Supplementary Material

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

Index.

1. Dataset: Overview of the selected dataset.

2. List of Stopwords: Complete list of empirically extracted stopwords.

3. Annotated Data: Overview of the files selected for manual annotation.

4. Automatic detection:

- Algorithm used for SOE extraction.
- Features used for CF-SOE extraction.

1. Dataset. Average starting age, range of months covered, number of transcripts per child 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.

Corpus	start age	age range	files p. child	child utt. p. 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**

2. List of stopwords. We considered as stopwords all function words amongst the 100 most frequent words in the dataset. Complete list: *a, about, and, at, because, big, but, down, for, good, he, her, here, his, I, if, in, is, it, just, me, my, no, not, now, of, oh, okay, on, out, right, s, she, so, t, that, the, them, then, there, they, this, to, too, up, we, well, with, yeah yes, you, your*.

3. Annotated data. Overview of the transcripts selected for manual annotation together with the age of the child in each file.

Corpus – Child	Files	Age
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
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

4. Automatic detection. The algorithm / features used for automatic detection of SOEs and CF on SOEs.

SOE detection.

Algorithm 1 was used for classifying SOEs .

```
Data: Set of manually annotated child
      utterances.
Result: Classification into SOE vs. non-SOE
for utterance in list do
    $p_1$(utt) = no SUBJ in dependency parse;
    $p_2$(utt) = negation or phoneme falsely
    identified as SUBJ;
    $p_3$(utt) = first word a noun;
    $p_4$(utt) = INCROOT dependency on proper
    name;
    if [$p_1(utt)$ OR $p_2(utt)$] AND $\neg p_3(utt)$ AND
    $\neg p_4(utt)$ then
        return SOE;
    else
        return non-SOE;
    end
end
```

Algorithm 1: Classification of SOE vs. non-SOE

CF-SOE detection.

The features passed to the SVM for CF-SOE classification were the following.

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 and does not recognise the subject. Those parental questions starting with a form of 'to be' are most often followed by an added subject correcting the child's missing one. 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.