children do already master language to a certain degree in the used files. As the exact time at which certain features of language are acquired can be very diverse, a measure representing proficiency was chosen for this, rather than relying on age. Thus transcripts were excluded if the mean length of utterance - in words - (MLU) of the child was below 2. With an average of more than one word per utterance the child is at a stage where it needs to apply grammatical rules to combine words into interpretable sequences. Next, very short files were excluded. That is, those that contained less than 50 child utterances or less than 100 exchanges in total. Consequently we wanted to assure that a considerable development of the child's language proficiency is observable in the files, hence we excluded folders if the age of the corresponding child did not range over at least one year. And last, to make sure enough data points were available, children's files were deleted if they did not still have a total of at least 10 files and a file density of at least five files per year. Theoretically, almost all these files could be in the first month of the first year with only one other file at the very end. Plotting the count of the number of files against the age of the child, showed that this was never the case. Figure 4.1 shows two representative examples of such plots, the others are in Appendix

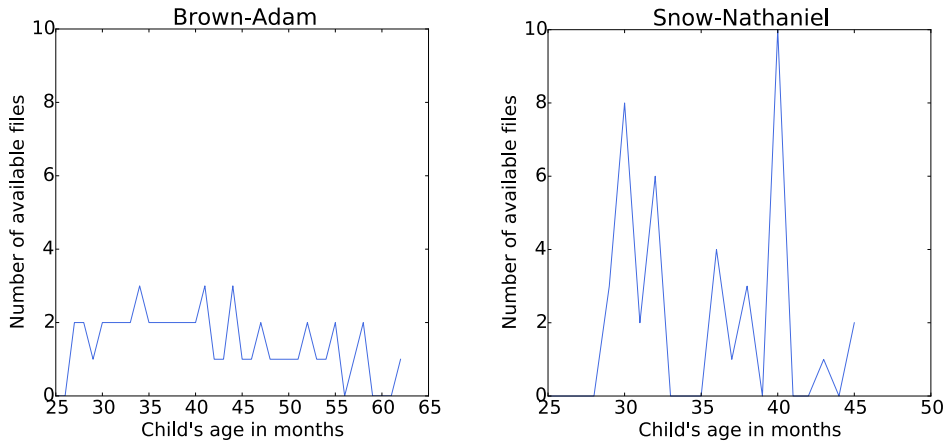


Figure 4.1: The number of available files against month of the child's age for Adam in the Brown corpus (Brown, 1973) and Nathaniel in the Snow corpus (MacWhinney, 2000b).

B.1.

The complete list of folders which were used in the final analysis is given in Appendix B.2. A total number of 1,683 files from 25 different children in 15 corpora were used, containing 628,988 child utterances. The average age which the children had at the time of the first transcribed conversation lay around two years. The mean difference between the child's age in the first and the last gathered file varies considerably more between corpora, but overall also lies around 2 years. If this seems rather high, recall that the minimum amount of time that needed to be covered was limited to 1 year, so all the corpora taken into consideration have a longitudinal focus. These properties are summarized in table 4.1, together with the mean number of files per child and mean number of child utterances per file. We can see that both the amount and the length of files vary significantly between corpora. This will need to be kept in mind later on, when we consider measures of the child's language abilities at certain points in time. In corpora with a considerably lower density of the contained information single outlier files containing non-representative samples will have a greater influence.

Overall, we made sure that the selected files contain enough meaningful information for our investigation.

Corpus	start age	agerange (months)	# files per child	# child utt. per file
Lara	2;1	14.5	106	477.6
Thomas	2;2	33.2	238	641.3
Belfast	2;4	22.5	13.5	232.5
Bloom70	1;11	14.1	15	1597.5
Braunwald	1;5	64.8	95	296.7
Brown	2;4	33.3	82.5	560.0
Clark	2;2	12.0	47	386.5
Demetras	2;0	23.0	26	268.3
Kuczaj	2;4	31.6	203	111.4
MacWhinney	2;4	61.3	226	147.6
Providence	1;9	21.8	41.5	506.7
Sachs	1;10	34.5	50	238.9
Snow	2;5	15.5	40	335.0
Suppes	2;0	15.7	49	633.9
Weist	2;4	27.5	36.75	268.9
Overall	**2;1**	**26.9**	**67.32**	**373.7**

Table 4.1: Average starting age, covered agerange, number of files and number of child utterances per file for the corpora used in the investigation. In total 628,988 utterances from 25 different children were analysed.

4.2 Data Preparation

The selected files do not all contain the same amount of additional information extending the mere transcription of utterances. To insure a uniform analysis this was adjusted where needed. Consequently further relevant information was added. A toolbox with many useful programs for this purpose is available for the CHILDES database. This toolbox is called CLAN, an acronym for Computerized Language ANalysis. It provides programs for the automatic examination of conversations transcribed in the CHAT format, which is the format used in the CHILDES database (MacWhinney, 2000a). Additionally, certain tools are

aimed specifically at the analysis of children's language data, which is beneficial in our case.

To make later examples understandable, it will be helpful to specify a few properties of the CHAT format. A file starts out with a header, each line of which is preceded by *@*. The header contains information on the speakers, the situation, etc. For every speaker a three letter label is specified, where the target child will always be *CHI*. Consequently the conversation is transcribed. One line contains one utterance; a turn can span several lines. Utterances are preceded by *, followed by the three letter label of the speaker, a colon, tab, and then the transcribed text. It is also possible to add tiers containing additional information, for example on morphology of the preceding utterance, or simply comments on the situation. These lines are preceded by %, followed by a three letter label specifying the type of tier, colon, tab and the comment. A file ends with an *@END* line.

4.2.1 Removing non-uttered words

Concerning the additional information provided in the various corpora, first of all in some transcripts a lot of emphasis was put on making the child utterances understandable despite missing words. Thus these words were transcribed and marked as missing using a preceding *0*. However, this information is ignored in the syntactic dependency parsing consequently performed on the utterances. While very likely resulting in more accurate parse trees, this feature is counterproductive for our analysis, as we precisely need information on omitted words. Therefore the non-uttered words were removed from the transcripts.

4.2.2 Morphological analysis, part of speech tagging

Furthermore, most but not all of the files are endowed with a morphological decomposition and part of speech tagging. This information will be necessary later on and was therefore added if necessary. From all the files selected for our investigation, only the corpora consisting of data from Lara (Rowland and Fletcher, 2006) and Thomas (Lieven et al., 2009) did not already contain this analysis. The difference between these two corpora and the others will be marked below by referring to the two folders as ENG (without initial morphological annota-

tion) and ENG-MOR (with initial morphological annotation). The tool used to obtain the additional information present in the ENG-MOR folder is available from the CLAN toolbox. It is called MOR and adds a tier, labeled %mor, after each utterance. In this tier, each word occurring in the preceding utterance is equipped with a part of speech tag and decomposed into its morphemes. This tool was thus used on the files in the ENG folder. As an example, an utterance and its analysis, taken from one of Lara's transcripts.

(1) *MOT: they're called marshmallows .
%mor: pro:sub|they~aux|be&PRES part|call-PASTP n|marshmallow-PL .

They is identified as a pronoun, more specifically a subject pronoun. The tilde ~ marks a junction of words using an apostrophe. The *re* is analysed as the present form of the auxiliary *be*, *called* is decomposed as the particle *call* suffixed by the past perfect marker. Finally, *marshmallows* is identified as the noun *marshmallow*, suffixed by its plural marker.

To allow detailed analysis of the transcripts in the ENG folder all occurring word stems needed to be in the dictionary. Approximately 3,000 previously unknown words had to be manually added to the lexicon. Overall, 61 part of speech tags were differentiated in our transcripts. Regarding only the super-classes of tags, where in the above example *they* is only viewed as a pronoun and the *:sub* specification is ignored, 40 different tags occurred.

For files in the ENG-MOR folder, where this analysis was already present, we still needed to remove non-uttered words from the corresponding tier.

4.2.3 Syntactic dependency parsing

Next, the utterances were extended with a syntactic dependency parse. This information was already available for the files in the ENG-MOR folder. However, as explained above, the parser takes non-uttered words into account. Thus it had to be rerun on the modified files in the ENG-MOR folder, with the non-uttered words removed, as well as on the files in the ENG folder. The parser employed here is MEGRASP, developed by Sagae et al. (2007). It is this parser which was used previously for the files in the ENG-MOR folder, and it is available in CLAN.

MEGRASP is specifically aimed at analysing CHILDES data. Extending their previous work in this direction, which still used the manual annotation from Wall Street Journal for training (Sagae et al., 2004), Sagae et. al trained this parser explicitly on CHILDES data. The first 15 files of the Eve corpus (Brown, 1973) were manually annotated as a training set. Cross-validation of the parser on the manually annotated files gave an overall labeled accuracy score of 92.0 % and an unlabeled accuracy score of 93.8 %, with slightly higher scores on the adult than on the child utterances. The focus on incorporating child speech correctly in the parsing together with these solid accuracy scores make this parser a good choice for our purposes. MEGRASP differentiates 37 possible syntactic relations, of which 35 occurred in our transcripts. Again, it is useful to look at an example analysis, where we will re-use example (1).

(2)
*MOT: they're called marshmallows .
%mor: pro:sub|they~aux|be&PRES part|call-PASTP
n|marshmallow-PL .
%gra: 1|3|SUBJ 2|3|AUX 3|0|ROOT 4|3|OBJ 5|3|PUNCT

The %mor tier is used as input for the dependency parsing. Words are accessed via their indices, where 0 is the implicit root word and punctuation counts as the last occurring word. Note that here *they* and *are* are correctly counted as two words. Relations are represented by triples $i|j|R$, where i is the index of the dependent word, j the index of the head word and R a description of the relation. Thus the above annotation represents the tree depicted in figure 4.2. *Called*, as the main verb in this sentence, is the sole dependent of the root, and the sole head of the punctuation. These relations are labeled ROOT and PUNCT, respectively. *Are* is identified as the auxiliary of *called*, *they* as the subject, and *marshmallows* as the object.

4.2.4 Overlap in child-adult exchanges

Subsequent to the morphological and dependency analysis we extracted information on the matching, added and deleted words between a child and following parent utterance. To this end we employed the CHIP program, equally provided in CLAN. It was designed by Jeffrey Sokolov and focuses on the analysis of adult-

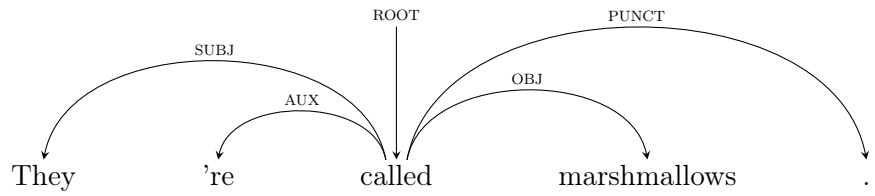


Figure 4.2: Dependency tree extracted by the MEGRASP parser for the sentence “They’re called marshmallows ."

child interaction (Sokolov and MacWhinney, 1990). Speakers can be labeled as adult or child by the user and consequently pairs of utterances, in any constellation of speakers for the source and the response utterance, are analysed. A tier containing the results of the analysis is added after the response utterance, where the label of the tier represents the constellation of speakers (child-adult, adult-child, child-child or adult-adult).

The information contained in the tier lists the added, deleted and exactly matching words between the source and response utterance, after the codes $ADD, $DEL, and $EXA, respectively. Next, in case the response is an exact match, an expansion, or a reduction of the source the codes $EXACT, $EXPAN, or $REDUC are included. The distance - in utterances - between the source and response is given as an integer after the comment $DIST. Finally, the comment $REP specifies the number of words in the response utterance which matched exactly with words from the source utterance divided by the total number of words in the response utterance. That is, $REP specifies the amount of repetition as a value between 0 and 1. Note that $EXACT and $REP = 1.00 do not represent the same information, as $REP = 1.00 also occurs in the case of a reduction.

The CHIP program enables limiting the source utterance speaker - response utterance speaker constellation by suppressing output of unneeded tiers. As we only require information concerning adult responses to child utterances this was the only comment added. The corresponding tier is labeled %adu. The target child is labeled CHI in all CHILDES files, hence this was the only speaker specified as child. For the adults the case is more diverse. Different notations are used for mother and father, such as MOT, FAT, MOM, DAD. Additionally,

sometimes neither of the direct parents participated in the conversation but instead the grandmother, the grandfather, or an investigator. The CHIP program does not allow for an unlimited amount of adult speakers, therefore the participants had to be extracted and adjusted separately for each child. By default CHIP takes as input the %mor tier after an utterance so that morphological information is taken into account. This is very useful as certain overlapping words, such as *do* and *don't* would otherwise not be recognised. Chip extracts only the relevant utterance pairs to a new file, so this information was subsequently merged with the previously obtained annotation. The following example of an utterance pair together with its analysis is taken from Adam in the Brown corpus (Brown, 1973).

(3)
*CHI: there go one .
%mor: adv|there v|go pro:indef|one .
*MOT: yes there goes one .
%mor: co|yes adv|there v|go-3S pro:indef|one .
%adu: $EXA:there-go-one $ADD:yes $EXPAN $MADD:-3s $DIST = 1 $REP = 0.75

There, *go*, and *one* are identified as exactly matching words, *yes* is added. On the morphological level the third person singular form of *go* is added. Overall, the adult utterance is an expansion of the child's. The child utterance comes directly before the adult utterance, the distance between the two is 1. The fraction of repeated words in all words of the adult utterance is 0.75.

4.2.5 Summary of data preparation

To conclude, all utterances were equipped with a morphological decomposition, part of speech tags, and syntactic dependency analysis. Additionally information on the amount of overlap as well as the added, deleted, and matching words between child-adult utterance pairs was extracted.

4.3 Annotation

Having selected the data to use and equipped it with the necessary additional information the next step in the empirical analysis was to manually annotate a subset of files for the presence of corrective feedback. These annotated instances will then work as a training set for our classifier.

4.3.1 Selection of files for annotation

For each set of files (ENG, ENG-MOR) four files each from two children - thus sixteen files in total - were selected for annotation in the first round. The first criterion to be met was that the length of the file, in child utterances, did not diverge more than 20 utterances from the average of the corresponding folder. Afterwards it was ensured that an agerange of over one year per child was covered by those files still under consideration. Similarly, children for which the starting age was very high (above 4 years) were disregarded.

After this preselection, two children of those still available were randomly chosen per language using the *random* module in python (Matsumoto and Nishimura, 1998). However, for the ENG folder only Thomas fulfilled all the criteria. Taking into consideration also transcripts diverging in length up to 50 child utterances from the language mean, Lara's files covered over one year and were also used. Overall, in the ENG folder files from Lara (Rowland and Fletcher, 2006) and Thomas (Lieven et al., 2009) were used. In the ENG-MOR folder Trevor in the Demetras corpus (Demetras, 1989) and Emily in the Weist corpus (Weist et al., 2009; Weist and Zevenbergen, 2008) were selected.

To choose four entries per child the available transcripts were manually presorted into four groups according to the age of the child at the time of recording. From each group one file was selected randomly, again using the *random* module in python. In a second round, another set of transcripts from the same children was selected for annotation. The aim was to have a comparable starting MLU and comparable age intervals covered to conduct a preliminary analysis of the effects which would be later on investigated in large scale.[1] It turned out that

[1] The results of this initial inspection are not as informative as the ones obtained later on and will therefore not be reported.

Emily had a very large MLU consistently over all files. Hence only 2 files each from Thomas, Lara and Trevor were selected for further annotation.

The corpora, children and transcripts which were annotated, together with the age of the child in each file, are summarised in table 4.2.

4.3.2 Selection of exchanges for annotation

Before the actual annotation was executed further preliminary steps were taken to reduce complexity of the files. Corrective feedback is not a highly common phenomenon. When classifying child-adult utterance pairs into corrective feedback vs. non-corrective feedback, the second class will be much bigger. Thus it would be very hard to find a good machine learning classifier on unfiltered data, and we needed to reduce the amount of non-corrective feedback instances by employing an appropriate pre-selection of exchanges. This has the additional advantage that annotation will be faster.

Recall from section 3.2 that we defined adult utterances incorporating corrective feedback as containing some but not complete overlap with the corrected child utterance. Thus by definition it suffices to solely look at those adult responses to child utterances which contain a partial repetition. The necessary information to automatically execute this selection is contained in the tier extracted by the CHIP program: an exchange has partial overlap if the repetition value, specified after the $REP comment, lies above 0 and the comment $EXACT is not mentioned.

However, not all child-adult utterance pairs containing overlap do so due to the presence of an intended parental repetition of part of the child's statement. Most obviously, certain words are simply so common that they can easily occur in consecutive statements by pure chance. Thus a list of stopwords was derived empirically, and exchanges were excluded if they contained a maximum number of two words overlap, both of which were in this list of stopwords. It was evident from a preliminary round of annotation that this procedure was indeed innocuous and did not exclude any exchanges which exhibited corrective feedback. Now for the procedure employed in creating the list of stopwords: first of all a list of the 100 most frequent words in all files was extracted. Consequently those words in the list which were too *meaningful* were manually deleted from it.

Folder	Corpus - Child	Files	Age
ENG	Thomas-Thomas	Thomas-2-07-29.cha	2;07
		Thomas-2-09-03.cha	2;09
		Thomas-2-11-05.cha	2;11
		Thomas-3-01-15.cha	3;01
		Thomas-3-06-01.cha	3;06
		Thomas-4-04-06.cha	4;04
	Lara-Lara	Lara-2-01-25.60.cha	2;01
		Lara-2-06-16.45cha	2;06
		Lara-2-10-22.105.cha	2;10
		Lara-2-11-10.90.cha	2;11
		Lara-3-01-26.60.cha	3;01
		Lara-3-03-10.45.cha	3;03
ENG-MOR	Demetras-Trevor	tre02.cha	2;00
		tre04.cha	2;01
		tre07.cha	2;06
		tre09.cha	2;08
		tre21.cha	3;03
		tre28.cha	3;11
	Weist-Emily	emi03.cha	2;07
		emi07.cha	2;09
		emi19.cha	3;04
		emi21.cha	4;03

Table 4.2: The files selected for manual annotation.

Being too meaningful is defined here as being easily imaginable in an exchange with only one or two of these words overlap which does contain corrective feedback.[2] Excluded words are for example: question words (*who when how why what where*), verbs (*do are can have go want see know, ...*), numbers (*one, two*) and frequency words (*all, some, little, more, ...*). The complete lists is presented in Appendix B.3.

Finally, after a preliminary manual qualitative scan of the child-adult utterance pairs it became visible that in cases where the child utterance contained only one different word (one type) it is very hard to decide whether a following adult utterance presents corrective feedback or not. Therefore these exchanges were also labeled as non-corrective feedback and excluded from the annotation.

Summary of the selection process

Overall, child-adult utterance pairs were extracted for annotation if they had a percentage of repetition $REP greater than 0, did not exactly match, did not contain only maximally two words of overlap both of which were in the stopword list, and in which the child utterance contained more than one different word. This resulted in a total of 2,627 pairs of child-adult utterances to be annotated (1,817 from the ENG and 810 from the ENG-MOR folder).

4.3.3 Annotation scheme

Annotation was performed using the provided coder mode in CLAN. This mode enables the automatic creation of tiers with a fixed selection of possible entries. Our tier was labeled %cof, a name previously unused and short for corrective feedback. The possible entries first of all list whether the exchange contains corrective feedback, a clarification question or neither.[3] Principally every kind

[2] The *easily* in this description is of importance, as we are concerned with a large scale empirical analysis. While extreme cases can occur in which an exchange contains only words from the currently used stopword list as overlap and still does contain corrective feedback, these cases will be rare and can therefore be ignored for our case.

[3] While we do not investigate clarification questions in the present analysis these were still included in the annotation for possible use in later work. Clarification questions form another schema of parental reaction which could point to a child's grammatical error. It might therefore be useful to investigate them later on.

of corrective feedback can also be seen as a clarification question, presuming the right intonation. This overlap is avoided here by defining clarification questions as containing a certain amount of overlap with the preceding child utterance but replacing some part of the sentence with a wh- form or similar question. Thus, a sentence can only contain both corrective feedback and a clarification question if these refer to different errors. This first selection of annotation labels is depicted in the first three lines of the decisiontree in Figure 4.3.

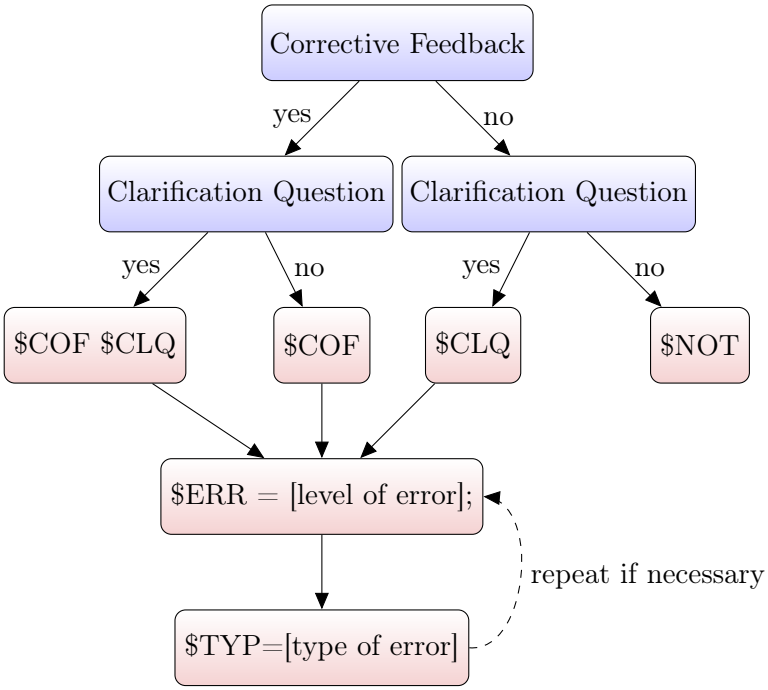


Figure 4.3: Decision tree for the annotation. Blue nodes represent decisions, red nodes corresponding annotations.

Subsequently the affiliation of the instance of corrective feedback with the appropriate class of the taxonomy presented in Section 3.3 was also specified. Specific comments list the levels as well as types of corrected errors. For instances of corrective feedback which did not fit into this differentiation the level and type of error *other* were added. This same subdivision was applied to clarification questions.[4] Several errors are listed by adding one comment each.

[4]The corrective feedback and clarification question markers expect a specification of the error. Thus, if no error was visible prior to a clarification question, for example due to

Figure 4.3 shows the complete decision tree for the annotation, together with the labels used to designate the described markers.

To get an impression of these annotations, let us look at some examples again, taken from different transcripts:

(4)
*CHI: what about kiss ?
*DAD: what about a kiss ?
%adu: $EXA:what-about $EXA:kiss $ADD:a $EXPAN $DIST = 1 $REP = 0.75
%cof: $COF $ERR = umorph:det;$TYP=omission

(5)
*CHI: just orange one .
*DAD: just what ?
%adu: $EXA:just $ADD:what $DEL:orange-one $DIST = 1 $REP = 0.50
%cof: $CLQ $ERR = umorph:det;$TYP=omission

(6)
*CHI: I want in a bottle .
*MOT: please may I have in a bottle .
%adu: $EXA:i $EXA:in-a-bottle $ADD:please-may $ADD:have $DEL:want $DIST = 1 $REP = 0.57
%cof: $NOT

(7)
*CHI: got big parcel here for you .
*MOT: you've got a big what for me ?
%adu: $EXA:get $EXA:big $EXA:for $ADD:you've $ADD:a $ADD:what $ADD:me $DEL:parcel-here $DEL:you $MSUB:&pastp $DIST = 1 $REP = 0.43
%cof: $COF $CLQ $ERR = synt:subj;$TYP=omission $ERR = umorph:det;$TYP=omission

Example (4) contains corrective feedback, and is thus labeled with $COF. The corrected error is determiner omission, in the set of errors in unbound morphology according to our list in section 3.3. Exchange (5) is a standard

imprecise pronunciation which was correctly transcribed, then still a *0* error comment was added to specify this. This does by definition not occur for corrective feedback.

example of what is labeled as clarification question here, hence the marker $CLQ. The child error which seems to have elicited the clarification request is determiner omission, thus this error is annotated. Arguably, in the case of clarification questions, extracting the error which led to the question is not necessarily well-defined. Inter-annotator agreement on this task will have to be closely monitored when clarification questions get investigated. Example (6) contains neither corrective feedback nor a clarification question, and is thus labeled $NOT. Finally, Example (7) shows a case where both corrective feedback and a clarification question occur in the same utterance.

4.3.4 Annotation reliability

To be able to evaluate the quality of the annotation the child adult utterance pairs preselected from 2 files (one from Lara and one from Thomas in the ENG folder) according to the procedure described in Section 4.3.1 were also annotated by my supervisor, Raquel Fernández. A total of 350 exchanges fit the preselection criteria in these two files. After a first round of annotation the mismatches were evaluated qualitatively. For some of the disagreements consensus was easily found. Others showed structural issues. One common source for controversy were exchanges containing a non-sentential child utterance. This is often correct in spoken dialogue, with the semantics of the full sentence being clear from the context. In written language however, non-sentential utterances would be considered false. Therefore judgements diverged on whether their corrections are to be judged corrective feedback (recall that we defined corrective feedback as responding to an *error*). One example of such a disagreement is the following, taken from Lara.

(8)

*CHI:	just like those.
*MOT:	we'll make one a bit like those .
Annotator 1:	$NOT
Annotator 2:	$COF $ERR = synt:subj;$TYP=omission
	$ERR = synt:verb;$TYP=omission

Another sort of child utterances often giving rise to disagreements are very incomprehensible ones. The following example from Thomas shows this.

	$CLQ	$COF	$NOT
$CLQ	⟨ 9 ⟩	–	12
$COF	–	⟨ 46 ⟩	17
$NOT	1	5	⟨ 260 ⟩

Table 4.3: Confusion matrix for the judgements on $COF, $CLQ and $NOT. Cohen's κ is 0.71.

(9)
*CHI:	a more bus tin more sweets in that one .
*MOT:	but but the sweets in the bus tin are for later , aren't they ?
Annotator 1:	$COF $ERR = umorph:det:$TYP=omission
	$ERR = umorph:prep;$TYP=omission
Annotator 2:	$NOT

Overall, of the 57 disagreements in the first round 18 were resolved in discussion. Inter-annotator agreement was computed using Cohen's κ, a chance-corrected coefficient that takes annotator bias into account.[5] The confusion matrix is depicted in table 4.3, Cohen's κ is at 0.71 for the three categories $COF, $CLQ and $NOT. Not taking clarification questions into account and only considering corrective feedback vs. non-corrective feedback judgements the κ value is 0.77.

4.4 Properties of Corrective Feedback

Using the manually annotated files we can extract certain statistical properties of corrective feedback. For this, the numbers observed in the ENG and ENG-MOR folder are averaged to give an impression of the properties in the English language overall, which are not influenced by the morphological information present or absent in the original transcriptions. First of all, the percentages of corrective feedback, clarification questions and neither in those exchanges which were annotated are summarised in Table 4.4. Corrective feedback has a much higher frequency than clarification questions. Still, neither of the two do absolutely occur often. In the annotated files, 14.8% of all exchanges exhibited corrective feedback, and 3.8% a clarification question. Looking at these numbers

[5] See Artstein and Poesio (2008) for a discussion of agreement measures.

it should be kept in mind that the annotated exchanges were already largely limited in favor of containing corrective feedback. Additionally, recall that an utterance can in rare cases contain both corrective feedback and a clarification question, hence the reported fractions do not sum precisely to one.

Next we can also look at statistics concerning the linguistic level and type of errors. For adults - both first language speakers and second language learners - the most common error type are substitutions (Foster, 2007). They know that a word needs to appear but pick a wrong implementation. Contrastingly, as was already mentioned above, in child speech most errors would be expected to be omissions. This expectation was confirmed by the present data. The distribution of the types of errors is presented in Table 4.5 and the detailed distribution of both location and type is listed in Table 4.6. These numbers were computed looking only at exchanges which contain corrective feedback or clarification questions. This introduces a bias into the numbers: they are not representative of the overall distribution but only of the distribution of errors picked up by the parents. However, the annotation scheme did not consistently require errors to be noted in the absence of corrective feedback and clarification questions. Thus using the overall distribution in the annotated files would introduce a subjective bias based on the annotator's preference for marking certain errors, instead of the structural bias now included. Most errors are made at the level of syntax (47% of all errors) and unbound morphology (41%), and by far most error types are omissions (87.5%). The most common overall errors are subject omission (27%), main verb and auxiliary verb omission (15% and 18% respectively) and determiner omission (13%). It is also interesting to note that errors for which the type could not be classified into omission, addition or substitution are extremely rare and were almost exclusively also errors which did not fit into the labels we chose for the linguistic levels. In the latter case then child utterances were simply in general rather incomprehensible. This distribution indicates that the differentiation of error types is almost exhaustive.

Another feature visible from the annotated data which does not come surprising and was shown before (Saxton et al., 2005) is that the frequency of corrective feedback decreases over time. Considering that children simply make fewer mistakes when they grow up, a necessary precondition for corrective feed-

Type of feedback	Absolute number of occurrences	Fraction
$COF	389	0.148
$CLQ	100	0.038
$NOT	2,146	0.817

Table 4.4: Distribution of corrective feedback and clarification questions in the annotated exchanges. Total number of exchanges: 2,627.

Type of error	Absolute number of occurrences	Fraction
omission	532	0.875
addition	14	0.023
substitution	54	0.089
other	7	0.011

Table 4.5: Distribution of the observed types of errors that get corrected with corrective feedback or taken up in clarification questions. Total number of errors: 608.

back, parents will also present fewer corrections. This is confirmed in the current files. Figure 4.4 shows the frequency of corrective feedback against the child's age for the four children from which files were annotated. The frequency of corrective feedback is computed as the number of exchanges containing corrective feedback divided by the number of exchanges that were annotated. Again, as the utterances which were annotated are largely limited in favor of containing corrective feedback this percentage is not representative of the overall number. However, as the files in all ageranges were prepared in the same way this should not influence the presented development over time. It is visible that for three out of the four children the development is as expected, with a large negative Pearson correlation of the age and the amount of corrective feedback given. The value for Lara is -0.91, for Thomas -0.87 and for Trevor -0.95. Emily in the Weist corpus presents an exception, with a Pearson correlation of +0.32. Possibly related to this, Emily also consistently has a very high MLU. A plot of the development of MLU against the child's age is shown in Figure 4.5. We can see that Emily starts with a MLU at age 2;5 which the other children achieve much later. This might indicate that she is rather advanced in her development

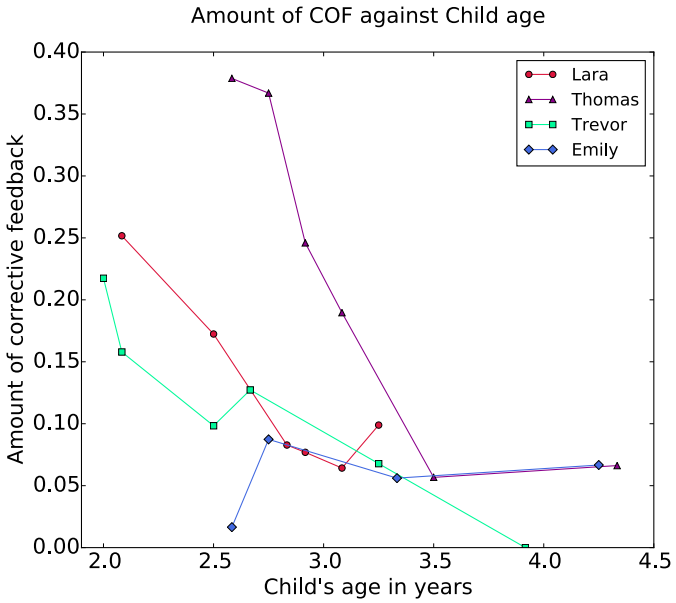


Figure 4.4: Development of the fraction of exchanges in the annotated files containing corrective feedback against the child's age. Pearson coefficient between these two measures for Lara is -0.91, for Thomas -0.87, for Trevor -0.95 and for Emily +0.32.

in general, which could also be related to the fact that the amount of corrective feedback which she receives is non-typical.

4.5 Summary of the Chapter

In this chapter we developed the selection and preparation of data to be used in the further analysis. All relevant transcripts from the CHILDES database (MacWhinney, 2000a) in the English language were used. We described in detail what was considered as *relevant* in Section 4.1. The transcripts had to contain free speech from normally developing children in conversation with adult interlocutors. Additionally, the files had to cover a sufficient agerange in the development of the child with a sufficient density. Subsequently, in Section 4.2 we presented the employed preliminary addition of necessary information. Utterances were equipped with morphological analysis and part of speech tagging as well as a syntactic dependency analysis. Child-adult utterance pairs were provided with an analysis concerning the overlap between them. As input into

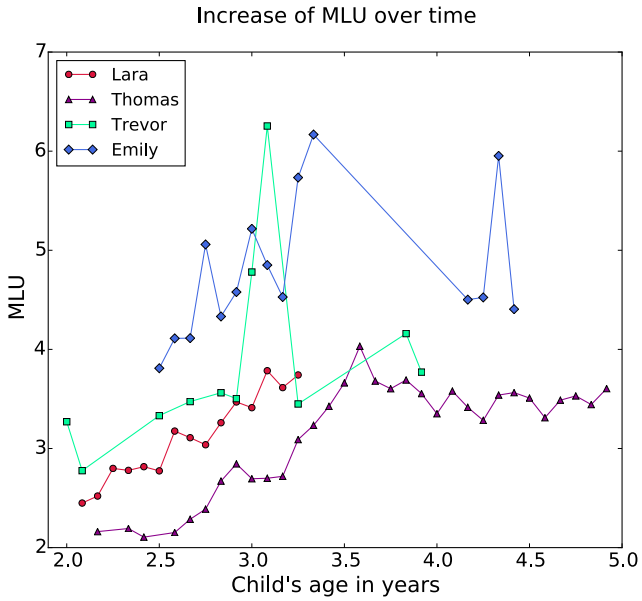


Figure 4.5: Development of MLU against the child's age for the children from which files were manually annotated.

a classification algorithm we need exchanges manually annotated for the presence of corrective feedback. In Section 4.3 we first of all developed the selection criteria for files to be annotated and consequently preselection criteria for candidate corrective feedback utterance pairs. Only exchanges showing a certain amount of non-stopword overlap and containing a childutterance with at least two different words were considered for annotation. Following this preselection, we established the annotation scheme. In addition to marking the presence or absence of corrective feedback the level and type of the corrected error were also noted, corresponding to the differentiation of child errors presented in Section 3.3. Finally, the annotated files were analysed and certain statistical properties were extracted. It was visible that corrective feedback is overall not a highly common phenomenon, occurring in only 15% of the annotated exchanges. The most common types of error in child speech are omissions, the most common linguistic levels of errors are at subject, verb, determiner and auxiliary verb position. Following this, in Chapter 5 we will devise methods for automatically classifying corrective feedback instances based on the annotated files.

Level \ Type	Total	Omission	Addition	Substitution	Other
syntactical:					
subject	**163 (0.27)**	**162 (0.27)**	0	1 (0.002)	0
verb	**94 (0.16)**	**92 (0.15)**	1 (0.002)	0	1 (0.002)
object	22 (0.04)	22 (0.04)	0	0	0
other	0	0	0	0	0
noun morphology:					
poss -'s	6 (0.01)	5 (0.008)	1 (0.002)	0	0
reg. plural	3 (0.005)	0	3 (0.005)	0	0
irr. plural	3 (0.005)	0	0	3 (0.005)	0
other	0	0	0	0	0
verb morphology:					
3rd pers	5 (0.008)	5 (0.008)	0	0	0
reg. past	11 (0.02)	10 (0.02)	1 (0.002)	0	0
irr. past	5 (0.008)	1 (0.002)	0	4 (0.007)	0
other	4 (0.007)	3 (0.005)	0	1 (0.002)	0
unbound morphology:					
determiner	**87 (0.14)**	**81 (0.13)**	0	6 (0.01)	0
preposition	35 (0.06)	22 (0.04)	1 (0.002)	12 (0.02)	0
aux. verb	**117 (0.19)**	**111 (0.18)**	5 (0.008)	1 (0.002)	0
pres. progr.	9 (0.02)	9 (0.02)	0	0	0
other	1 (0.002)	1 (0.002)	0	0	0
other	43 (0.07)	9 (0.01)	2 (0.003)	26 (0.04)	6 (0.01)

Table 4.6: Detailed distribution of levels and types of errors in the annotated exchanges which get picked up with corrective feedback or clarification questions in English language. Numbers outside the brackets represent absolute occurrence counts, numbers in brackets show the fraction in all errors. Total number of errors: 608.

Chapter 5

Automatic Extraction

The candidate instances of corrective feedback selected and annotated according to the criteria described in Chapter 4 were next used as the training set for an automatic classifier deciding whether child-adult utterance pairs exhibit specific child errors and parental corrective feedback. In Section 5.1 we explain briefly how the classifier employed here - a support vector machine - works.

In a first step, we tried to achieve the extraction of instances of corrective feedback globally, for all locations and types of child errors. A set of features aimed at being able to represent the relevant contrast between the child and adult utterances was developed. The process for this is described in section 5.2.1. The achieved accuracy scores, presented in section 5.2.3, however, showed that the phenomenon of corrective feedback is too diverse for the attempted general approach. It was discernible that the applied extraction method was unable to capture all the different semantic, syntactic and pragmatic influences showing whether any given exchange contains a corrective reformulation.

We therefore proceeded by limiting the investigated phenomenon to a more clearly confined space: instances of corrective feedback correcting children's subject omission errors. We saw in Section 4.4 that this combination of location and type of error is most frequent over all errors. We hoped that by constraining the search space it would be more feasible to find features able of capturing the desired contrast between the child and adult utterance. Next to the extraction of exchanges containing corrective feedback on a subject omission error, we also automatized the extraction of the error itself. This process is described in Section 5.3. In Section 5.4 we develop the classification method for corrective

feedback on this specific error type. An automatic extraction of instances of children's errors as well as responses containing corrective feedback will enable us to assess the effect of this response pattern on the acquisition of the correct grammatical form. Finally, in section 5.5 we give a concise recap of the chapter.

5.1 Support Vector Machines

A support vector machine (SVM) is a machine learning algorithm that performs two-class classification.[1] Classification is achieved via the position of datapoints relative to a separating hyperplane. This plane is what is learned by the algorithm. It is selected such that the distances (margins) between the plane and the closest points in each dataset are maximal. Figure 5.1 shows an example of two linearly separable classes of datapoints together with a separating hyperplane and the margins. Clearly, this approach only works well if the datapoints are indeed linearly separable. However, for the case of non-linearly separable data classification is still possible. The datapoints are projected into a higher - possibly infinite - dimensional space, and the separating hyperplane is found in this projected space.[2][3]

Generally, a support vector machine takes as input a set of instances with corresponding labels. The separating hyperplane is learned from these training points. Consequently the label of unseen datapoints can be predicted. Input instances are given in form of a matrix, with one row representing one instance, one column giving the values of one feature. The labels are represented as a binary vector.

We chose to use a support vector machine as the classifier here because it gives reliable results even when the number of features is relatively large in com-

[1] The algorithm can be extended to the multiclass case using several classifiers and a voting scheme.

[2] See for example Bishop (2006) or Burges (1998) for a more detailed introduction into support vector machines.

[3] As an easy figurative exemplification of why projection into higher dimensional spaces can enable linear separation, imagine two 2-dimensional datasets, one inside and the other outside of a 1-sphere with center (0,0). These sets are not linearly separable in the 2-dimensional space. However, projecting them onto an elliptic paraboloid centered around the origin in 3-dimensional space they do become separable by a hyperplane.

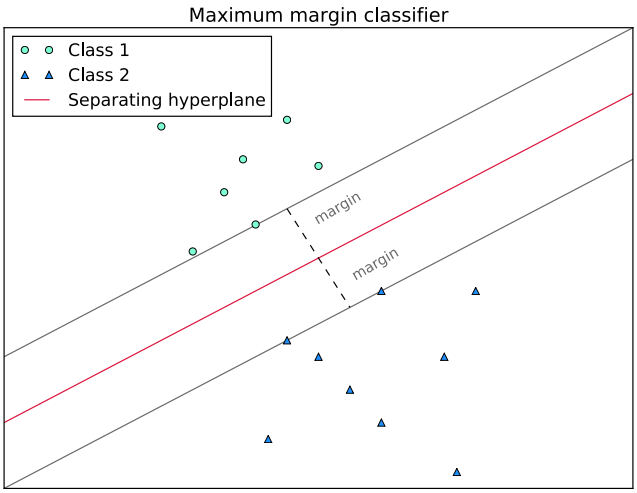


Figure 5.1: An example of a two-class classifier on linearly separable data, showing the separating hyperplane and the margins between the plane and the closest datapoint in the two classes.

parison to the number of instances. This was the case in the global extraction of corrective feedback instances. The computations will be executed using the provided implementation in the scikit-learn module in python (Pedregosa et al., 2011).

5.2 Global Extraction of Corrective Feedback

For the global extraction of corrective feedback exchanges we compiled a large list of possibly meaningful features to represent the data and used these as input into a support vector machine. The extracted features are described in Section 5.2.1. In Section 5.2.2 we describe how the accuracy of the classifier is monitored. In Section 5.2.3 we present the obtained accuracy scores. These show that corrective feedback in general is too complex of a phenomenon to be classified with the employed means.

5.2.1 Features

The task of choosing which features to extract from the textual data for presentation to the classification algorithm is not a trivial one. A lot of information is contained in the exchanges, so we decided to start out by compiling a large list of features before possibly reducing these again according to predictive strength. Considering that the accuracy of the classification algorithm based on these features did not meet our expectations, the intended reduction was abandoned. Also, the extraction of features will not be described in detail here. A survey of the features is presented, and more detail is given in Appendix C.

Establishing the contrast between child and parent utterance

The intuitively most striking element present in all instances of corrective feedback is that the child and adult utterance use different means to represent the same statement. The two accounts discussed in Section 3.4 give ideas on how the relevant contrast can be established: due to a difference in form despite a converging *meaning* or converging *semantic function*. The fact that the parent utterance diverges in form from the child utterance was already represented by our preselection of those utterance pairs which exhibit non-complete overlap.

Semantic similarity between the two utterances was portrayed by computing the distance between semantic representations of each utterance. A semantic vector representation of each word was obtained using word2vec, an implementation which has proven to yield competitive results on semantic similarity tasks (Mikolov et al., 2013a,b). Subsequently, the vectors for single words were combined into a vector for the whole utterance using addition. Despite clearly being inaccurate, due to for example commutativity, this method generates a good approximation while additionally being computationally simple (Mikolov et al., 2013b). Distance between the two obtained vectors was computed via two measures: as cosine distance, the standard measure for similarity between semantic vector representations dependent on the angle between the two vectors, and as euclidean distance, taking the length of the vectors into account.

Syntactic similarity between the utterances was represented via the tree edit distance of the syntactic dependency trees obtained from the MEGRASP parser. To compute this, the algorithm presented by Zhang and Shasha (1989) was used.

Features related to added, deleted and exactly matching words

The output generated by the CHIP program, concerning added, deleted and exactly matching words between the child and adult utterances, presents another rich source of information. First of all, the fraction of added and exactly matching words in the adult utterance, and the fraction of deleted words in the child utterance were computed. Next, the part of speech tags of added, deleted and matching words were extracted individually. Finally, also the semantic relations the three sets of words are involved in, as analysed by the MEGRASP parser, were obtained separately for each set.

Overall, four different feature matrices were extracted from the annotated files: with detailed or rough part of speech tags, and with binarised values for the tags and semantic relations or with frequency counts. Subsequently a development set was split from the feature matrix and label vector, to be able to evaluate the accuracy of a predictor with fine-tuned parameters on an independent test set. The test set contained approximately 20% of all instances. Next, features with zero variance were removed, as these do not contribute any information. The same is true for duplicate features, which were also removed.

5.2.2 Training and evaluation setup

The set of features described in Section 5.2.1 together with the correct class labels were used to train a support vector machine. As described in Section 5.1, this results in a two-class classifier. Here, the aim was to distinguish corrective feedback from non-corrective feedback instances.

Accuracy of the classifier was monitored using 5-fold cross validation.[4] That is, the input was split into five parts. Following this, the labels for each part were predicted using the other four parts to train the classifier. These predictions were subsequently compared to the actual labels to measure the quality of the obtained prediction. As we are interested in a classifier that correctly selects instances of corrective feedback, the quality was measured via precision, recall and f-score for the corrective feedback class. Features and parameters are fine tuned to increase explanatory power of the prediction.

[4]The amount of available data is not extremely large and a more finegrained cross validation would lead to very small sets of left out data.

To make sure that the final scores report how well the given approach generalises, the set of all instances is split into a development and a test set. The test set is disregarded during the tuning of features and parameters. Finally predictive accuracy is evaluated on this wholly unseen test set. The number of available instances is not very large, hence the test set was picked to be rather small, containing slightly below 20% of all candidate exchanges. The locations from which these instances were taken were randomly selected.

5.2.3 Results

In the first round, the prediction was run using the full matrices described in Section 5.2.1 and without specifying class weights. This resulted in a classification of all instances as non-corrective feedback. Considering that this class is much larger than the corrective feedback class, this prediction yields comparably high accuracy scores. It was thus selected by the classifier, despite not being informative for our purposes.

Next we modified the class weights such that misclassification of corrective feedback instances as non-corrective feedback receives a higher penalty and is dispreferred. The penalty was increased by the factor 1.5, 2, 5 and 10. Additionally, the classes were weighted negatively proportional to their size, which is labeled as *'auto'*. In our case this lies close to multiplying the penalty for misclassification of corrective feedback instances by 5. F-scores for these modified classifiers are presented in figure 5.2. We see that the classifier mildly increases descriptive strength compared to the previous classification of all instances as non-corrective feedback. However, overall, the obtained scores are still too poor to enable meaningful deductions from the obtained classifications.

Finally, we reduced the feature sets according to empirically derived thresholds for the variance in the features or the correlation between features and labels. Neither of these approaches resulted in improvements of predictive accuracy.

Corrective feedback is a very diverse phenomenon. Additionally, in certain cases the fact that any given exchange contains a corrective reformulation is revealed not only by syntactic or semantic but also pragmatic considerations. We therefore decided not to continue tuning the features on this general classifier

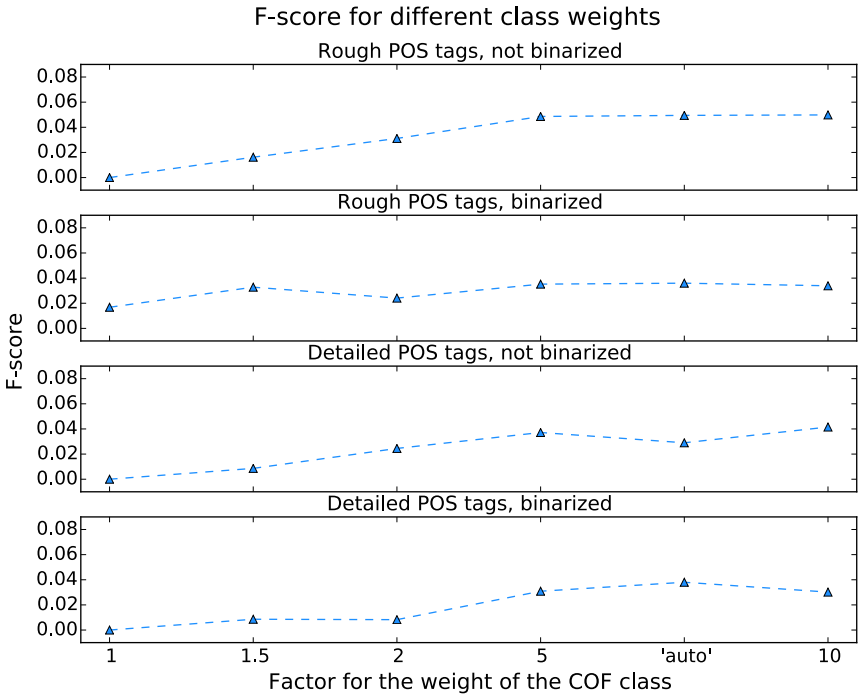


Figure 5.2: F-score of the classifier for different classweights, for all four different feature matrices.

but instead to constrain the search space and focus on a more clearly contoured phenomenon. This procedure is described in Sections 5.3 and 5.4.

5.3 Subject Omission Errors

We saw in Section 4.4 that by far the most common error over all child utterances is subject omission. When deciding how to constrain the investigated phenomenon we therefore chose to analyse this error and the effect of corrective feedback on it. To this end, we first of all devised a method for automatically extracting child utterances containing subject omission errors from transcripts.

5.3.1 Base set for the extraction

The base set for the extraction of children's subject omission errors are all annotated exchanges which were necessarily labeled with the error in the child utterance, i.e. that contained corrective feedback or clarification questions. 489 exchanges meet this criterion. However, as the presence of corrective feedback or clarification questions can take a different value for several adult responses to the same child utterance certain child utterances occur multiple times. These were thus excluded to make sure each child utterance occurs only once. 477 utterances are available after this reduction. Next, the errors marked according to the annotation described in Section 4.3 are those picked up by the adult in the correction, not all errors made by the child. Thus in certain cases the annotation does not mention a present subject omission error. Hence the files were manually marked for the occurrence of this error. The baselines for overall accuracy as well as precision and recall for the subject omission error class taking random classifiers are given in Table 5.1. We split off a test set of 47 utterances taken from random places in the base set, leaving a development set of 430 utterances.

5.3.2 Extraction method

We saw above that a predefined selection of features, which is chosen without examining predictive effectiveness, can fail to represent the necessary relation between the child and adult utterances. Hence we decided here to select and

Features	Precision	Recall	Accuracy
randomly classified $P(1) = P(0) = 0.5$	0.5	0.38	0.5
randomly classified according to class weights	0.38	0.38	0.53

Table 5.1: Baselines for precision and recall for the subject omission class and overall accuracy in the classification of subject omission versus no subject omission.

tune the features in a decision tree setting while qualitatively monitoring the falsely classified instances.

Overall, we were interested in a high-precision classifier. Having a high precision, utterances labeled as erroneous most likely indeed contain a mistake, while some utterances are unnoticed despite an error. Thus the amount of error occurring in the child speech is estimated conservatively.

The first obvious idea for extracting a child's subject omission error was to select those utterances without a subject dependency in the dependency parse. That is, the dependency label *SUBJ* does not appear. This already gave a rather good approximation. However, looking at the falsely classified instances it became apparent that in some cases the dependency parser mistakenly identified negations or phonemes void of meaning as subjects. An example for this is the following, taken from Thomas.

(1) *CHI: not need a (ba)nana .
%mor: neg|not v|need det|a n|banana .
%gra: 1|2|SUBJ 2|0|ROOT 3|4|DET 4|2|OBJ 5|2|PUNCT

Hence in the next step we selected exchanges as subject omission errors if they did not contain a subject dependency label, or if they did contain a subject dependency but the dependent word was a negation or meaningless phoneme, as classified according to the part of speech tags. Subsequently we looked at features to reduce the number of false positives. In this respect, it was visible that sometimes the dependency parser did not recognise the subject of the sentence correctly. Mostly, this was related to other errors in the child utterance

obscuring the structure. One example is the following, taken from Trevor.

(2) *CHI: I going cut your hair .
%mor: pro:sub|I part|go-PRESP v|cut&ZERO pro:poss:det|your n|hair .
%gra: 1|0|INCROOT 2|1|XMOD 3|2|OBJ 4|5|MOD 5|3|OBJ 6|1|PUNCT

Thus utterances were additionally only labeled as subject omissions if they did not start with a noun or pronoun, excluding proper names. Another reason for a missing *SUBJ* dependency relation despite the subject being present in the utterance are incomplete sentences in which the head of the subject relation is missing. If the root of an incomplete sentence (marked *INCROOT*) was directly linked to a proper name, this would often times be the subject. In this case utterances should not be considered subject omission errors despite the absence of a *SUBJ* dependency relation.[5] This consideration constituted the last refinement of features for the extraction of subject omission errors in child utterances.

5.3.3 Results

The precision, recall and f-score values for the subject omission error class in various sets of features employed for the classification are given in table 5.2.

We can see in the obtained scores for the most basic decision feature that this already gave a rather good approximation. As expected, the first added feature mostly increased recall as it leads to more utterances being labeled as error instances. The next two added features decreased the number of utterances selected as erroneous, and hence increased precision.

Overall, with the selected classifier precision lay at 83% on both the development and the test set, and recall at 86% on the development set, versus 80% on the test set.

[5]Substituting 'nouns or pronouns' for 'proper names' in the above selection criterion was also tested, but gave weaker results.

Features	Precision	Recall	F-score
no SUBJ	0.74	0.93	0.82
no SUBJ OR subject is neg / phon	0.75	**0.97**	**0.85**
(no SUBJ OR subject is neg / phon) AND no starting noun	0.82	0.88	**0.85**
(no SUBJ OR subject is neg / phon) AND no starting noun AND not INCROOT dependency on proper name	**0.83**	0.86	**0.85**
→ on the test set	0.83	0.8	0.82

Table 5.2: Precision, recall and f-score for the subject omission class in the classification of subject omission versus no subject omission, for various combinations of features. Overall, 430 instances were classified in the development set and 47 in the test set.

5.4 COF on Subject Omission Errors

Having devised a reasonably reliable method for classifying children's subject omission errors in Section 5.3 we next developed a classifier selecting child-adult utterance pairs containing corrective feedback on a subject omission error.

5.4.1 Base set for the extraction

The base set for this classification were all those annotated candidate corrective feedback utterance pairs which do contain a subject omission error in the child utterance. At first, this latter selection was executed using the automatic classification devised in Section 5.3. 820 exchanges fit this criterion. To make sure we were monitoring the predictive accuracy solely on the corrective feedback classification task, without taking into account the misclassifications of child errors, the exchanges were manually marked for the presence of a child's subject omission error. Exchanges with no subject omission error in the child utterance were without further consideration classified as non-corrective feedback. This

Features	Precision	Recall	Accuracy
randomly classified $P(1) = P(0) = 0.5$	0.5	0.502	0.5
randomly classified according to class weights	0.502	0.502	0.5
Parent utterance contains a SUBJ dependency	0.38	0.88	0.54

Table 5.3: Baselines for precision and recall for the corrective feedback class and overall accuracy in the classification of corrective feedback on a subject omission error versus no corrective feedback on a subject omission error.

resulted in a higher number of correct negatives, but as we were monitoring precision and recall on the corrective feedback class this did not unjustifiedly boost our scores. A test set of 100 exchanges randomly distributed in the base set was split off, leaving a development set of 720 exchanges. The baselines of precision and recall for the corrective feedback class as well as overall accuracy for random classifiers and the most basic feature as sole predictor are given in Table 5.3.

Again, the feature tuning was executed in a decision tree setting before proceeding to use the meaningful features as input into a support vector machine. After a considerable amount of feature selection it became visible that many of the false positives, i.e. exchanges classified as corrective feedback on a subject omission error without being marked as such, could indeed be considered positives. Thus all exchanges were re-annotated with a preference for corrective feedback in doubtful instances.

5.4.2 Feature selection

Finding features to represent corrective feedback was more challenging than finding features to represent a child's subject omission error. The latter is simply a structural deficiency, while for the former we needed to capture the relevant interaction between the child and adult utterance. Hence a more extensive feature search was necessary, which will not be described in detail here. The features selected in the end were the following (all of which are binary values):

1. The child utterance and the overlapping words are not solely non-words (as identified by the part-of-speech tag).
2. The adult utterance contains a *SUBJ* dependency relation.
3. The first word in the adult utterance after a word with part of speech tag *neg* or *co* (words like 'yeah', 'mhm', 'ehm') is an added noun.
4. If the adult utterance is a question (identified by the punctuation mark), in the above statement also a verb or auxiliary verb can occur before the added noun.
5. The dependent of the *ROOT* or *INCROOT* dependency in the child utterance is an exactly matching word.
6. The dependent of the *ROOT* or *INCROOT* dependency in the adult utterance is an exactly matching word.
7. The adult utterance starts with a form of the verb *to be*, and this verb is the head of a predicate or object dependency relation.
8. The adult utterance contains a *SUBJ* dependency relation and the head of this relation is an exactly matching word.
9. The adult utterance contains a *SUBJ* dependency relation and the head of this relation is a word which does not exactly match, but has as its dependent a matching word.
10. An overlapping word is identified as a verb in the adult utterance.
11. An overlapping word is the dependent of an object dependency relation in the adult utterance.

Why features 1. to 4. were included should be clear. Features 5. and 6. are aimed at capturing the fact that the overlap between the child and adult utterance is at an important structural part of the sentence. Feature 7. was added because the dependency parse is often erroneous on questions. Hence the subject is not detected. Patterns like the following, where parents start out a question with a form of *to be*, are very common.

(3)
	*CHI:	go asleep .
	%mor:	v|go adv|asleep .
	%gra:	1|0|ROOT 2|1|JCT 3|1|PUNCT
	*MOT:	is it going to sleep ?
	%mor:	aux|be&3S pro|it part|go-PRESP prep|to n|sleep ?
	%gra:	1|0|INCROOT 2|1|OBJ 3|2|XMOD 4|3|JCT 5|4|POBJ 6|1|PUNCT

Feature 8. is aimed at representing an added subject as a dependent of a matching verb, feature 9. at representing the case of both subject and verb being added. Similarly for features 10. and 11. Some of these features are mutually exclusive, but that does not pose any problems.

5.4.3 Results

The features described in Section 5.4.2 together with the correct labels were used as input into a support vector machine (see Section 5.1). Explanatory power of the predictor was monitored as described in Section 5.2.2, using 5-fold cross validation.

As for the error extraction task we are interested in a high-precision classifier, to have a conservative estimate of the amount of corrective feedback. Thus we increased the weight for misclassification of non-corrective feedback instances to enhance precision. The obtained precision, recall and f-score values for varying classweights are summarized in Table 5.4. Increasing the penalty for misclassification of non-corrective feedback instances by 3 gave the highest possible precision while maintaining an overall f-score of above 0.5. Hence this classifier was selected. The predictions using the extraction of the classifier with equal class-weights and lower precision will be compared to the predictions using this high-precision classifier at an exemplary stage in Chapter 6.

5.5 Summary of the Chapter

In the present chapter we developed methods for automatically extracting instances of specific types of children's errors and corrective feedback from transcripts. In Section 5.1 the employed classification algorithm, a support vector

Classweights	Precision	Recall	F-score
{0 : 1, 1: 1}	0.82	0.43	0.57
{0 : 2, 1: 1}	0.84	0.39	0.53
{0 : 3, 1: 1}	0.89	0.36	0.51

Table 5.4: Precision, recall and f-score for the corrective feedback class in the classification of corrective feedback on a subject omission error, versus no corrective feedback on a subject omission error. Classification was performed using a support vector machine with different class-weights. Overall, 720 instances were classified.

machine, is explained. It is a non-probabilistic algorithm performing two-class classification by deriving a separating hyperplane such that the distance between the plane and each class is maximal. New datapoints are classified according to their location relative to the separating plane. In Section 5.2 the attempted method for classifying general corrective feedback instances is described. The scores of the classifier, presented in Section 5.2.3, did not allow for a meaningful continuation based on this classification. Therefore the phenomenon to be investigated was refined. Subject omission errors are the most common child errors, hence we decided to analyse this error and the effect of corrective feedback on it. Section 5.3 describes the feature selection process for classifying children's subject omission errors. Most features are directly related to the presence of a subject dependency or a noun in the common subject position (at the beginning of the sentence). Precision of the selected classifier on the test set was 0.83. Subsequently, parental corrective feedback on subject omission errors was extracted. This process is described in Section 5.4. Here, the feature selection was more difficult, and the employed features were more diverse. This relates to the fact that corrective feedback is an interactive phenomenon. To extract it, an interaction between the child and adult utterance needs to be captured. Finally, precision for the classification of the corrective feedback class lay between 0.82 and 0.89. Next, these extraction mechanisms will be applied to all transcripts selected in Section 4.1 to derive predictions concerning the influence of corrective feedback on the acquisition of the corresponding structure.

Chapter 6

Towards Language Acquisition

In the present chapter we will investigate the effect of corrective feedback after subject omission errors on the learning of correct subject inclusion by the child. To this end we will use the extraction mechanisms developed in Sections 5.3 and 5.4 on all transcripts selected for analysis in Section 4.1. As the extraction mechanism for corrective feedback works only on a preselection of child-adult utterance pairs, this preselection is also incorporated. Thus, exchanges are interpreted as corrective feedback if they meet the criteria, and consequently are classified as corrective feedback by the algorithm. In Section 6.1 we will start out by describing in detail the experimental setup, as well as how outcomes should be interpreted. Mainly, a correlation analysis between the amount of corrective feedback received and the improvements after a certain timespan give a first indication into the relation. Next, a linear regression analysis with several other input factors is employed to investigate the specific effect of corrective feedback over other predictors. In Section 6.2 we will present the results of the analysis. In Section 6.3 the main conclusions we can draw from these observations are discussed. Finally, in Section 6.4 we briefly summarise the chapter. In the following we will sometimes refer solely to *corrective feedback*, *errors*, or *learning*. This, according to the extraction mechanisms provided, is always limited to corrective feedback on subject omission errors, subject omission errors, and learning to include subjects.

6.1 Experimental Setup

To develop the experimental setup several decisions need to be taken. First of all, we need to select a procedure to group together transcripts into datapoints. A first idea for this would be to consider one transcript as one datapoint. However, as we saw in Section 4.1, the number of available files per child as well as the individual file length vary considerably. Hence a means of organizing these is developed in Section 6.1.1. Next the experimental procedures are worked out in Section 6.1.2. Finally we will discuss what would be expected as results if corrective feedback does influence learning in Section 6.1.3, to facilitate understanding of the presented outcomes.

6.1.1 Datapoints

We first of all explored measuring features as an average taken over all available files in one month. However, even with this averaging, a considerable fluctuation of measures such as mean length of utterance (MLU), amount of errors and amount of corrective feedback is observable. The left graph in Figure 6.1 shows this exemplarily for Adam in the Brown corpus (Brown, 1973).[1] The amount of subject omission errors is computed as the fraction of the number children's erroneous utterances divided by the number of all child utterances. The amount of corrective feedback is computed as the number of instances of corrective feedback on a subject omission error divided by the number of child subject omission errors. As the three features (MLU, rate of error, rate of corrective feedback) lie in different ranges, three different sets of labels are specified on the y-axes.

The fluctuations in this graph are caused by the fact that we are looking at samples. The child's development, however, can reasonably be assumed to be comparatively steady. Hence employing a measure which reflects this steadiness is preferable over a measure which results in a high fluctuation. Therefore we computed measures as an average over 3 consecutive available instances each covering information from one month. This averaging procedure assumes that

[1] The graphs showing the development for the other children are presented in Appendix D.1.

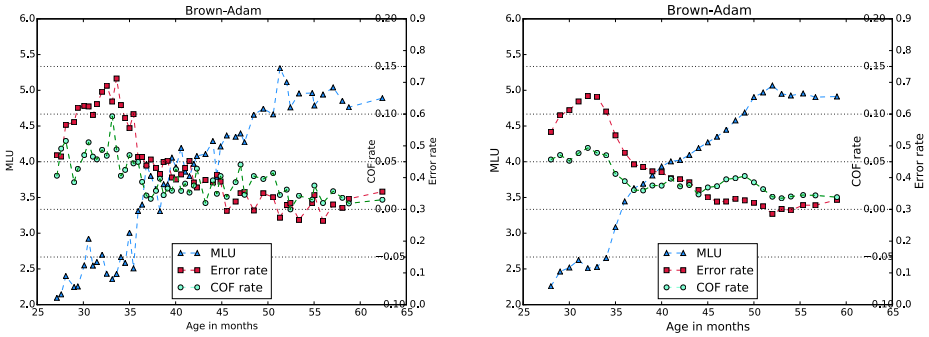


Figure 6.1: Development of MLU, amount of subject omission errors and amount of corrective feedback on subject omission errors of Adam in the Brown corpus when measured in one month (left) or as an average over the data available in three consecutive instances each containing data of one month (right).

over a limited timespan of a few months the child's development is approximately linear.[2] The right graph in Figure 6.1 shows the improved steadiness of the described measuring procedure. In the rest of the chapter, all reported features will be computed using this means of averaging.

Next, we need to define how to measure advances in time. It is well known that children reach certain levels of language proficiency at largely varying ages. Thus, often, children are grouped into classes of similar developmental stages not according to age but according to different measures, such as MLU. Here, however, we do look at the development over time as measured in months of the child's age. This is due to the fact that by considering other possibly predictively meaningful measures we would incorporate many unwanted interdependencies into our predictions. These would be hard to disentangle afterwards.

6.1.2 Experiments

Now that we have specified how to measure the features we are interested in and the evolvement over time, we need to develop which methods to employ to extract predictions from the data. We are interested in the effect of corrective

[2]This is certainly not true over a timespan of several years. But as we average solely over three consecutive instances each covering one month, we hope that this procedure does not introduce too large of a bias.

feedback on learning. Hence we will always look at the development observable over several datapoints $t_0, t_1, \ldots$. In a first step we will test for simple correlation between the amount of corrective feedback available at time t_0 and the improvement achieved by a later time t_1. This will be done for fixed starting ages and a fixed timespan, fixed starting ages and variable timespans, as well as variable starting ages and variable timespans.

A significant relation is observed, so we consequently test the influence of other predictors. That is, we use a linear least squares regression analysis taking features from time t_0 as input to predict the decrease of the frequency of errors between t_0 and t_1.[3] Explanatory power of the featureset is evaluated using the R^2 value. This value tells us how much better the obtained prediction for the target values is than a prediction which always reports the mean for the target. Thus, a model which always predicts the mean of the target value, irrespective of the input features, will have R^2 score 0. Maximal value for R^2 is 1.

We examine how much variance can be explained by adding corrective feedback as predictor versus using only other predictors. In the first comparison we test how predictions using only the amount of observed error as input compare to predictions using amount of error and amount of corrective feedback. Next, we look at how adding corrective feedback as predictor compares to a larger set of features. The additional features are:

1. Child MLU
2. MLU of child-directed speech
3. Percentage of words uttered by the child out of all words in the transcripts
4. Size of child vocabulary
5. Size of vocabulary in child-directed speech

Two featuresets obtained from these are considered: containing or excluding corrective feedback as predictor. The size of the vocabulary is computed as the number of types (different words) divided by the number of tokens (words), so that longer files do not immediately give a higher value. Additionally, it is

[3]The implementation of this algorithm in the python module scikit-learn (Pedregosa et al., 2011) was used for computation.

computed using all observed words up to the given datapoint, instead of only at the given datapoint. We can reasonably assume that words used in a previous transcript are still in the vocabulary later on even if they happen to not occur in the sample.

6.1.3 How to interpret results

It will be useful to think about what possible outcomes imply. To this end, we will go through what the expected outcomes would be if corrective feedback did indeed have a favourable influence on learning. In the first step we are looking at the correlation between the amount of corrective feedback at t_0 and the improvement achieved by a later time t_1. This improvement is measured as the amount of error observable at t_1 minus the amount of error observable at t_0. If learning takes place this value should be negative, as the frequency of the error decreases. More learning is then visible from a lower number (in the directed, not in the absolute sense). If corrective feedback has a positive effect on learning, then *more* corrective feedback should lead to a bigger improvement, i.e. a *lower* number for the difference in the amount of errors. Hence *negative correlation* shows a positive influence of corrective feedback on learning.

As for the linear regression analysis, if corrective feedback assists in learning, then we should be able to obtain a model with more predictive power from using both corrective feedback and the amount of error as input than from using solely the amount of error. Equivalently in the larger featuresets, we should be able to predict more variance with all features than if we remove corrective feedback as a predictor.

6.2 Results

In the present section we will present the results of the analysis described in Section 6.1.

6.2.1 Observations unrelated to learning

The first observation visible in Figure 6.1 is that the amount of corrective feedback decreases over time, even when correcting for the fact that the frequency

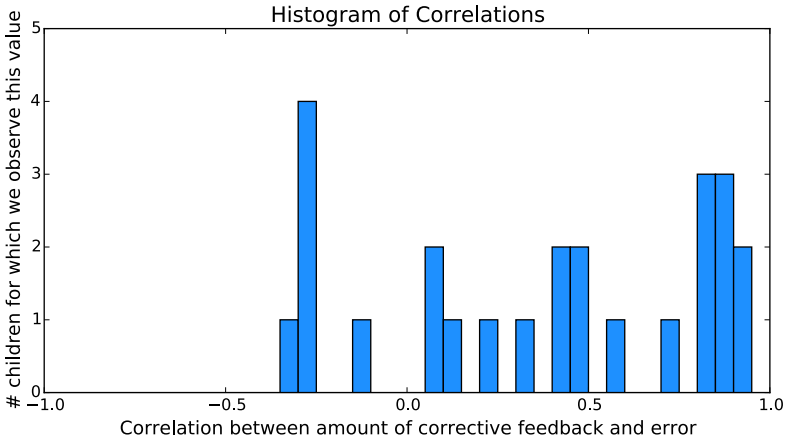


Figure 6.2: Histogram of correlations between the amount of subject omission errors and corrective feedback on subject omission errors for each child.

of errors, and thus the possibility for giving corrective feedback, decreases. In Section 4.4 we saw the same development, but could not yet exclude that it was solely based on the decrease of the amount of errors made by the child. This trend can now be confirmed for most children, for the case of corrective feedback on subject omission errors.

Next, we also see in Figure 6.1 that the amount of corrective feedback and the error frequencies seem to correlate. To test whether this is generally the case we extracted the correlations between these two values for all children. A histogram showing for how many children certain ranges of correlation were observed is presented in Figure 6.2. It is visible that indeed for many children this value is largely positive. Corrective feedback is counted as the fraction of errors which do get corrected; the overall amount of error is factored out. Hence this correlation was not to be expected a priori. Interpreting this fact is not trivial, due to the many interactions happening between the conversational participants.[4]

[4]Just to name two possible interpretations: This relation could be one indicator showing that parents adjust their input according to their child's needs, giving more corrections if they have the impression that their child struggles with a certain structure. But it could also be that corrective feedback, being affirmative due to the repetition, inclines the child to keep making a certain error.

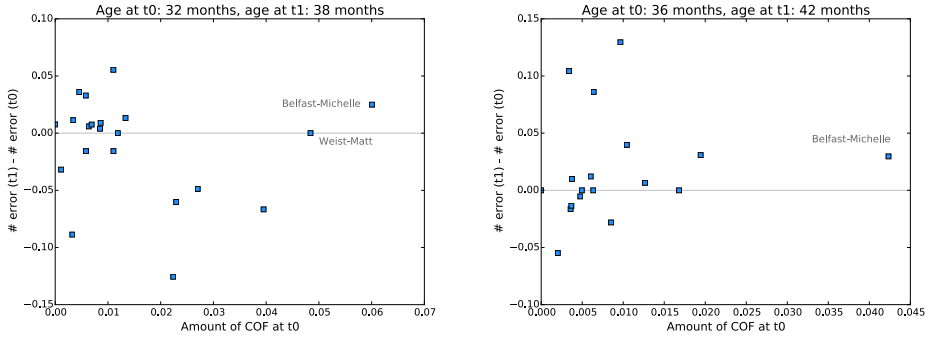


Figure 6.3: Scatter plots showing the amount of corrective feedback at t_0 and the difference in the frequency of the error after 6 months for each child, for starting ages 32 and 36 months. The correlation coefficient for starting age 32 months is -0.15,for starting age 36 months -0.52.

6.2.2 Correlation analysis

In a next step we investigated the relation between the amount of corrective feedback given at a starting point t_0 and the difference in the amount of errors between t_1 and t_0 for a fixed timespan between these two. Figures 6.3 and 6.4 show exemplary outcomes for starting ages 32 and 36 months in the form of scatter plots. One point represents the data taken from one child. These ages were chosen as we have enough data for meaningful predictions roughly between the starting ages of 30 to 38 months. The two selected points give illustrative scores for the development in this range. Figure 6.3 shows the correlation scores if t_1 lies 6 months after t_0. Figure 6.4 shows the relation after a lag of 13 months. Outliers were labeled, to enable subsequent investigation into possible reasons for the observed irregularities. In all four cases a negative correlation is discernible. The correlation between amount of corrective feedback and difference in the amount of error for starting age 32 months and a difference in time between t_0 and t_1 of 6 months is -0.15. For a difference of 13 months it is -0.52. Similarly, for a starting age of 36 months the correlation is -0.52 for a difference of 6 months and -0.65 for a difference of 13 months. Is higher correlation after a bigger time lag a common pattern?

To investigate this we computed the same correlation values for variable timespans covered between t_0 and t_1. The results are depicted in Figure 6.5,

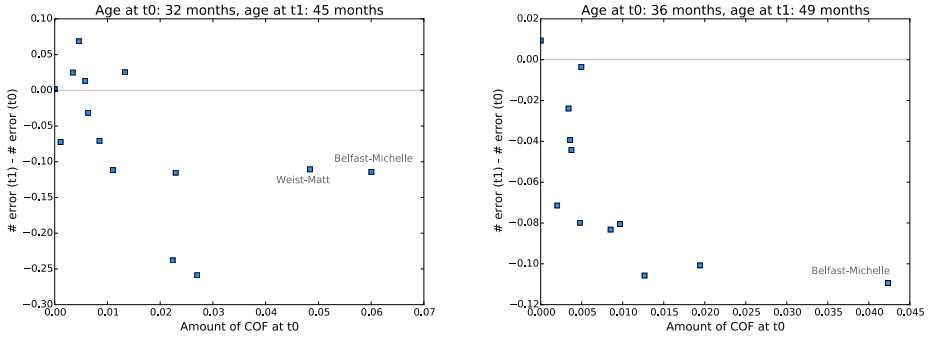


Figure 6.4: Scatter plots showing the amount of corrective feedback at t_0 and the difference in the frequency of the error after 13 months for each child, for starting ages 32 and 36 months. The correlation coefficient for starting age 32 months is -0.52,for starting age 36 months -0.65.

again for fixed starting ages of 32 and 36 months. The corresponding graphs for other starting ages are presented in Appendix D.2. Here we need to take into account that for large differences between start and end age we do not necessarily have data from all children. Thus those values computed using at least 10 different datapoints were marked.

Certain trends are observable, which represent a general pattern. First of all, for younger children and a low age-difference no noteworthy connection between the amount of corrective feedback and the decrease of the amount of error can be established. This changes for older children or somewhat larger differences between start and end age. Here the observed correlation fits in with the theory that corrective feedback assists learning. However, the relation could be caused by many influencing factors. No conclusion other than that the observations do not contradict the hypothesis can be drawn from this. Finally, for a large difference between start and end age of 1.5 to 2 years again no correlation between the amount of feedback in the start and the learning process until the end can be established. This seems reasonable, as many other influences will be available to the child during this large timespan.[5]

[5] Why the onset of an effect for younger children takes so long can only be speculated here. Considering the complexity of the phenomenon of corrective feedback we observed in Section 5.2 younger children might lack the grammatical proficiency to detect it. This idea, as said before, is wholly speculative.

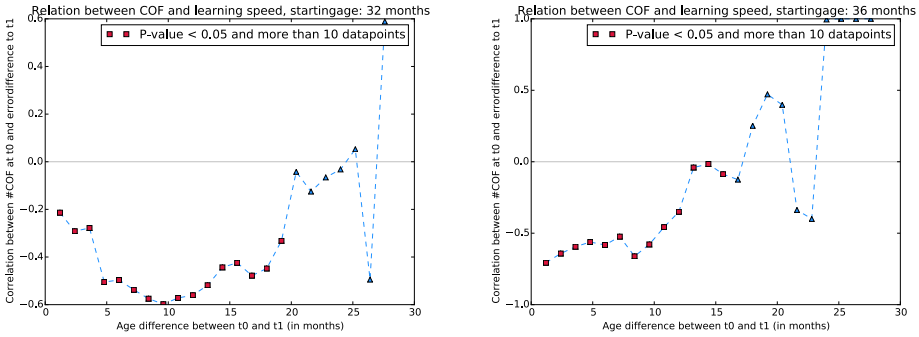


Figure 6.5: Correlation between the amount of corrective feedback at t_0 and the difference in the amount of error between t_0 and t_1 over all children against the timespan between t_0 and t_1, for starting ages 32 and 36 months.

Subsequently we computed the same correlation scores over all starting ages, solely taking into account the difference in ages between t_0 and t_1. These values are presented in Figure 6.6. Here, the information on one timespan can contain data from the same child, with different starting ages. For comparison, we also present the values obtained from extracting corrective feedback instances using a lower-precision classifier. In Section 5.4 we showed that we can obtain higher precision by increasing the penalty for misclassification of non-corrective feedback instances, and decided to use this classifier. The graph presented on the left of Figure 6.6 gives the values obtained from classification with equal class weights. The graph on the right was obtained using the higher-precision classification method. We see that the latter classifier, which as described is preferable, results in a more pronounced relation between corrective feedback and learning. Additionally, we again see that this observable correlation between corrective feedback and reduction in error increases for longer timespans between the two observation points. It peaks after a lag of about one year, and subsequently decreases again.

6.2.3 Linear regression analysis

Now that we have established the presence of a significant correlation between the amount of corrective feedback received at t_0 and the difference in the amount of error between t_0 and t_1, for various t_0 and t_1, we need to investigate whether

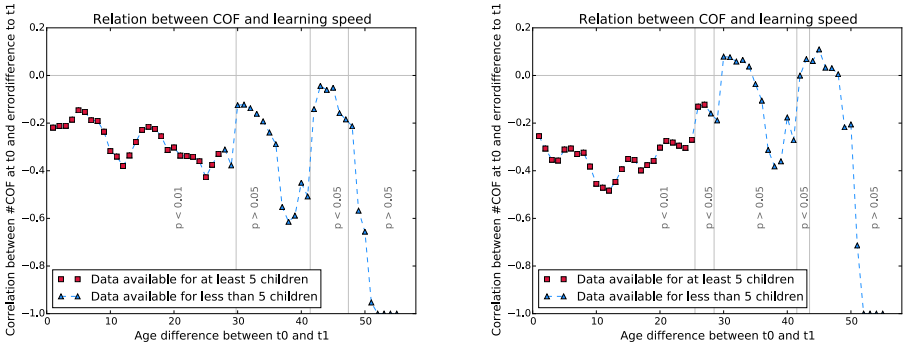


Figure 6.6: Correlation between the amount of corrective feedback at t_0 and the difference in the amount of error between t_0 and t_1 over all children against the timespan between t_0 and t_1, disregarding the starting age (i.e. taking all starting ages into account). On the left are the scores when classification of corrective feedback is completed using a lower-precision classifier (0.82), on the right are the scores according to classification using the higher-precision classifier (0.89) employed also in all other extractions.

the explanatory power of corrective feedback exceeds that of other predictors. This is where we employ the described linear regression analysis. In a first step, the explanatory power of a model obtained from using as input only the amount of error at t_0 is compared to the one obtained from using both corrective feedback and the amount of error as input. The results for starting ages 32 and 36 months, for varying timespans between t_0 and t_1, are given in Figure 6.7.[6] We see that overall the predictions for the improvement after a short timespan are not reliable.[7] For a time lag of less than 9 months the prediction does not at all improve by adding corrective feedback as an input feature. However, after a bigger time difference between t_0 and t_1 adding corrective feedback as a feature does indeed account for more variance in the target value. Additionally, the above observation is confirmed: explanatory power of corrective feedback peaks after a lag of a little over one year, with the difference in time being slightly

[6]The graphs for other relevant starting ages are presented in Appendix D.3.

[7]If we predict solely the amount of error at t_1 this changes considerably; the amount of errors made immediately after t_0 can be predicted very well. Considering that we take the amount of error at t_0 as input feature this should not be surprising, assuming a somewhat steady development.

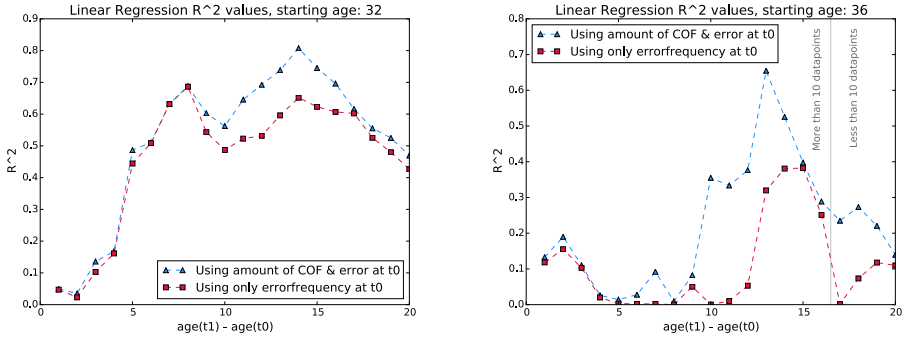


Figure 6.7: R^2 values for linear regression predicting the amount of error at t_1 using different featuresets, against the difference in age between t_0 and t_1, for starting ages 32 and 36 months. All depicted scores for start age 32 months were obtained using more than 10 datapoints.

higher for younger children.

Next we evaluated the change in predictive strength of the model if corrective feedback is added to the larger set of features described in Section 6.1. The observed R^2 scores for starting ages 32 and 36 months are presented in Figure 6.8.[8] Additionally, the results obtained from using only corrective feedback and the amount of error are depicted again here. The same observations as before are visible. Predictions of the decrease of the amount of error after a short timespan are very unreliable, even with this larger set of features. Again, up until a lag of 9 months, adding corrective feedback as a predictor does not increase the explanatory power of our model. However, after this timespan corrective feedback is able to account for a variance in the difference of error above what can be explained with the other features. Another previously observed characteristic which is confirmed here is that explanatory power of corrective feedback peaks after a lag of over one year, with the difference in time being slightly higher for younger children.

[8] Again, results for other starting ages are given in the Appendix D.3.

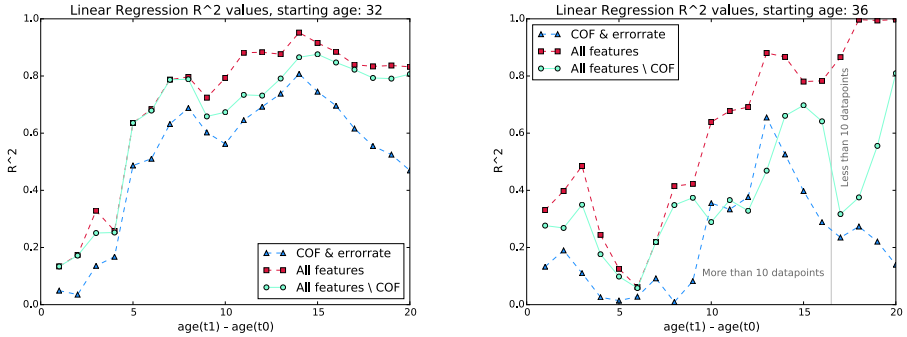


Figure 6.8: R^2 values for linear regression predicting the amount of error at t_1 using different featuresets, against the difference in age between t_0 and t_1, for starting ages 32 and 36 months. All depicted scores for start age 32 months were obtained using more than 10 datapoints.

6.3 Discussion

We observed that the amount of corrective feedback given on subject omission errors at time t_0 negatively correlates with the difference in the amount of subject omission errors made at time t_1 versus t_0, for many values of t_0 and t_1. This is in accordance with the hypothesis that corrective feedback on a given error assists the child in learning the corresponding structure. However, no information as to causation can be deduced from this correlation value. For younger children, correlation sets in after a lag of around 5 months between t_0 and t_1. This conforms to the findings of Saxton et al. (2005), who showed that on subject error locations no influence of corrective feedback was discernible after a lag of 12 weeks in a sample of children with mean starting age 2;0. For both younger and older children the correlation observed here peaks in intensity after a timespan of a little over one year between t_0 and t_1 and then starts decreasing. This finding was confirmed by the linear regression analysis; R^2 values of the linear regression using corrective feedback as predictor reach their highest point for the same timespan covered between t_1 and t_2. This is a long time, especially for developing children. No immediate justification for this development is apparent.

Finally, the linear regression analysis also showed that models obtained from

adding corrective feedback as a predictor to other sets of features can account for more variance in the difference between rates of error than if corrective feedback is not included as a feature. This development only sets in after a lag of 9 months, for both considered sets of additional features.

Overall, the presented results indicate that corrective feedback on subject omission errors is indeed helpful for the child learning to correctly include subjects.

6.4 Summary of the Chapter

In Section 6.1.1 we first of all developed means for measuring relevant features in the child's development. The fact that every transcript presents only a sample of the child's abilities at that time, leading to fluctuations in the values for naturally rather steady measures, was corrected for by taking average values over consecutive available months. Subsequently, in Section 6.1.2 we decided which analyses to use to obtain predictions from the features measured in the data. We first of all a conducted a simple correlation analysis between the amount of corrective feedback at a starting point and the decrease in error until a later point in time. Significant correlation was visible in the direction which would be expected if corrective feedback facilitates learning. The highest correlation was observed for lags of about one year. This is striking, as one year in the life of a language acquiring child is a lot of time.

After significant correlation was detected, the explanatory power of corrective feedback in relation to other predictors was investigated using a linear regression analysis. First of all the above finding was confirmed: corrective feedback is most influential after a lag of a little over one year. Additionally, after a lag of 9 months adding corrective feedback as a feature results in higher scoring models, for both sets of additional features that were considered. This indicates that corrective feedback on subject omission errors does facilitate learning to include subjects.

Chapter 7

Conclusion

7.1 Summary

The present study investigates parental reformulations as a possible source of *negative input* for language learning children. This analysis is empirically based and incorporates comparatively large amounts of data. Computational methods were developed to enable the processing of this data.

We first of all defined what exactly constitutes instances of corrective feedback: an exchange containing an erroneous child utterance, followed by an adult response which repeats part of the child utterance but modifies it to give a correction of the error. Subsequently we devised a more finegrained taxonomy of instances of corrective feedback based on the linguistic level and the type of error observed in the child utterance and corrected by the adult interlocutor. Following this we selected the appropriate data to be used as input from the English section of the CHILDES database (MacWhinney, 2000a). We equipped the selected transcripts with the necessary additional information to be able to extract informative features. This included manually annotating a selection of candidate corrective feedback exchanges for the presence and specific instantiation of corrective feedback.

Next, we attempted to develop an automatic classifier to detect all instances of corrective feedback, using the subset of manually annotated exchanges as a training set. The obtained predictive performance showed that the general phenomenon is too diverse for the selected classification method. Hence the phenomenon to be investigated was limited to a single category of the distin-

guished instantiations of corrective feedback: reformulations of a child's subject omission error. We established reliable classification algorithms extracting both child utterances containing subject omission errors and utterance pairs containing corrective reformulations of such errors.

Finally these extraction mechanisms were applied to the whole dataset to analyse the effect of corrective feedback on the acquisition of subject inclusion. We investigated the relation between the amount of corrective feedback on a subject omission error at a certain time t_0 and the difference in the amount of error made at a later stage t_1 compared to the starting age t_0. In a first step a correlation analysis showed that more corrective feedback was related to a higher decrease of the amount of error after a certain period. This relation is strongest after a lag of a little over one year, and indicates that corrective feedback does positively influence learning. These observations were confirmed by a linear regression analysis. Additionally, the latter analysis also showed that corrective feedback can predict variance in the difference of the amount of error above what can be predicted from other features. This effect is visible after a timespan of at least 9 and and most 17 months in between the analysed points t_0 and t_1. Thus, overall, the observed results suggest that corrective feedback facilitates the learning process of subject inclusion. What is the significance of this in the initial setting in which we asked the question, namely, the nature versus nurture debate? If corrective feedback facilitates learning, then this might be due to the fact that children do perceive it as a correction, and thus as negative input. Consequently, this diminishes the support for the *no negative feedback hypothesis*, used to justify the *poverty of the stimulus* argument brought up on the "nature" side of the debate. However, the hypothesis is not disproven, as we solely showed that corrective feedback has a positive effect on language learning, which might be but is not necessarily caused by the child perceiving corrective feedback as negative input. Additionally, our investigation was limited to subject omission errors and the acquisition of subject inclusion. Other kinds of child errors might reveal different outcomes.

7.2 A Note of Caution

The obtained results have to be interpreted with caution. First of all, the general observations concerning significance and correlation need to be considered. Namely, the paired t-test employed to check for statistical significance is two-sided. All that can be derived from the p-value is that the observed results would be very unlikely if the observed variables were indeed independent. The correlation gives an indication in which direction the dependency between them goes in the observed sample. However, this does not constitute evidence that we can assume the dependency to be in the observed direction with the certainty provided by the p-value.

More specifically related to the present investigation, the results have to be considered with prudence as the employed classification methods are non-ideal. We had to settle on good, but not perfect, extraction methods for both subject omission errors and reformulations correcting subject omission errors. All results are based on classifications done using these tools, and are therefore imperfect by extension.

7.3 Future Work

The first question that comes to mind when looking at the presented analysis is why the onset of the effect of corrective feedback takes so long. It is not immediately evident, however, how this question could be answered fully.

More apparent extensions of the presented work are to investigate corrective feedback on other levels and types of child errors, to test whether the observed relations are a general trend or error specific. Additionally, other languages can be analysed to examine possible differences. Examining second language learners could help in establishing whether the observed results are related to the general cognitive development of the child.

Finally, instead of using correlation and linear regression analysis to investigate the effect of corrective feedback on the *difference* in the amount of error made at certain moments it would be interesting to employ an analysis that is able of discerning how *fast* the corresponding structure is learned. Event history analysis (for example employed in this way by Tamis-LeMonda et al. (2001))

enables this. Using this procedure one can examine whether corrective feedback is related to a faster decrease of the amount of error made, as compared to the overall mean of the development. As a conclusion, the observed results indicate that corrective feedback does indeed have a positive effect on the acquisition of subject inclusion after a lag of at least 9 months, but more investigation is needed to discern whether this also holds for other kinds of errors, and whether the observed lag is caused by the general cognitive development of the child or related to other reasons.

Appendices

Appendix A

Gold's Proof

In Section 2.3.1 we discussed the reasoning behind two proofs for the necessity of negative input during language acquisition. Here we will give details of the proof presented by Gold (1967), concerning a formalised model of language learnability. Given a definition of a language learnability model for Turing machines he shows that the class of grammars to which natural languages are often taken to belong cannot be learned from only a *text*, that is, from what we described as positive input. As this proof is very formal it will first of all be necessary to define the terms used in it before proceeding to explain the ideas in the proof.

Definition 2 *An* alphabet A *is a set of symbols.*

Throughout the rest, A will be considered fixed and finite. We can imagine it as containing words or letters. ΣA is the the set of all finite strings over A.

Definition 3 *A* language L *is a subset of* ΣA.

Thus, those sentences in L are the grammatical ones, those not in it are ungrammatical.

Definition 4 *A* language learnability model *consists of the following triple:*

1. *A* definition of learnability
2. *A* method of information presentation

3. A naming relation*, which assigns names (perhaps more than one) to languages*

Only one definition of learnability is considered: identifiability in the limit. Time is separated into steps, which are enumerated by the natural numbers. At each time step $t_1, t_2, \ldots$ the learner is presented with information $i_1, i_2, \ldots$ according to the method of information presentation. The learner is a function G that will guess a name - according to the naming relation - of the presented language depending on the information so far available:

$$g_t = G(i_1, \ldots, i_t)$$

Identifiability in the limit applies to classes of languages. A class of languages is identifiable in the limit if there is an algorithm for G such that, for every language in the class and for every order in which the information is presented, the corresponding guesses are all the same after a finite number of steps, and they are correct.

Gold considers a set of different methods of information presentation. We will only differentiate two main classes here: *text* and *informant*. Information presentation by text is what we described as positive input. It is a sequence of strings $x_1, x_2, \ldots$ from L such that every string of L occurs at least once. Depending on the function from $\mathbb{N}$ to $\overline{x}$ the text is called *arbitrary, recursive* or *primitive recursive*. An informant gives information concerning both the grammatical but also the ungrammatical sentences. Whether this takes place via a list of all possible sentences together with their grammaticality value or via responses to queries from the learner does not affect learnability. A learner presented with either of these two possibilities can model the other.

Two possible naming relations for a language are differentiated: *tester* or *generator*. A generator naming relation will give a way of generating all strings in the language; a tester naming relation will give a way of testing whether a given string is in the language. Both times the way this is achieved is by naming a Turing machine which accomplishes the generating or testing task.

Now the important theorem is the following:

Theorem 1 *Using information presentation by primitive recursive text and the tester naming relation, any class of languages which contains all finite languages and at least one infinite language L* IS NOT *identifiable in the limit.*

The proof of this theorem proceeds by assuming a guessing algorithm for the class of languages containing all finite languages and at least one infinite one and constructing an input text such that the language guessed from this text will change an infinite number of times. That is, the corresponding language is not identified in the limit, contradicting the assumption that there is an appropriate guessing algorithm in the first place.

Note that as primitive recursive text is a subset of both recursive and arbitrary text this theorem implies that the same result holds for those two methods of information presentation. If the constructed counter example is available from the set of primitive recursive text it is also available for arbitrary or recursive text. Second, if a class of languages is not identifiable in the limit then neither are any classes it is contained in. This is due to the fact that identifiability in the limit is a notion that applies to a class of languages if every language in it has the desired property. Third, the tester naming relation is more informative than the generator naming relation, thus in this respect the proof does not extend.

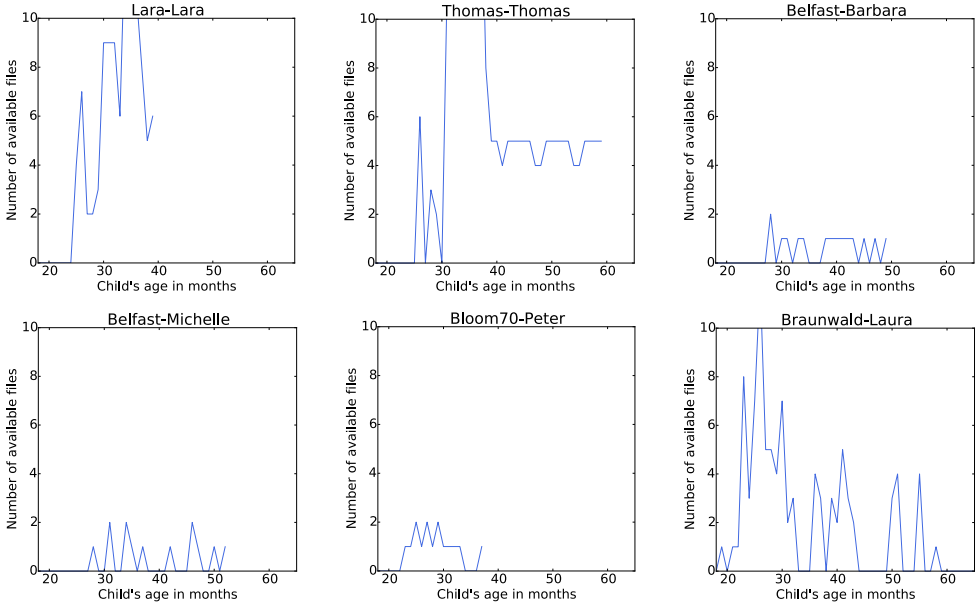
To extend this property of formal language learnability to natural language acquisition a few additional assumptions are needed. First, that the class of natural languages is a superset of the set containing all finite languages and at least one infinite one. Second, that children learn a tester for their corresponding language. Third, that identifiability in the limit is a model of learnability for natural languages. Granting this, the above proof implies that either negative input is necessary or that the class of all possible languages is a proper subset of the set here described - that is, the search space is initially limited by a universal grammar.

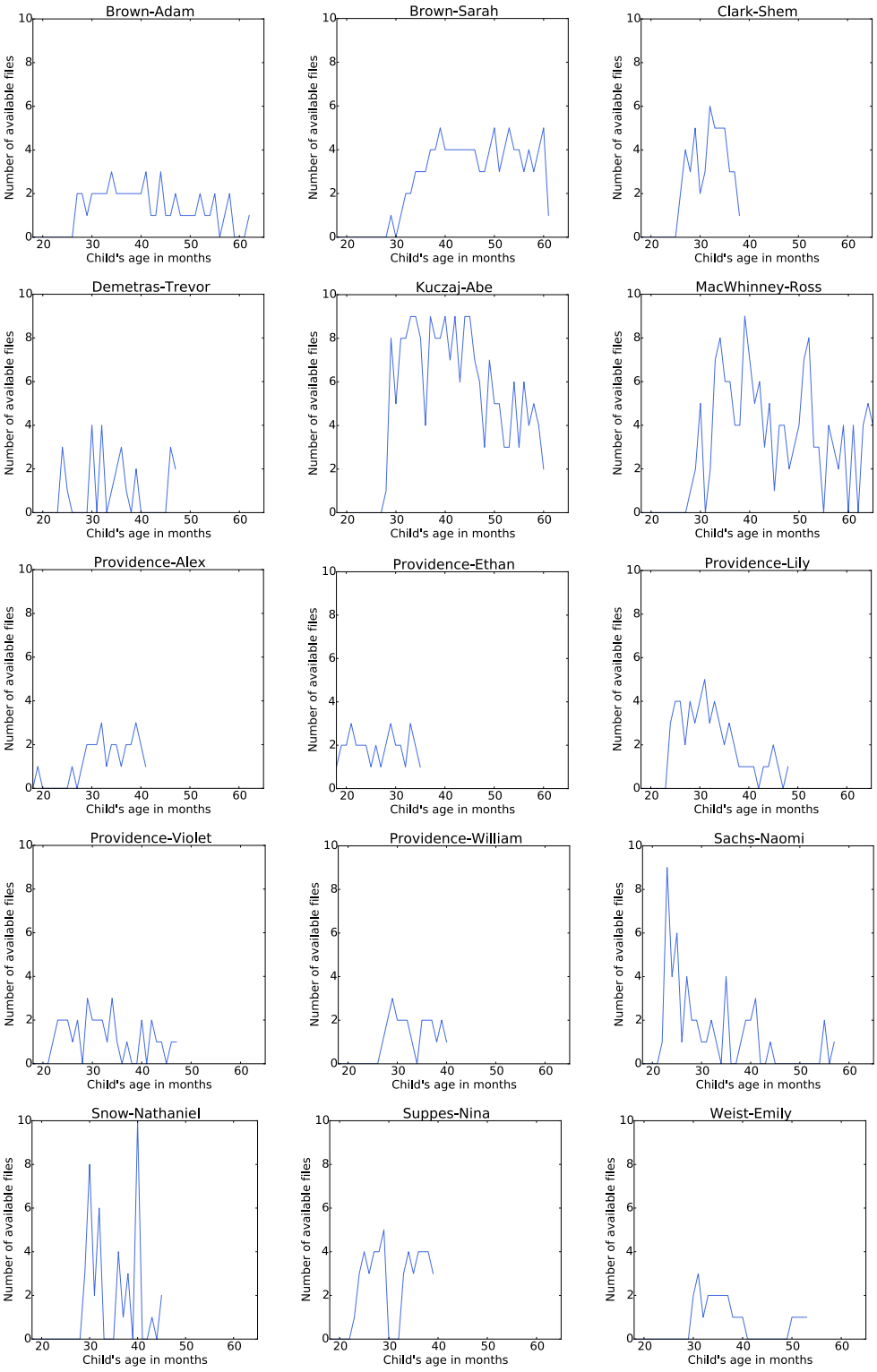
Appendix B

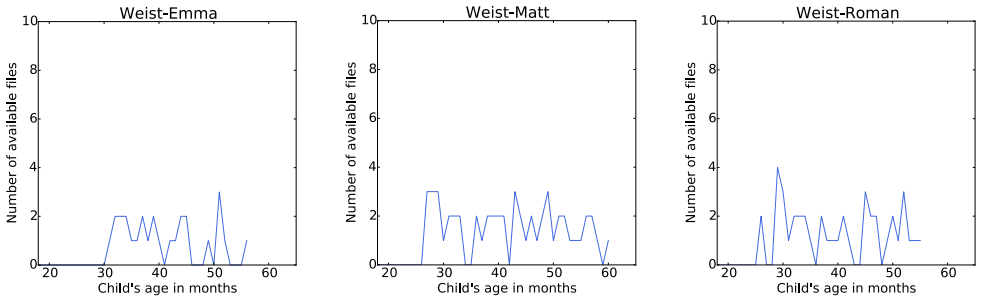
Data Selection

B.1 File Density

In the following graphs the number of available files is plotted against the child's age (in months).




Brown-Adam
Brown-Sarah
Clark-Shem
Demetras-Trevor
Kuczaj-Abe
MacWhinney-Ross
Providence-Alex
Providence-Ethan
Providence-Lily
Providence-Violet
Providence-William
Sachs-Naomi
Snow-Nathaniel
Suppes-Nina
Weist-Emily
Number of available files
Child's age in months



B.2 Selected Files

Table B.1 lists all the corpora and children in the CHILDES database from which files were used for analysis after the extensive preselection described in Section 4.

B.3 Stopwords

Table B.2 gives the empirically derived stopwords which were used to exclude child-adult exchanges from annotation that contain maximally two words of overlap both inside the list.

Corpus	Reference	Child
Lara	Rowland and Fletcher (2006)	Lara
Thomas	Lieven et al. (2009)	Thomas
Belfast	Henry (1995)	Barbara
	Wilson and Henry (1998)	Michelle
Bloom 1970	Bloom (1970)	Peter
Braunwald	Braunwald (1971)	Laura
Brown	Brown (1973)	Adam
		Sarah
Clark	Clark (1978)	Shem
Demetras	Demetras (1989)	Trevor
Kuczaj	Kuczaj (1977)	Abe
MacWhinney	MacWhinney (2000b)	Ross
Providence	Demuth et al. (2006)	Alex
		Ethan
		Lily
		Naima
		Violet
		William
Sachs	Sachs (1983)	Naomi
Snow	MacWhinney (2000b)	Nathaniel
Suppes	Suppes (1974)	Nina
Weist	Weist et al. (2009)	Emily
	Weist and Zevenbergen (2008)	Emma
		Matt
		Roman

Table B.1: Corpora, references and (nick-) names of the children from which files were used for the final analysis.

you	s	the	I	it	a	that	and	to
is	no	in	oh	we	this	on	yeah	there
of	your	okay	they	here	yes	right	not	my
for	she	with	now	up	just	well	at	so
then	because	out	them	down	good	big	if	her
too	his	t	he	me	but	about		

Table B.2: Empirically derived stopwords in the English language.

Appendix C

COF Features

In the present section, the features used in the automatic classification of general corrective feedback versus non-corrective exchanges are explained in detail.

C.1 Semantic Similarity

One of the features we want to extract from the exchanges in our dataset is the semantic similarity between the child and adult utterances. In order to get to this, we extracted a distributed vector representation of both utterances and consequently measured similarity in two ways: via cosine distance, the standard measure used for similarity between semantic vector representations which depends on the angle between the two vectors, and additionally via euclidean distance, to take into account also the spatial distance. For obtaining a vector representation of the sentences in the utterances we need a vector representation of the words occurring in them as well as a combination function for obtaining a representation of sentences from the representations of words.

For the representations of words a pre-trained set of vectors available from `https://code.google.com/p/word2vec/` was used. It contains 300-dimensional distributed vectors for 3 million words and phrases and was trained using the neural network implementation word2vec on part of the Google News dataset (about 100 billion words). Unlike more simple representations, such as for example a one-in-$|V|$ encoding, where $|V|$ is the vocabulary size, these distributed representations capture important semantic features such as similarity between words. Furthermore, the obtained vectors are much smaller in size and not

sparse. The word2vec model is a simplification of the neural network language model presented in Bengio et al. (2003). This latter approach primarily aims at optimising a language model, that is, a probability distribution for predicting the next word given the previous words in a sentence. Distributed representations for words which capture similarities between these are used as a tool to be able to generalise to unseen data. The standard N-gram models, for example, are unable to give the appropriate high probability to an unseen sequence which is obtained from a known sequence simply by replacing one word with another very similar one, such as *cat* with *dog*. To overcome this, Bengio et al. (2003) employ a feed-forward neural network with an input, projection, hidden and output layer and train simultaneously word representations and the language model. While resulting in good predictions this approach is computationally very expensive. Therefore Mikolov et al. (2013a) reduce model complexity by removing the hidden layer from the neural network and use this framework to first of all extract only the word vector representations. Consequently these representations can be used as input into a more complex neural network to obtain a language model. However, this last step is not necessary for our current purposes. As for the training of the word vectors using the simplified model, two different structures of the model are possible: either the current word is predicted given the surrounding words, or the context is predicted given the current word. The latter yields much better results on semantic similarity tasks for the resulting vectors. As an extension to the described model, Mikolov et al. (2013b) extract not only representations for single words but also for phrases. That is, standard expressions made up from several words for which the meaning cannot be obtained by combining the meanings of its compounds. An example is the name of the newspaper *Boston Globe*. Thus, overall, a distributed representation for words and phrases, which captures semantic as well as syntactic features is obtained.

Having acquired a vector representation for single words we next need a composition function for computing sentence representations from these. Despite being clearly inaccurate from a formal semantics viewpoint, due to for example commutativity, vector addition has proven a good approximation for combining meanings (Mikolov et al., 2013b). Therefore this simple procedure

was employed here. In total, two features, the cosine distance and the euclidean distance between the vector representations of the child and adult utterances, were extracted to represent semantic similarity.[1]

C.2 Syntactic Similarity

Another feature we want to represent to the classification algorithm is the syntactic similarity between the child and adult utterances. From the application of MEGRASP we already have a dependency tree representation of the two utterances, between which we can measure similarity. The standard measure for similarity between trees is the tree edit distance, which has as its sole deficiency the computational cost it involves. As the sentences in our dataset are mostly very short this does not pose a problem here. Hence this procedure was chosen. Zhang and Shasha (1989) present an algorithm for computing tree edit distance, which is implemented in the python module zss and was used here. Three different kinds of operations are distinguished: adding a node, deleting a node and renaming a node. The former two received equal cost of 1. The latter, renaming of a node, has as its default cost the string edit distance, but received zero cost in our implementation. Thus only the structure of the trees is compared. To understand why this represents the information adequately recall that the output of MEGRASP is structured in triples $i|j|R$, where i and j give the *indices* of the corresponding words in the utterance. Hence the actual words are not present in our trees. Now consider the following exchange, the trees for which are depicted in figure C.1.

[1] Words for which no semantic vector is available, such as for example certain function words, were ignored in the computation of the meaning of the whole utterance. In one single case this resulted in the null vector for the whole sentence, as none of the words was in the dictionary. Consequently, no distance between this vector and another one can be computed. The utterance in question is **CHI: pretendy@c pretenty@c pengy@c pengy@c .*, which is indeed incomprehensible. It was therefore decided to exclude the exchange in question from the analysis.

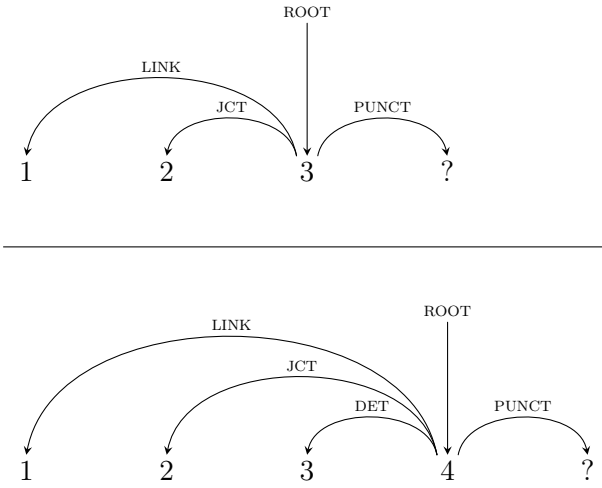


Figure C.1: The dependency tree representations of the sentences “What about kiss ?" (top) and “What about a kiss ?" (bottom)

(1) *CHI: what about kiss ?
%gra: 1|3|LINK 2|3|JCT 3|0|INCROOT 4|3|PUNCT
*DAD: what about a kiss ?
%gra: 1|4|LINK 2|4|JCT 3|4|DET 4|0|INCROOT 5|4|PUNCT

Due to the insertion of a determiner inside the sentence the index for *kiss* switches from 3 to 4. However, this should not be counted as an editing move, which is why the cost for renaming nodes was set to 0. Overall, one feature was extracted to represent syntactic similarity between the child and parent utterance: the edit distance between the bare structure of the dependency trees corresponding to the two sentences. Further syntactic features concerning the actual words in the utterances were extracted with regard to whether these words were added in the adult utterance, deleted from the child utterance, or matched exactly between both. These features are described in the next section.

C.3 Features related to CHIP output

As a final set of features which we extract from the utterance pairs there is a lot of information related to the output of the CHIP program. First of all,

whether the adult utterance is an extension or a reduction of the child utterance, which are both represented as binary variables. Next, the fraction of the words in the parent utterance which are a repetition of ones in the child utterance, which is computed by the program and represented as a floating point number between 0 and 1. The extreme cases will not occur as we filtered out those exchanges in a preliminary step, but this is of no importance and the value still gives meaningful information. Finally the tier created by the CHIP program also contains lists of the words which were added, deleted and which matched exactly between the child and parent utterance. For the first two we mirrored the value already calculated by the program for exactly matching words and computed the fraction of added (deleted) words relative to the length of the adult (child) utterance.

Consequently we wanted to represent more information about the specific words which were added, deleted and exactly matched. We considered a representation using 1-in-$|V|$ encoding, where $|V|$ is the vocabulary size, as too finegrained and instead chose to represent only the part of speech tags in the same way. Thus for each of the three sets of words a separate 40-dimensional or 61-dimensional vector was extracted, using either the rough differentiation of superclasses of parts of speech tags or the more finegrained one including all differentiations. Each index received an entry with the number of times the corresponding part of speech tag occurred in the current set of words. Alternatively, this count was binarised to a simple yes/no occurrence information. For the added and exactly matched words these were taken from the morphological annotation following the adult utterance, and for the deleted words from the morphological annotation following the child utterance. In total, this leads to four different sets of features which were extracted: representing rough or detailed part of speech tags, and binarised or non-binarised counts.

Finally, in the computation of the edit distance between the dependency trees of the two utterances, which is described in Section C.2, the actual relations are not taken into account. Thus another feature was added to include these. Overall there are 35 different dependency relations, so two 35-dimensional vectors representing the occurring relations were extracted, one for the added and one for deleted words. In these two sets it is clear from which utterance

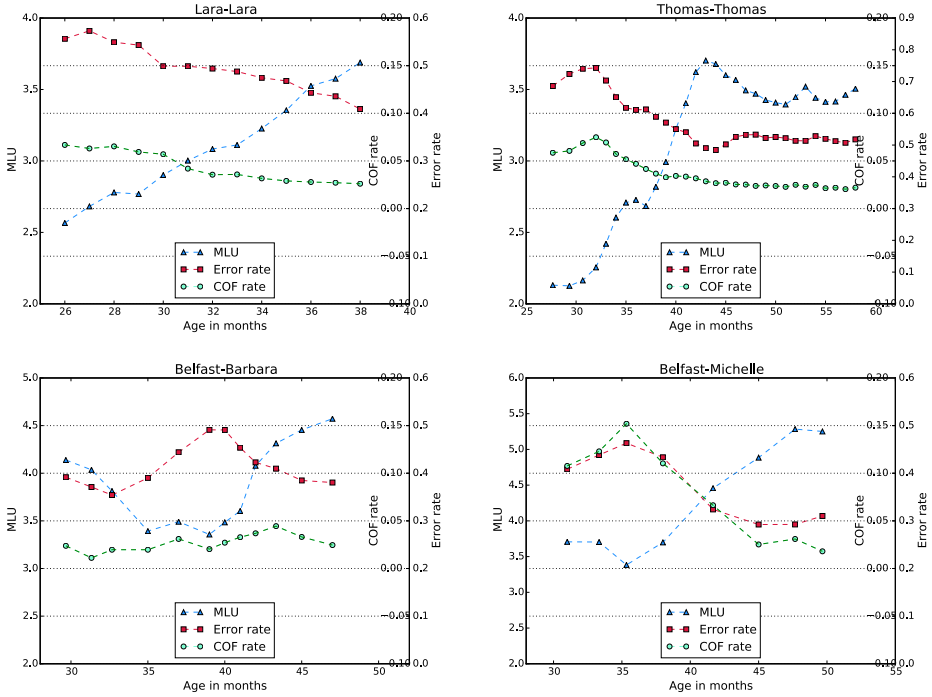
to take this information. For those words matching exactly the case is more complicated, as they occur in both utterances with possibly different syntactic relations. We decided to extract those relations which differ for the two utterances. That is, a word occurring in both statements is involved in this relation in one sentence but not in the other. Using positive and negative values to signify different directions it was possible to represent this in the same vector. This gives a third 35-dimensional vector. Again, the numbers were represented either as occurrence counts or as binarised yes/no entries, depending on the representation which the part of speech tags received.

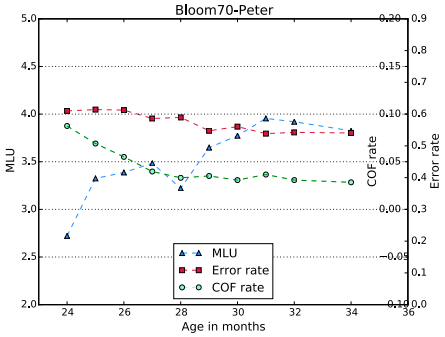
Appendix D

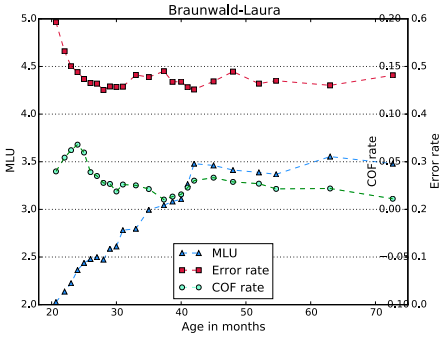
Experimental Investigation

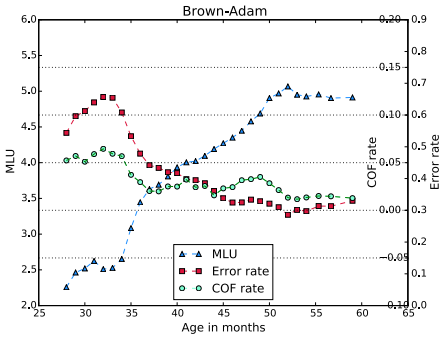
D.1 Preparation

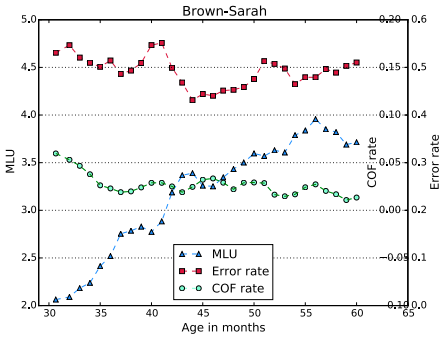
In the present section we present the graphs showing the develpment of MLU, errorrate and rate of corrective feedback for all children. The values were computed using the averaging described in Section 6.1.1.

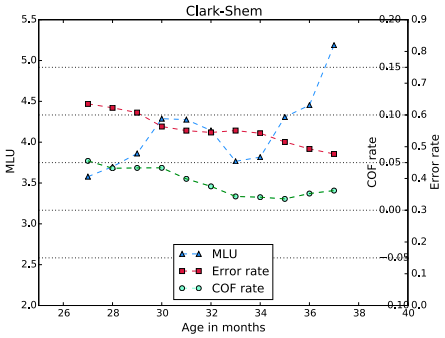


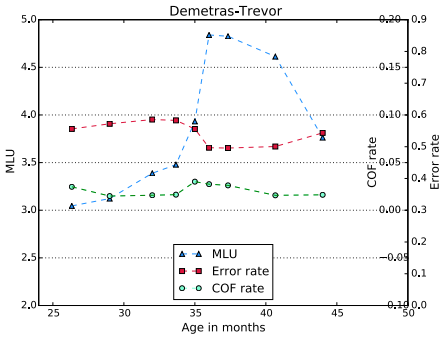

Bloom70-Peter
MLU
COF rate
Error rate
Age in months
MLU
Error rate
COF rate

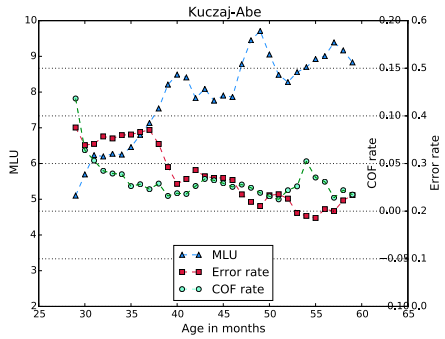

Braunwald-Laura
MLU
COF rate
Error rate
Age in months
MLU
Error rate
COF rate

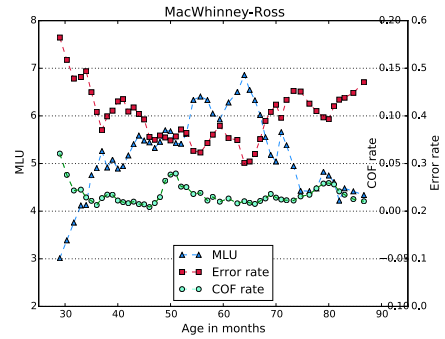

Brown-Adam
MLU
COF rate
Error rate
Age in months
MLU
Error rate
COF rate

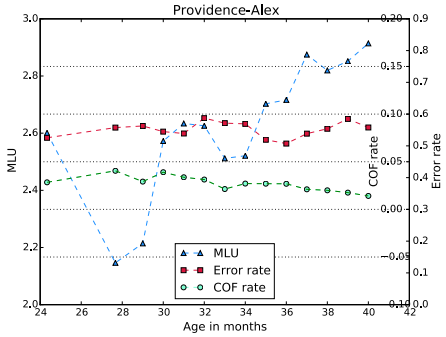

Brown-Sarah
MLU
COF rate
Error rate
Age in months
MLU
Error rate
COF rate

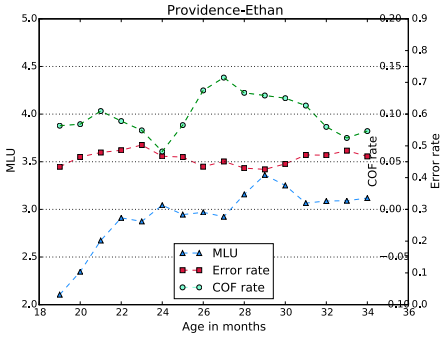

Clark-Shem
MLU
COF rate
Error rate
Age in months
MLU
Error rate
COF rate

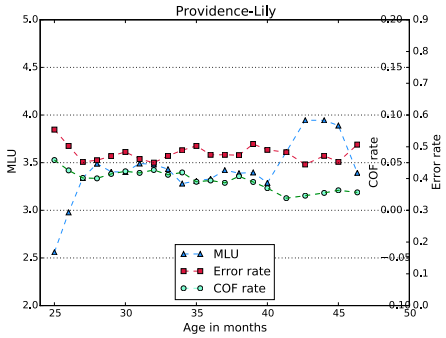

Demetras-Trevor
MLU
COF rate
Error rate
Age in months
MLU
Error rate
COF rate

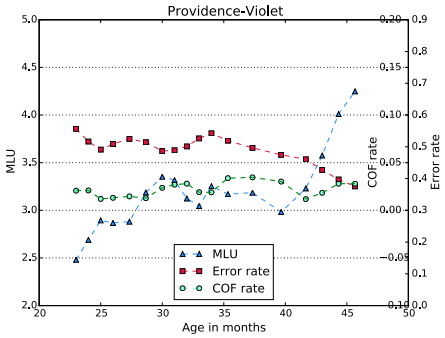

Kuczaj-Abe
MLU
COF rate
Error rate
Age in months
MLU
Error rate
COF rate

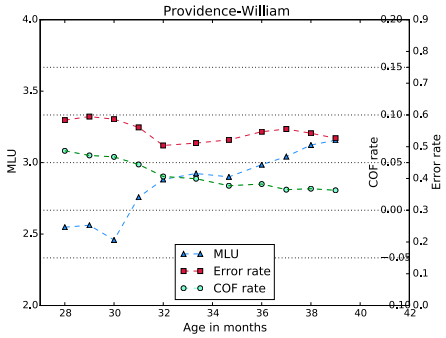

MacWhinney-Ross
MLU
COF rate
Error rate
Age in months
MLU
Error rate
COF rate

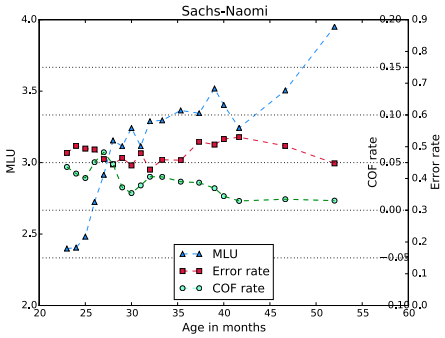

Providence-Alex
MLU
Error rate
COF rate
Age in months

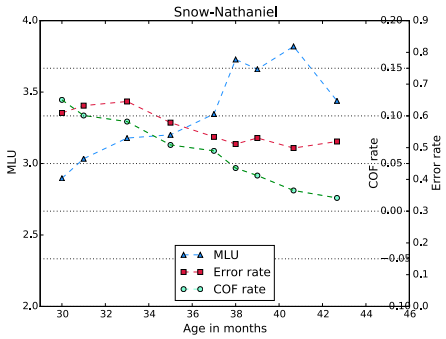

Providence-Ethan
MLU
Error rate
COF rate
Age in months

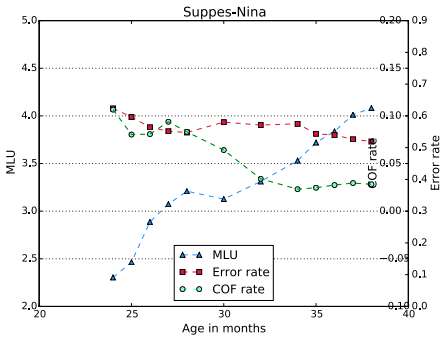

Providence-Lily
MLU
Error rate
COF rate
Age in months

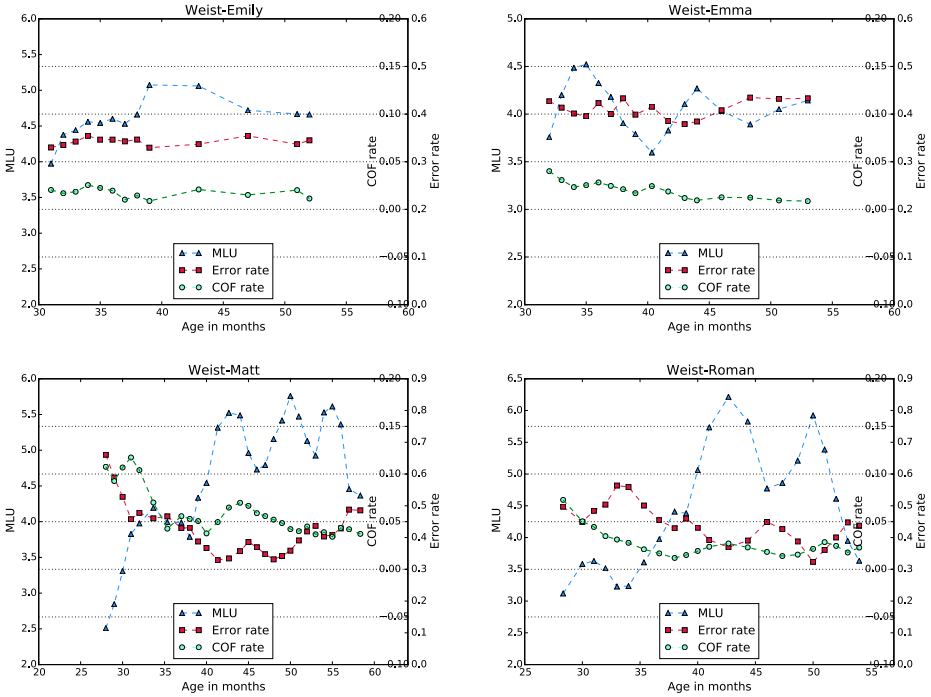

Providence-Violet
MLU
Error rate
COF rate
Age in months


Providence-William
MLU
Error rate
COF rate
Age in months


Sachs-Naomi
MLU
Error rate
COF rate
Age in months

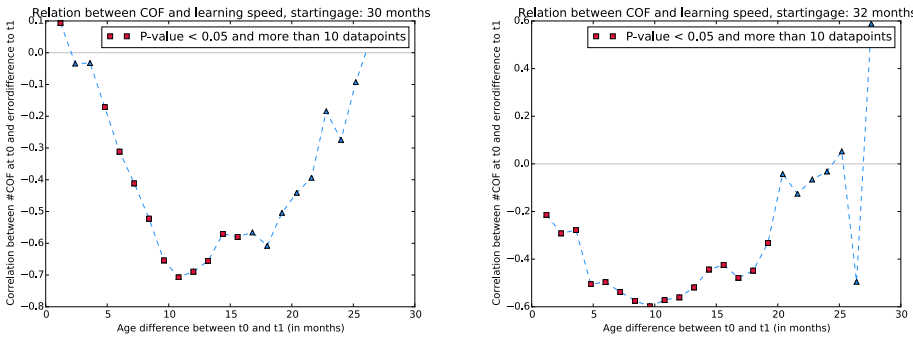

Snow-Nathaniel
MLU
Error rate
COF rate
Age in months

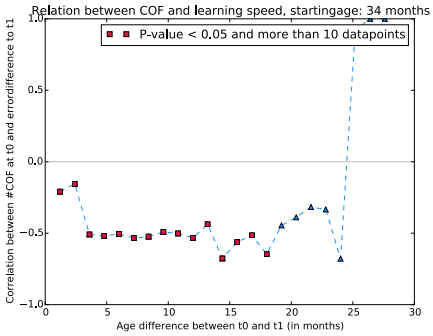

Suppes-Nina
MLU
Error rate
COF rate
Age in months

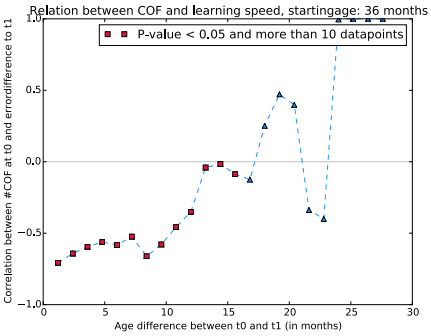


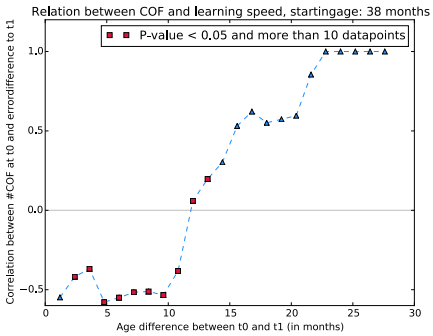
D.2 Correlation Analysis

In the following are the graphs showing correlation between the amount of corrective feedback at t_0 and decrease of error between t_0 and t_1 for varying agedifferences, for all startingages. Note that for very low or very high starting ages not enough data is available to make significant conclusions. Thus only startingages for which sufficient data was available are depicted.



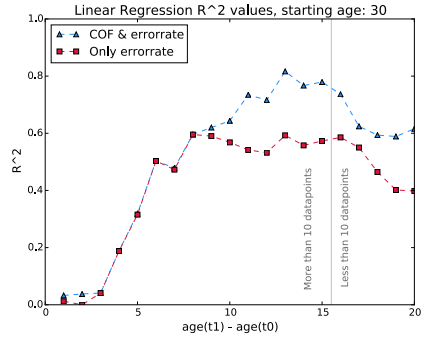


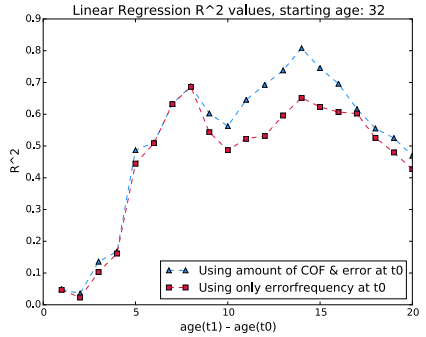


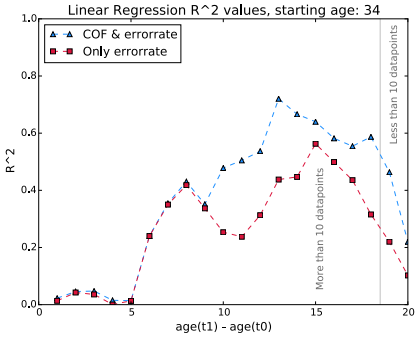


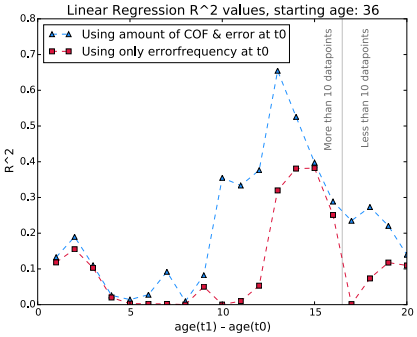
D.3 Linear Regression

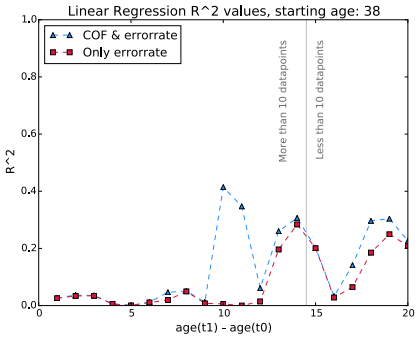
In the following we present the R^2 scores obtained in the linear regression analysis for varying sets of features and varying starting ages. The precise description of the featuresets is given in 6.1.2. First, we show the graphs that contain the scores of models using corrective feedback and errorfrequency or only errorfrequency as input.



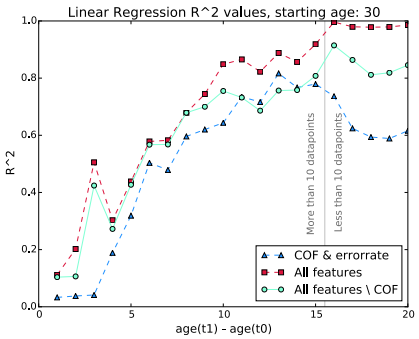


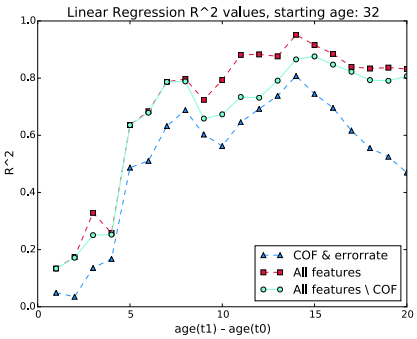


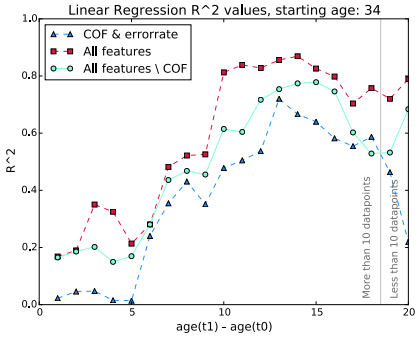


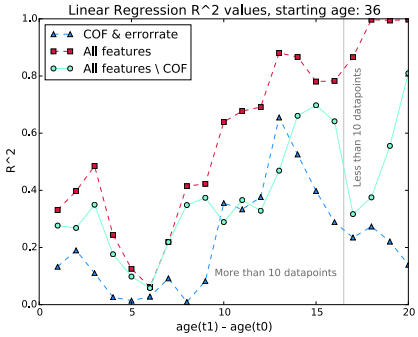


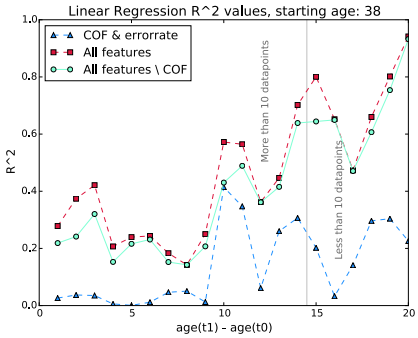
In the following we show the graphs comparing explanatory power of models obtained using corrective feedback and error score as input to those obtained using larger featuresets.






Linear Regression R^2 values, starting age: 34
COF & errorrate
All features
All features \ COF
R^2
age(t1) - age(t0)
More than 10 datapoints
Less than 10 datapoints


Linear Regression R^2 values, starting age: 36
COF & errorrate
All features
All features \ COF
R^2
age(t1) - age(t0)
More than 10 datapoints
Less than 10 datapoints


Linear Regression R^2 values, starting age: 38
COF & errorrate
All features
All features \ COF
R^2
age(t1) - age(t0)
More than 10 datapoints
Less than 10 datapoints

Bibliography

Ron Artstein and Massimo Poesio (2008). Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, 34(4):555–596.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent and Christian Jauvin (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137 — 1155.

Christopher M. Bishop (2006). *Pattern Recognition and Machine Learning*. Springer-Verlag New York.

Lois Bloom (1970). *Language development: Form and function in emerging grammars*. MIT Press.

Susan R. Braunwald (1971). Mother-child communication: The function of maternal language input. *WORD*, 27(1-3):28–50.

Roger Brown (1973). A first language: The early stages. *Journal of Child Language*, 1(2):289–307.

Roger Brown and Camille Hanlon (1970). Derivational Complexity and Order of Acquisition in Child Speech. In: John R. Hayes (ed.), *Cognition and the Development of Language*. John Wiley & Sons, Inc.

Christopher J. C. Burges (1998). A tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, 2:121–167.

Noam Chomsky (1965). *Aspects of the Theory of Syntax*. The MIT Press, Cambridge, Massachussets.

Noam Chomsky (1980). Initial States and Steady States. In: Massimo Piattelli-Palmarini (ed.), *Language and Learning. The Debate between Jean Piaget and Noam Chomsky*. Routledge & Kegan Paul.

Michelle M. Chouinard and Eve V. Clark (2003). Adult Reformulations of Child Errors as Negative Evidence. *Journal of Child Language*, 30(03):637 – 669.

Eve V. Clark (1978). Awareness of Language: Some Evidence from what Children Say and Do. In: Anne Sinclair, Robert J. Jarvella and Willem J. M. Levelt (eds.), *The Child's Conception of Language*, vol. 2, pp. 17–43. Springer Berlin Heidelberg.

M. Demetras (1989). Working parents' conversational responses to their two-year-old sons. *Working Paper. University of Arizona*.

Katherine Demuth, Jennifer Culbertson and Jennifer Alter (2006). Word-minimality, epenthesis, and coda licensing in the acquisition of english. *Language and Speech*, 49(2):137–174.

Jennifer Foster (2007). Treebanks gone bad: Parser evaluation and retraining using a treebank of ungrammatical sentences. *International Journal of Document analysis and Recognition (IJDAR)*, 10(3-4).

E. Mark Gold (1967). Language Identification in the Limit. *Information and Control*, 10:447 – 474.

Paul Grice (1975). Logic and Conversation. In: Peter Cole and Jerry L. Morgan (eds.), *Syntax and Semantics. 3: Speech acts*, pp. 41 – 58. New York: Academic Press.

Alison Henry (1995). Belfast English and Standard English: Dialect variation and parameter setting. *Language in Society*, 25(3):471–476.

Stan A. Kuczaj (1977). The acquisition of regular and irregular past tense forms. *Journal of Verbal Learning and Verbal Behavior*, 16(5):589–600.

Elena Lieven, Dorothe Salomo and Michael Tomasello (2009). Two-year-old children's production of multiword utterances: A usage-based analysis. *Cognitive Linguistics*, 20(3):481–507.

Howard Maclay and Charles E. Osgood (1959). Hesitation Phenomena in spontaneous English speech. *Word*, 15:19–44.

Brian MacWhinney (2000a). *The CHILDES Project: Tools for analyzing talk*. Mahwah, NJ: Lawrence Erlbaum Associates, 3 ed.

Brian MacWhinney (2000b). Guide to the CHILDES database north american english corpora.

Makoto Matsumoto and Takuji Nishimura (1998). Mersenne twister: A 623-dimensionally equidistributed uniform pseudorandom number generator. *ACM Transactions on Modeling and Computer Simulation*, 8(1):3–30.

Tomas Mikolov, Kai Chen, Greg Corrado and Jeffrey Dean (2013a). Efficient estimations of word representations in vector space. In: *In Proceedings of Workshop at ICLR*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado and Jeffrey Dean (2013b). Distributed representations of words and phrases and their compositionality. In: *In Proceedings of NIPS*.

Ernst L. Moerk (1983). A Behavioral Analysis of Controversial Topics in First Language Acquisition: Reinforcements, corrections, modeling, input frequencies, and the three-term contingency pattern. *Journal of Psycholinguistic Research*, 12(2):129 – 154.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Massimo Piattelli-Palmarini (ed.) (1980). *Language and Learning. The Debate between Jean Piaget and Noam Chomsky*. Routledge & Kegan Paul.

Caroline F. Rowland and Sarah L. Fletcher (2006). The Effect of Sampling on Estimates of Lexical Specificity and Error Rates. *Journal of Child Language*, 33(4):859–877.

Jacqueline Sachs (1983). *Talking about the there and then: The emergence of displaced reference in parent–child discourse.*, vol. 4. Hillsdale, NJ: Lawrence Erlbaum Associates.

Kenji Sagae, Eric Davis, Alon Lavie, Brian MacWhinney and Shuly Wintner (2007). High-accuracy Annotation and Parsing of CHILDES Transcripts. In: *Proceedings of the ACL-2007 Workshop on Cognitive Aspects of Computational Language Acquisition*. Prague, Czech Republic.

Kenji Sagae, Brian MacWhinney and Alon Lavie (2004). Adding Syntactic Annotations to Transcripts of Parent-Child Dialogs. In: *LREC*.

Matthew Saxton (2010). *Child language. Acquisition and Development*. SAGE Publications.

Matthew Saxton, Phillip Backley and Clare Gallaway (2005). Negative input for grammatical errors: Effects after a lag of 12 weeks. *Journal of Child Language*, 32(03):643 – 672.

Barbara C. Scholz (2004). Gold's Theorems and the logical problem of language acquisition. *Journal of Child Language*, 31:959 – 961.

Jeffrey L. Sokolov (1993). A local contingency analysis of the fine-tuning hypothesis. *Developmental Psychology*, 29(6):1008 – 1023.

Jeffrey L. Sokolov and Brian MacWhinney (1990). The CHIP framework: Automatic coding and analysis of parent-child conversational interaction. *Behaviour Research Methods, Instruments and Computers*, 22(2):151 – 161.

Patrick Suppes (1974). The semantics of children's language. *American Psychologist*, 29:103–114.

Catherine S. Tamis-LeMonda, Marc H. Bornstein and Lisa Baumwell (2001). Maternal Responsiveness and Children's Achievement of Language Milestones. *Child Development*, 72(3):748–767.

Richard M. Weist, Aleksandra Pawlak and Karen Hoffman (2009). Finiteness systems and lexical aspect in child Polish and English. *Linguistics*, 47(6).

Richard M. Weist and Andrea A. Zevenbergen (2008). Autobiographical memory and past time reference. *Language Learning and Development*, 4(4).

John Wilson and Alison Henry (1998). Parameter setting within a socially realistic linguistics. *Language in Society*, 27(1):1–21.

Kaizhong Zhang and Dennis Shasha (1989). Simple fast Algorithms for the Editing Distance between Trees and Related Problems. *SIAM journal of computing*, 18:1245–1262.