A Data-driven Investigation of Corrective Feedback on Subject Omission Errors in First Language Acquisition

Sarah Hiller and **Raquel Fernández**
Institute for Logic, Language & Computation
University of Amsterdam
sarah.hiller@posteo.de | raquel.fernandez@uva.nl

Abstract

We investigate implicit corrections in the form of contrastive discourse in child-adult interaction, which have been argued to contribute to language learning. In contrast to previous work in psycholinguistics, we adopt a data-driven methodology, using comparably large amounts of data and leveraging computational methods. We conduct a corpus study on the use of parental corrective feedback and show that its presence in child directed speech is associated with a reduction of child subject omission errors in English.

1 Introduction

It is widely agreed that children learn how to use language through interaction with their caregivers and peers. There is, however, a long-standing discussion concerning the exact nature of the learning mechanism enabling this. For example, while there is no doubt that children are exposed to *positive input* (i.e., grammatically correct utterances in context), it is an open question whether they also receive *negative input*—evidence about the inadequacy of their erroneous utterances.

What seems clear is that explicit disapprovals are very rarely used to correct grammatical mistakes, as already shown by Brown and Hanlon (1970). However, certain contrastive constructions may provide negative input in an implicit way. For instance, in the following exchange between 2-year-old Lara and her father, from the corpus by Rowland and Fletcher (2006), the father picks up the child's erroneous utterance in the following turn and presents a form with the appropriate preposition:

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

The contrast between the two forms is particularly noticeable and could potentially lead the child to recognise and correct their own error. It has thus been argued that this type of construction presents children with negative evidence and, unlike explicit corrections, it does so in the course of conversation, without disrupting the dialogue flow.

Researchers have used different terms to refer to this phenomenon, including *recast* (Brown, 1973), *reformulation* (Chouinard and Clark, 2003), *embedded correction* (Clark, 2003), and *corrective input* (Saxton, 2000). In this paper, we adopt the term *corrective feedback* (CF), which we will define precisely in Section 3, and analyse its effect on first language learning—in particular on retreating from subject omission errors in English. In contrast to previous work in psycholinguistics, we investigate these questions in a data-driven manner, using comparably large amounts of data from the CHILDES Database (MacWhinney, 2000a) and developing computational methods to support linguistically motivated studies. More concretely, we make the following contributions:

- We present a taxonomy of child error types that can receive CF and an annotation scheme for coding instances of CF.
- We report the results of a corpus study showing that subject omission errors make up the largest proportion of errors met with CF.
- We develop classifiers to automatically detect subject omission errors and CF on those errors, training on the manually annotated data.
- Using automatically processed data, we investigate the impact of CF on learning subject inclusion in English with a series of linear regression models, showing that CF has predictive power over a variety of control factors. Our results indicate that the effects of CF are most noticeable after a period of about 9 months.

Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning (CoNLL), pages 105–114,
Berlin, Germany, August 7-12, 2016.

2 Related Work

Adult repetitions of child speech with different degrees of variation are a hallmark of child-adult dialogue. As mentioned before, this kind of phenomenon is often referred to as *recast*, and was first studied by Brown and Bellugi (1964). Brown and Hanlon (1970) were the first to suggest that recasts had corrective potential and thus "could be instructive". Since then, several studies have pointed out that contrastive discourse of this kind is very common (e.g., Demetras et al. (1986), Strapp (1999), Chouinard and Clark (2003), Saxton et al. (2005)). There is however no full agreement regarding what motivates the production of corrective feedback. According to Chouinard and Clark (2003), reformulations follow from general cooperative principles: Since children's contributions are often difficult to comprehend due to their limited linguistic abilities, adults explicitly check up on their intended meaning by rephrasing the child's utterance. In contrast, for Saxton et al. (2005), most often corrective feedback does not arise from semantic uncertainty but rather "it is only the linguistic *form* that the adult might take issue with". In any case, researchers generally agree that it is unlikely that adults consciously recast child speech, despite doing it very frequently.

The fact that negative input is available, however, does not immediately mean that it contributes to language development. Some small-scale studies have indeed found an association between corrective feedback and language growth (e.g., Newport et al. (1977), Nelson et al. (1984), Eadie et al. (2002), i.a.). Chouinard and Clark (2003) investigated children's immediate responses to parental reformulations for 5 children and found that acknowledgements and repeats by the child were very frequent. This indicates that children attend to their parent's corrective input and can immediately revert to the correct form. However, it is less clear whether this has a long-term effect. In this respect, Saxton et al. (2005) recorded interactions between 12 mother-child pairs at two time periods with a lag of 12 weeks and found that reformulations had a positive effect on the correct use of 3 out of 13 investigated grammatical structures after this time lag.

Despite the existing evidence, teasing apart the effect of corrective feedback from all other sources of input available to the child is not an easy matter empirically, which explains why the influence of corrective feedback on learning remains controversial (Tomasello, 2009). Here we aim to shed some light on this debate by investigating this phenomenon using much more data than ever before, taking advantage of NLP techniques.

3 Corrective Feedback

We start by defining what we mean by corrective feedback (CF) and by providing a taxonomy of errors in child language that CF can target.

3.1 Definition of corrective feedback

An utterance by the child followed by an utterance by an adult constitutes an instance of corrective feedback if all the following constraints are met:

(C1) The child's utterance contains a grammatical anomaly.

(C2) There is some degree of overlap between the adult and child utterances: the adult's response is anchored to the child utterance through at least one exactly matching word.

(C3) The adult utterance is not a mere repetition of the child's, i.e., there is some contrast.

(C4) This contrast offers a correct counterpart of the child's erroneous form.

While this definition does not make any claims regarding the intentions of the adult nor the possible uptake by the child, which may be considered a simplification, arguably it can be operationalised in a corpus study, which is our main goal here.

3.2 Taxonomy of errors

Now that we have a general definition of the phenomenon, we can proceed towards a more fine-grained classification of types of CF. Mainly, the exchanges can be discriminated via the kind of error in the child utterance and via the kind of correction employed by the adult. Here we focus on the type of error observable in the child utterance, restricting ourselves to grammatical errors, i.e., syntactic and morphological.[1] We differentiate between the linguistic *level* at which the error occurs and the *type* of error observed. For *level*, we follow Saxton et al. (2005) in distinguishing four main classes sub-divided into a total of 13 categories. Regarding *type*, inspired by Sokolov (1993) we distinguish between *omissions*, *additions*, and *substitutions*:[2]

[1] We do not consider phonological and lexical errors as they are not easily identifiable in a transcribed corpus.

[2] Note that Sokolov (1993) uses these terms to characterise the way in which parental utterances diverge from child utter-

- Error Level:
  - *Syntactic:* subject, object, verb[3]
  - *Noun morphology:* possessive, regular plural, irregular plural
  - *Verb morphology:* 3rd person singular, regular past, irregular past
  - *Unbound morphology:* determiner, preposition, auxiliary, present progressive
- Error Type:
  - *Omission*, *Addition*, *Substitution*

The following examples illustrate some of the categories above:

(2) a. *synt: subject, omission*
CHI: don't want to .
MOT: you don't want to ?
b. *v. morph: irregular past, substitution*
CHI: he falled out and bumped his head .
MOT: he fell out and bumped his head .
c. *u. morph: auxiliary, addition*
CHI: I'm read it .
DAD: you read it to mummy .

4 Dataset

Selection. Our dataset consists of a selection of files from the English language section of the CHILDES Database (MacWhinney, 2000a) including all transcripts of conversations between adults and unimpaired, naturally developing children that contain a minimum of 50 utterances by the child and 100 utterances in total, and where the mean length of utterance (MLU; in words) of the child is at least 2. This ensures that the child is already at a stage where grammatical constructions are starting to be used. Finally, since we are conducting a longitudinal study, in order to make sure there is enough data per child we consider only transcripts of children for which there is data over at least one year, with a file density of at least 5 transcripts per year and a minimum of 10 transcripts overall per child.

The resulting dataset contains a total number of 1,683 transcripts from 25 different children, with 1,598,838 utterances overall. The average child age at the time of the first transcribed conversation lays around 2 years, with very little variation. The mean difference between the child's age in the first and the last gathered transcript varies considerably more across children, but overall also lies around 2 years.[4]

	Total	Avg. per child
transcripts	1,683	67.32
utterances	1,598,838	63,953.52
candidate CF pairs	136,152	5,446.08

Table 1: Overview of our dataset containing longitudinal data from 25 different children.

Preprocessing. Most of the transcripts in the dataset already include part-of-speech tagging, morphological analysis, and dependency parsing. We used the CLAN toolbox (MacWhinney, 2000b) and the MEGRASP dependency parser (Sagae et al., 2007) to add POS tags and to morphologically and syntactically parse the transcripts where this information was not available. We also automatically coded each adult response to a child utterance with information on overlap using the CHIP programme (Sokolov and MacWhinney, 1990), also part of the CLAN toolbox. CHIP provides information on added (`$ADD`), deleted (`$DEL`), and exactly matching (`$EXA`) morphemes in the source and response utterances, as well as the proportion of morphemes in the response utterance which match exactly morphemes in the source (`$REP`). Figure 1 shows a sample child-adult exchange with all the layers of information computed during the preprocessing stage.[5]

Selection of candidate CF utterance pairs. In order to investigate the effect of CF on language learning, we need to quantify the CF exchanges present in the corpus. That is, we need to find mechanisms for automatically detecting these. We use the overlap information to extract candidate instances of CF. In line with constraints (C2) and (C3) in our definition, we consider candidate instances all child-adult utterance pairs with a percentage of repetition $0 < \texttt{\$REP} < 1$, where the overlap is not exclusively due to stopwords.[6] We also require that the child's utterance contains a minimum of two distinct words so that there is scope for a grammatical anomaly (C1).

An overview of the dataset in shown in Table 1.

ances, while we use them to characterise the type of error in a child utterance.

[3]We deviate from Saxton et al. (2005) slightly here by considering any main verb including copulas, rather than only copulative verbs as they do.

[4]Further details are given in the supplementary material available at `http://tinyurl.com/cf-conll2016`.

[5]The **manual annotation** layer in Figure 1 is discussed in the next section.

[6]The list of stopwords was empirically derived by taking the function words amongst the 100 most frequent words in the dataset. See the supplementary material for the full list.

CHI: I climb up daddy .	
– **POS & morph**	%mor: pro.sub|I v|climb prep|up n|daddy
– **dependency**	%gra: 1|2|SUBJ 2|0|ROOT 3|2|JCT 4|3|POBJ
DAD: you did climb over daddy .	
– **POS & morph**	%mor: pro|you v|do.PAST v|climb prep|over n|daddy
– **dependency**	%gra: 1|2|SUBJ 2|0|ROOT 3|2|OBJ 4|3|JCT 5|4|POBJ
– **overlap**	%adu: $EXA:climb $EXA:daddy $ADD:you did $ADD:over $DEL:i $DEL:up $REP=0.40
manual annotation	%cof: $CF $ERR=umorph:prep; $TYP=subst

Figure 1: Sample child-adult utterance pair with information layers automatically added during preprocessing, plus a Corrective Feedback layer manually annotated with the decision tree in Figure 2.

5 Corpus Study

The simple heuristic used to extract candidate instances of CF fares very well on recall, but it is of course not very precise: a large quantity of candidate utterance pairs are not instances of CF since (C1) and (C4) in our definition (Section 3.1) are not fully accounted for. We therefore manually annotated a subset of the data to have a reliable basis for analysis and to use as training data for an automatic classifier. For this annotation task, we selected a subset of data that was representative of the entire dataset. We randomly picked four children in the dataset and selected between four and six files per child that covered a minimum period of one year and did not diverge by more than 20 utterances from the average transcript length in the overall dataset. This makes up 25,191 utterances in total (of which 9,783 are child utterances).[7]

We run our heuristic for extracting candidate CF utterance pairs, which resulted in a total of 2,627 pairs of child-adult utterances to be annotated. Of these, 350 instances were annotated by two coders to test the reliability of the annotation. The annotation scheme used distinguishes between CF and non-CF pairs. It subsequently uses the taxonomy of corrective feedback presented in Section 3.2 to indicate the kind of error picked up by the parent in those pairs coded as CF. If several child errors are implicitly corrected in a single CF response, all of them are included in the annotation. Figure 2 shows a simplified version of the decision tree used by the annotators. Inter-annotator agreement was measured with Cohen's *kappa* and was reasonably high ($\kappa = 0.77$). The annotators discussed cases of disagreement and arrived at a consensus label for the final annotation. The annotated dataset as well as the complete annotation guidelines are available at `http://tinyurl.com/cf-conll2016`.

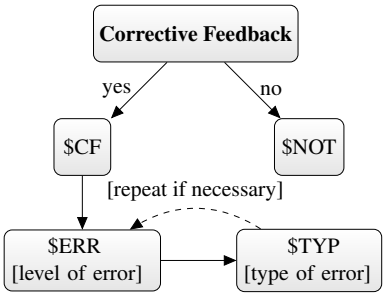

Figure 2: Decision tree for the annotation task.

Table 2 shows the results of the corpus study. Out of 2,627 candidate utterance pairs, 580 where coded as instances of CF. Most of the errors that receive corrective feedback are omissions. This should not necessarily be interpreted as omissions receiving a higher proportion of CF over other error types, but rather as a consequence of omission errors being predominant at this stage of development (Saxton et al., 2005). In particular, most instances of CF (30.8%) occur as a response to subject omission errors (SOEs); an example can be found in excerpt (2a), Section 3.2. In the remainder of the paper we thus focus on this type of error.

Figure 3 shows how the amount of CF received by the children (averaging over all types of errors) changes over time. Not surprisingly, corrective feedback has a clear tendency to decrease as children develop and make fewer errors. An exception amongst the four children targeted for the corpus study is the case of Emily, who has considerably higher MLU than the other three children at 2.5 years of age (MLU of 5 words vs. 4 for Lara and Trevor and 3 for Thomas) and is therefore more proficient, thus offering fewer opportunities for corrections.

[7] See the supplementary material for more details on the selected files.

	Om	*Add*	*Sub*	**Total**
Syntax				
subject	171	–	1	172
verb	90	1	–	91
object	13	–	–	13
N morph				
poss -'s	4	1	–	5
regular pl	–	3	–	3
irregular pl	–	–	3	3
V morph				
3rd person	4	–	–	4
regular past	10	1	–	11
irregular past	1	–	4	5
Unb. morph				
det	79	–	6	85
prep	21	1	12	34
aux verb	114	5	1	120
progressive	9	0	0	9
Other	4	2	19	25
Total	520	14	46	580

Table 2: Types of errors in exchanges coded as CF.

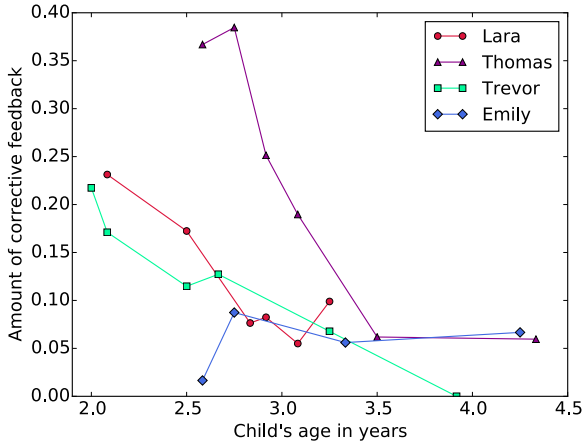

Figure 3: Amount of CF exchanges against child's age. Pearson's r=–0.90 (Lara), r=–0.88 (Thomas), r=–0.97 (Trevor), and r= +0.32 (Emily).

6 Automatic Detection

Our corpus study has provided insight on the types of errors that are met with CF, showing that subject omission errors (SOEs) make up the largest proportion. Our aim is to investigate the extent to which CF on SOEs plays a part in overcoming such errors during language acquisition. A large-scale data-driven investigation, however, requires the development of methods for the automatic extraction of these phenomena. We therefore use the manually annotated data to develop automatic extraction procedures to detect SOEs and CF on SOEs in order to extend our analysis to the entire dataset, beyond the manually annotated data.

For extracting SOEs, a rule-based algorithm was used, while for the more complex task of detecting CF on SOEs we trained a Support Vector Machine. Overall, we are interested in high-precision classifiers: Given our aims, it is preferable to have a conservative but reasonably accurate estimate of amount of error and of presence of corrective feedback in order to avoid unreliable boosting of the possible effect of CF on learning.

6.1 Subject omission errors

To construct a base set for developing a SOE classifier, we coded part of the annotated subset for the presence or absence of subject omission errors (recall that the original annotation only indicated the error type when this was present and received corrective feedback). The resulting base set contains 453 utterances, of which 206 are positive instances and 247 are negative instances of SOEs. This data was randomly split into a training set and a test set, roughly corresponding to 90% and 10% of the data, respectively.

We used the training set to derive a set of simple features that were combined in a non-probabilistic rule-based algorithm classifying utterances as SOE or non-SOE. Feature tuning was done by qualitative analysis of classification errors in a 5-fold cross-validation setting. Our baseline consisted in classifying all child utterances with no `SUBJ` node in the dependency parse tree as positive instances of SOE. This already achieves an F1-score of 0.82, with 0.74 precision and 0.93 recall. To improve precision while not lowering recall too dramatically, we experimented with additional features aimed at accounting for parsing errors. The algorithm that produced the highest precision results first considers utterances with no `SUBJ` node (baseline feature) or erroneously parsed utterances where the `SUBJ` tag is assigned to an implausible word such as a negation particle. It assigns them the positive class (SOE) if additionally the first word in the utterance is not a noun and the subject is not overlooked in the dependency parse due to a missing verb (captured by checking whether the `ROOT` node is assigned to a proper name). All instances that do not meet these constraints are classified as non-SOE.[8]

This classifier resulted in a precision of 0.83 on both training and test sets, and a recall of 0.86 on the training set and of 0.8 on the test set.

[8]The algorithm is spelled out in pseudocode in the supplementary material.

6.2 CF on subject omission errors

Given a child utterance with a SOE, the CF-SOE classifier is intended to detect whether or not the parental response contains corrective feedback for this type of error. Using the SOE detector described above, we extracted 514 utterance pairs where the child utterance exhibits a SOE. Manual inspection showed that this base set contains 250 positive instances and 264 negative instances of CF on SOEs. Again, we randomly split off the data into a training and a test set, corresponding to approx. 90% and 10% of the data, respectively.

To capture the interaction between the child and adult utterance needed for this classification proved to be harder than extracting the simple features representative of a SOE. For this task, we used the SVM implementation provided with the Python scikit-learn module (Pedregosa et al., 2011). Again, features were selected via qualitative analysis of wrongly classified instances in a 5-fold cross-validation setting over the training set. The final set of features used includes the presence of a `SUBJ` node in the dependency parse of the adult utterance, and exact matches in `ROOT` nodes (typically verbs) or `OBJ` nodes in the child and adult utterances.[9] In order to boost precision, we tuned the parameters of the SVM by giving a higher error-score to misclassifications of non-CF instances, while making sure the F-score did not fall below 0.5. The final class weights were 3:1 for the non-CF vs. CF class.

Overall, precision of the selected classifier reached 0.89 and recall 0.36 on the test set, compared to a majority class baseline of 0.49 for both values. Given the importance of precision for our aims, this classifier was preferred over a more balanced one.

7 CF and Language Learning

In this section, we investigate whether the presence of corrective feedback on subject omission errors contributes to the reduction of such errors in children's speech and thus has an impact on language acquisition.

7.1 Overview of the experimental design

Our experiments are designed as follows: We estimate the amount of SOEs at a particular period of time (defined in terms of child age in months) as the proportion of child utterances that contain a SOE. We compute the amount of SOEs at two different time periods, t_0 and a later time t_1. We then calculate the *relative error reduction* (`rer`) as the proportion of SOEs at t_0 that has been overcome at t_1:

$$\texttt{rer}(t_0, t_1) = \frac{SOE_{t_0} - SOE_{t_1}}{SOE_{t_0}} \qquad [1]$$

Our aim is to investigate the relationship between *relative error reduction* (`rer`) of SOEs at t_1 and the presence of corrective feedback on SOEs at t_0. The latter is calculated as the number of instances of CF on SOE at t_0 divided by the total number of child SOEs at t_0. We consider all possible instantiations of t_0 and t_1 per child in the corpus, with a minimum time distance of one month between the two. This allows us to investigate at what age CF seems more effective (different t_0 values) and how much time is needed for its effect to be noticeable on learning (distance between t_0 and t_1).

We construct several linear regression models, where $\texttt{rer}(t_0, t_1)$ is always the dependent variable we are interested in predicting and CF at t_0 is the independent variable whose predictive power we are investigating, while controlling for several other factors characterising child directed speech and children's own speech.

7.2 Setup details

Data. We apply the SOE and the CF-SOE high-precision detectors presented in Section 5 and trained on manually annotated data to the entire dataset (summarised in Table 1). This allows us to quantify the amount of SOE and CF on SOE a child receives at a given age. Overall, we detected 287,309 cases of child SOEs and 31,080 cases of CF on a SOE.

Control variables. To study the effects of CF we control for other features that may also contribute to predicting relative error reduction. We consider factors representative of the general language development exhibited by the child as well as of the quality and quantity of the input. To be precise, we consider the following factors:

(a) child age in months (`age`);

(b) mean length of utterance of child speech and of child directed speech (`chi.mlu`/`cds.mlu`);

(c) vocabulary size of child speech and of child directed speech (`chi.vocab`/`cds.vocab`);

[9] The complete list of features passed to the SVM, together with an explanation of what they represent, can be found in the supplementary material.

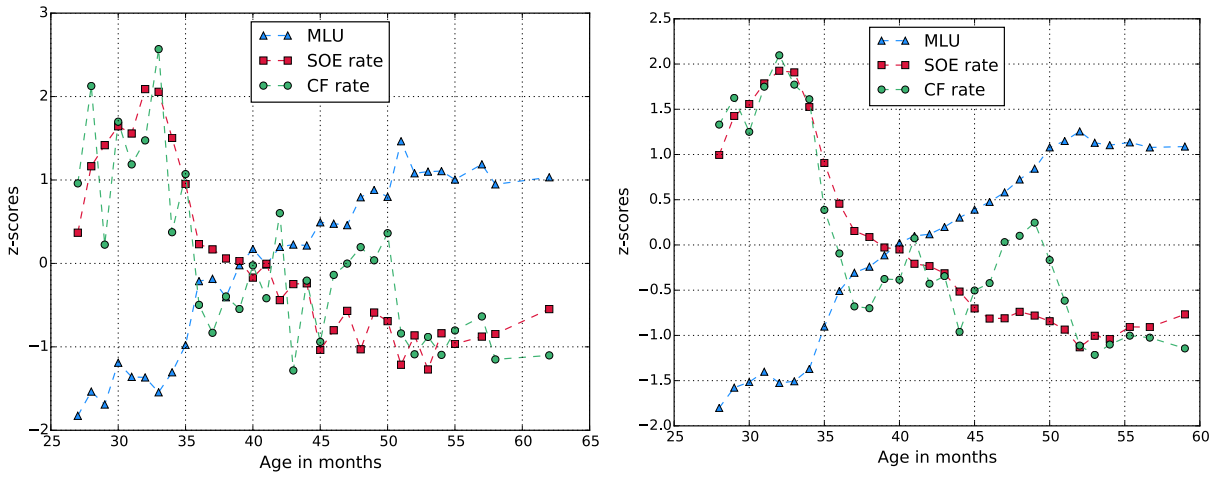

Figure 4: Development of child MLU, proportion of SOE and CF frequency (in standardised z scores) for Adam in the Brown corpus, when measured by month (left) and after smoothing (right).

(d) proportion of child SOEs (`chi.soe`);

(e) proportion of child directed utterances with subject omissions (`cds.so`);

(f) proportion of words uttered by the child over all words uttered in the child-adult interactions (`chi.speech`).[10]

All factors are computed at t_0, except for vocabulary size, which corresponds to the vocabulary of all transcripts available for a given child up to t_0. Note that we also encode how often adults produce utterances without an explicit subject (e), which does of course happen relatively often in spontaneous conversation (Fernández and Ginzburg, 2002). All of these features are expected to influence the degree to which a child learns subject inclusion in English. Our goal is to test to what extent (if at all) CF on SOEs has a positive influence on learning independently of these factors.

Feature computation. There is substantial variation amongst children in the corpus regarding the density and the length of transcripts. Therefore, to estimate the values of the variables above as well as amount of CF and `rer` at a given age, in months, t_i, we employ an averaging procedure over possibly several transcripts. This procedure consists first of all in averaging over all available transcripts in the same month. The resulting estimate for relatively steady measures such as MLU still shows a considerable fluctuation. We thus employ a smoothing procedure by averaging again over three consecutive available months t_{i-1}, t_i, t_{i+1}. Figure 4 illustrates the effect of such smoothing procedure for some of the variables.

[10]Given the nature of the CHILDES corpus (with unequal frequency and length of transcripts per child), it is not possible to properly estimate the amount of input received; we can only quantify the proportion of child speech vs. child directed speech.

7.3 Analysis and results

We first consider the entire dataset as a whole, taking all possible pairs (t_0, t_1) for the 25 children in the corpus (N= 2613). We observe that CF correlates positively with $\texttt{rer}(t_0, t_1)$ (Pearson's r=0.29, p<0.001); that is, the more corrective feedback at any given time t_0, the more error reduction at later times in development. A linear regression model controlling for the additional factors listed above shows that CF explains a significant proportion of the variance in relative error reduction of SOEs independently from all other factors.

Table 3 shows the standardised regression coefficients for the predictors considered, representing the change in $\texttt{rer}(t_0, t_1)$ associated with a change of 1 in the given predictor when all other factors are held constant. Note that since in this setting neither t_0 nor t_1 are fixed, we can include age at these two time periods as independent predictors in the model.

While this analysis shows that CF on SOEs contributes to predicting error reduction, it does not provide any information regarding the developmental period at which corrective feedback may be more effective or the time lapse required for learning to take place. The following analyses aim at offering insight on these aspects.

To investigate the time lag required for CF to have an impact on `rer`, we fix the distance be-

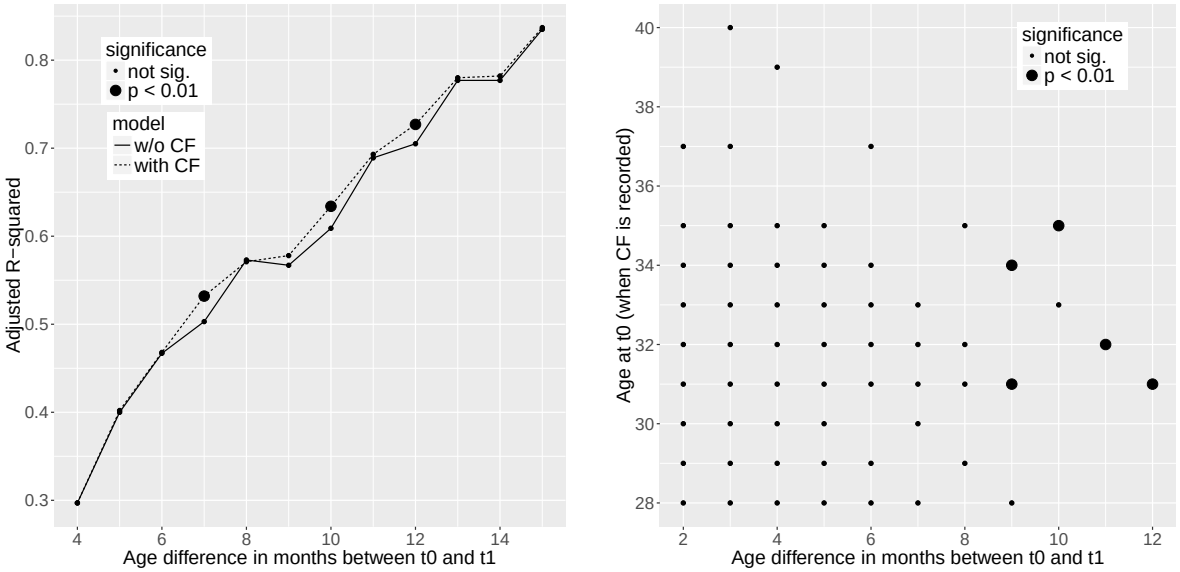

Figure 5: Impact of CF on SOE reduction after different time lags (left) and after different time lags for a fixed age at t_0 (right).

age.t0	-0.586 ***	age.t1	0.474 ***
chi.soe	0.843 ***	cds.so	0.101 **
chi.mlu	0.365 ***	cds.mlu	0.217 ***
chi.vocab	0.479 ***	cds.vocab	-0.450 ***
chi.speech	0.853 ***	cf.soe	0.123 ***

Table 3: Beta coefficients for linear regression model; *** p<0.001, ** p<0.01. Overall variance in rer captured by the model (adjusted R^2): 0.53.

tween t_0 and t_1 and construct two different linear regression models for each distance value, a model with all control factors without CF (and without age at t_1 as it is dependent on age at t_0 in this setting) and a model where CF is included. This allows us to test whether CF makes a significant contribution in accounting for the variance in rer, independently of the control factors. We should note that for each distance value (and thus for each pair of models) there are fewer datapoints than in the previous analysis, as not all distance values are available for all children in the corpus. The average number of datapoints per model is N=186.6 (sd=57.3). Figure 5 (left) shows the adjusted R^2 values for the pairs of models for a range of time spans between t_0 and t_1. As can be seen, CF has a significant effect after a time lapse of 7 to 12 months.

Finally, to investigate not only the time lapse but also the age at which CF may be more effective, we construct linear regression models with fix values for both (i) the difference between t_0 and t_1 and (ii) age at t_0. Again, the number of available datapoints drops substantially with this additional constraint; we consider only settings for which there are at least 10 datapoints (t_0 from 28 to 35 months). The results are shown in Figure 5 (right), where the larger dots indicate that the model that incorporates CF on top of the control factors has significantly more predictive power to explain the variance in rer.[11] We can see that the effect of CF is noticeable after a time lapse of 9 months (consistently with the previous analysis) and that this holds for any starting age t_0 for which there is available data. This is consistent with the results of Saxton et al. (2005), who did not find an effect of CF on this error type after a lag of only 3 months (the only time lag studied by these authors). Our results show that indeed, for SOEs, the learning effects are cumulative and can only be statistically appreciated at a later stage.

The results plotted in Figure 5 (left) may seem to indicate that after a lag of over 12 months this learning effect disappears, since we observe no significant difference between the two models (with and without CF). However, we believe that this is an artefact of the dataset. In our dataset a larger time lag between t_0 and t_1 coincides with a relatively advanced age at t_1: on average, children are 47 to 49 months old at t_1 for differences of 13 to 15 months between t_0 and t_1. The fre-

[11] Since the experiments reported in Figure 5 involve multiple statistical comparisons (thus increasing the chance of Type I errors), we use a stricter significance threshold (0.01) than the standard 0.05.

quency of subject omissions observed in the child speech converges towards a non-zero limit with increasing child age (a limit akin to the amount of subject omissions in child directed speech). Thus, predicting the relative error reduction when given the frequency of SOE at t_0 becomes easy at advanced child ages, with or without including CF as a predictor. The observed increase of adjusted R^2 scores as age distance becomes larger seems to support this hypothesis.

8 Conclusions

We have investigated the impact of corrective feedback on first language acquisition, in particular on the reduction of subject omission errors in English—a type of error which we found to be the most commonly met with CF in our corpus study. In contrast to previous small-scale studies in psycholinguistics, we have addressed this problem using a comparatively large data-driven setting. We have used machine learning methods trained on manually annotated data and then applied statistical modelling to the automatically extracted instances of CF on SOEs. The annotated dataset is publicly available at `http://tinyurl.com/cf-conll2016`.

Our results have shown that CF contributes to predict learning independently of other factors characterising the input received by the child and the child level of development. In our dataset, this is noticeable after a time lag of approximately 9 months. This suggests that CF does have an impact on long-term learning and not only on immediate responses by the child, as observed in other analyses with a few children (e.g., Chouinard and Clark (2003)). Of course, our results need to be interpreted with prudence, since the automatic classifiers for detecting SOEs and CF on SOEs in the whole dataset are far from perfect. Nevertheless, since we opted for high-precision classifiers, the results may in fact be a low estimate of the effect of CF on learning grammar.

Acknowledgments

We are grateful to the anonymous reviewers for their comments, as well as to Stella Frank and Willem Zuidema, who provided useful feedback on an earlier version of this work. This research has received funding from the Netherlands Organisation for Scientific Research (NWO) under VIDI grant nr. 276-89-008, *Asymmetry in Conversation*.

References

Roger Brown and Ursula Bellugi. 1964. Three processes in the child's acquisition of syntax. *Harvard educational review*, 34(2):133–151.

Roger Brown and Camille Hanlon. 1970. Derivational complexity and order of acquisition in child speech. In John R. Hayes, editor, *Cognition and the Development of Language*. John Wiley & Sons, Inc.

Roger Brown. 1973. *A first language: The early stages.* Harvard U. Press.

Michelle M. Chouinard and Eve V. Clark. 2003. Adult reformulations of child errors as negative evidence. *Journal of Child Language*, 30(3):637–670.

Eve V. Clark. 2003. *First language acquisition*. CUP.

Marty J. Demetras, K. Nolan Post, and C. E. Snow. 1986. Feedback to first language learners: The role of repetitions and clarification questions. *Journal of Child Language*, 13(02):275–292.

Patricia A. Eadie, M. E. Fey, J. M. Douglas, and C. L. Parsons. 2002. Profiles of grammatical morphology and sentence imitation in children with specific language impairment and down syndrome. *Journal of Speech, Language, and Hearing Research*, 45(4):720–732.

Raquel Fernández and Jonathan Ginzburg. 2002. Non-sentential utterances in dialogue: A corpus study. In *Proceedings of the Third SIGdial Workshop on Discourse and Dialogue*, pages 15–26, Philadelphia, USA. Association for Computational Linguistics.

Brian MacWhinney. 2000a. *The CHILDES Project: Tools for analyzing talk.*, volume 2: The Database. Lawrence Erlbaum Associates, 3 edition.

Brian MacWhinney. 2000b. *The CHILDES Project: Tools for analyzing talk.*, volume 1, Part 2: the CLAN Programs. Lawrence Erlbaum Associates, 3 edition.

Keith E. Nelson, M. M. Denninger, J. D. Bonvillian, B. J. Kaplan, and N. D. Baker. 1984. Maternal input adjustments and non-adjustments as related to children's linguistic advances and to language acquisition theories. In *The development of oral and written language in social contexts*, volume 13, pages 31–56. Ablex Norwood, NJ.

Elissa L. Newport, H. Gleitman, and L. R. Gleitman. 1977. Mother, I'd rather do it myself: Some effects and noneffects of maternal speech style. In *Talking to children*, pages 109–49. Cambridge U. Press.

Fabian 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.

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, November.

Kenji Sagae, E. Davis, A. Lavie, B. MacWhinney, and S. 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.

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

Matthew Saxton. 2000. Negative evidence and negative feedback: Immediate effects on the grammaticality of child speech. *First Language*, 20(60):221.

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.

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

Chehalis M. Strapp. 1999. Mothers', fathers', and siblings' responses to children's language errors: Comparing sources of negative evidence. *Journal of Child Language*, 26(02):373–391.

Michael Tomasello. 2009. The usage-based theory of language acquisition. In *The Cambridge handbook of child language*, pages 69–87. Cambridge Univ. Press.