The subject-object asymmetry revisited: Experimental and computational approaches to the role of information structure in children's argument omission

Journal:	*IEEE Transactions on Cognitive and Developmental Systems*
Manuscript ID	TCDS-2017-0177.R2
Manuscript Type:	SI: Language Learning
Date Submitted by the Author:	n/a

Complete List of Authors:	Graf, Eileen; NORC at the University of Chicago, Academic Research Centers Theakston, Anna; The University of Manchester, Division of Human Communication, Development & Hearing Freudenthal, Daniel; The University of Liverpool, School of Psychology Lieven, Elena; The University of Manchester, Division of Human Communication, Development & Hearing

Keywords:	attention mechanisms and development, cognitive system and development, language acquisition through development, theory of development

The subject-object asymmetry revisited: Experimental and computational approaches to the role of information structure in children's argument omissions

Eileen Graf, Anna Theakston, Daniel Freudenthal, Elena Lieven

Abstract—**In two studies we investigated the relation between information structure and argument omission in German child language in order to quantify to what extent the subject-object hypothesis (i.e., subjects are omitted more often than objects) is influenced by discourse pragmatics. Twenty-four children took part in an elicited production study in which they produced transitive SVO and OVS sentences. Both constructions are instances of a *topic-comment* information structure. The results showed that 3;6 year-old children omitted subjects and objects alike when the arguments assumed topics status and were placed in utterance-initial position. In a second study we then assessed whether a model of language learning implemented with a recency-bias (resulting in learning from the end of utterances) would produce similar omission rates of initial arguments. The model was found to be sensitive to the frequency with which both word orders occurred in the input: Initial objects were omitted more often than initial subjects, the pattern found in German caregiver speech. The results suggest that argument omission is heavily influenced by information structure and that a subject-object asymmetry per se does not exist.**

Index-terms—**Language development, argument omission, information structure, subject-object asymmetry, null-subjects, object drop, experimental pragmatics, computational models of language learning, Model of Syntax Acquisition in Children (MOSAIC).**

I. Introduction

Studies of language acquisition generally find that children omit subjects more frequently than they omit objects, both in pro-drop and in non-pro-drop languages [1], [2], [3], but see [4], [5] for a different view. The reasons for this asymmetry are unclear. Some theories posit performance limitations (*in production*: [6], or *in learning*: [7] while others offer grammatical accounts [8], [9], [10]) for the asymmetry. To some extent, all of these theories invoke the information structure distinction between *given* and *new,* which is argued to render subjects more susceptible to omission [10]. Subjects tend to contain known or recurring information [11] and thus are more expendable than objects [12].

A problem with most of these accounts is that they look at grammatical subjects and objects based on their prototypical information structural distribution. In English, the most frequent word order places subjects in preverbal position and objects in postverbal position. The prototypical English transitive sentence is an unmarked predicate-focus construction in which the focal element – the object – is generally expressed with a full noun phrase ([13], but see [11] for other languages) and carries primary stress. Based on the frequency of this SVO construction, it can be assumed that objects are inherently focal and subjects inherently topical [13], which might explain why objects are often expressed with lexical NPs (and not with pronouns) even when they contain given information. Given these different information structural properties, an asymmetry in subject-object omission in SVO sentences is not unexpected. However, an analysis of subjects and objects in their typical SVO positions therefore presents a far from ideal test case for the subject-object asymmetry.

Evidence that argument omissions are influenced by the information structure of a particular construction is presented by Allen [14]. Allen's data from child Inuktitut show that the rate of subject omissions differs significantly between transitive and intransitive constructions. Children are more likely to omit a transitive subject than an intransitive subject. Allen links her findings to Du Bois' Preferred Argument Structure account [11], which holds that constructions differ with regard to their canonical information structure: intransitive subjects provide the prime locus for new information, while transitive subjects express given information [15]. Clancy [16] reports similar findings from child Korean. In line with Preferred Argument Structure predictions [11], Clancy's data show that intransitive subjects serve to introduce new referents that are subsequently the topic of discussion. Transitive subjects, in comparison, are constant and likely to reflect the same/given referents for long stretches of the discourse. Transitive object arguments are characterised by a continuous stream of new referents. Clancy argues that the transitive subject "provides the stable 'ground', which can be left unspoken, against which the transitive object accommodates potentially changing 'figures', which call for overt mention" [16, p. 101].

Word order in German is more variable than in English and allows, among others, for these two constructions: an SVO construction, in which the subject is the topic, and an OVS construction, in which the object is the topic [17], [18]. In both constructions, subject and object arguments are topical in that they constitute those entities that the proposition expressed by the sentence is about [13], [19]. Following Lambrecht's definition, a referent constitutes the topic of a proposition if said proposition expresses "information which is relevant to and which increases the addressee's knowledge of this referent" [13, p. 131]. A constituent serves as a topic expression if its accompanying proposition is "pragmatically construed as being about the referent of this constituent" [13, p. 131].

In German, it is therefore possible to compare subjects and objects when both assume the same information structural role (topic) in the same sentential position (initial). Both word orders, SVO and OVS, are instances of a productive *topic-comment* construction in German, and enable us to un-confound sentence position and information status in a way not possible in English.

Spoken German, despite being a non-pro-drop language, allows null references for both subjects and objects in utterance initial position; examples (1) and (2) illustrate instances of topic drop with the omitted element displayed in brackets in the English translation (examples taken from the Szagun corpus [20]):

(1) [S]VO:
Ø kannst die Kasse ja mal eben aufmachen
Ø can [2nd ps] the cash register just perhaps open
[You] can just perhaps open the cash register.

(2) [O]VS:
Ø hab' ich mir überlegt
Ø have I me thought of
I have thought of [this] myself.

Previous research on German suggests that the subject-object asymmetry exists in German child language [21]. There is evidence that subject drop and object drop differ in German child language, though not in a way predicted by the subject-object asymmetry. Analysing a corpus of two German children, one from age 3;1 to 3;4 and another from 3;4 to 3;7 years of age, Hamann [9] examined all instances of argument omissions in obligatory contexts. She found that both argument types were omitted with similar frequencies overall, thus showing a symmetrical distribution. However, when the data were analysed according to the position of each argument in the utterance, a different picture emerged. With regard to objects, Hamann reported that the sentence final object drop rate (SV[O]) was lower than the sentence initial drop rate ([O]VS). She further showed that the two children in her corpus omitted sentence final subjects (OV[S]) more often than sentence final objects (SV[O]). Moreover, the omission of initial subjects ([S]VO) decreased over the developmental period, but the omission rates of initial objects in [O]VS remained constant. Thus, Hamann's findings indicated that initial objects were omitted more than initial subjects, a reversal of the typically discussed asymmetry. She argued that the higher rates of object drop could be explained by the fact that whenever an object is fronted it is because it is a topic, whereas pre-verbal subjects are not always topics. Further, Hamann posits that since postverbal subjects are not topics and yet were sometimes omitted in her corpus, argument omission in child German cannot be considered a purely pragmatically motivated topic drop phenomenon. Hamann's analyses, however, were purely descriptive. Additionally, based on a sample size of n=2, conclusions can only be drawn very tentatively.

There is as yet no conclusive evidence on the subject-object asymmetry in adult German, but there are some preliminary data that it might be similarly reversed, such that initial objects are dropped more often than subjects. Dittmar and colleagues [22] analysed a sample of 7032 utterances of mothers to six monolingual German children (aged between 1;8 and 2;5) taken from the Szagun corpus [20]. The authors identified 745 transitive sentences, 86 (11.5%) of which they categorized as fragments. Of these, 65 (75.6%) utterances were [O]VS structures with omitted objects and 21 (24.4%) were [S]VO structures with omitted subjects. Although this is a rather small sample, it nevertheless showed that objectless sentences in [O]VS are three times as frequent as subjectless sentences in [S]VO. In line with the Hamann study, the Dittmar et al. data suggest that, in German, when subjects and objects occupy the same information structural position, the subject-object asymmetry is reversed.

In our first – experimental – study, we examined the rates of argument omission by German children when subjects and objects exhibited these same information structural and discourse-pragmatic properties. That is, both argument types were placed in either sentence-initial or sentence-final position and construed as topical. In a second –computational – study we then modelled how argument omission of this type may be learnt and whether the output is reflective of the input in this regard.

II. Study 1: Eliciting Argument Omission

In the present study we used an elicited production/sentence repetition paradigm. In order to elicit SVO as well as OVS utterances, the children took part in a task in which they were asked to check a sticker book for a 3rd person referent (an elephant). The task was to tell the experimenter which stickers were still missing and which stickers were already in place in the elephant's sticker book. The children were provided with a verbal prompt modelled on the structure in question (depending on condition, either SVO or OVS) and had to decide which one was the correct answer.

A. Method

Participants. 24 typically developing German-speaking children aged between 3;3 and 3;8 (M = 3;6, 15 boys) were included in the study. A further 10 children were excluded because they either did not use the respective SVO or OVS structures according to condition for at least 2 out of 4 trials (n=5) or they did not pass the pre-test (n=5).

1) Materials

Pre-test. A sheet of A4 sized paper displayed an array of five double boxes and stickers with animal characters inside the boxes. In two of the five double boxes there was a picture of one single animal character (see Figure 1. for illustration), i.e., the sticker pair was incomplete. The remaining three double boxes contained the same animal character twice, one in each part of the double box, i.e., the sticker pair was complete (see Figure 2. for illustration).

Figure 1. *sticker pair incomplete*

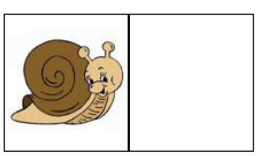
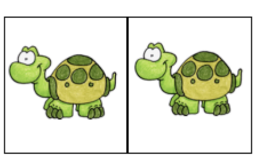


Figure 2. *sticker pair complete*



Test. Two sticker books were created for each child, one for each condition (SVO and OVS). The order in which the children received the two conditions was counterbalanced across participants. Each book contained eight pages, each depicting a double box. Four pages showed one single animal referent in one of the boxes and the remaining four pages showed the same referent displayed twice, one in each part of the double box (see Figures 1. and 2.). Thus, like in the pre-test, when there were two referents, the double box was filled (complete) and when there was only one animal referent, the double box was half empty (incomplete).

Four of the eight pages in each book corresponded to target trials, i.e., the children received a verbal prompt before they checked the elephant's book. The remaining four pages were not presented with verbal prompts and constituted filler items (both test and filler items had 2 complete and 2 incomplete sticker pairs each). Two additional fillers were inserted between each of the eight pages: One sheet displayed a geometric shape and one sheet was a coloured page. For an overview of the layout, see Table 1.

Table 1. *Book layout.*

Page #	Function	Content
0	cover	none
1	no prompt double box + page showing geometric shape + coloured page	sticker missing square red
2	prompt: *He's got it or he needs it.* double box + page showing geometric shape + coloured page	Sticker pair complete rectangle yellow
3	no prompt double box + page showing geometric shape + coloured page	sticker pair complete triangle blue
4	prompt: *This one*[1] *he wants or this one he owns.* double box + page showing geometric shape	sticker missing circle

[1] The English cleft construction is given here as an approximate translation. Cleft constructions exhibit different information structural properties. The left dislocated element [*this one*, in the above example] is scoped out in order to be the contrastive topic of the utterance. In both English and German this can be done to introduce new referents or to refer to a contrasting argument. Cleft elements can never be omitted. The OVS structure in the present study, however, topicalises the object in the same way an SVO structure topicalises the subject.

	+ coloured page	orange
5	no prompt double box + page showing geometric shape + coloured page	sticker missing hexagon pink
6	prompt: *This one he's got or this one he wants.* double box + page showing geometric shape + coloured page	sticker pair complete rhombus purple
7	no prompt double box + page showing geometric shape + coloured page	sticker pair complete trapeze brown
8	prompt: *He needs it or he owns it.* double box + page showing geometric shape + coloured page	sticker missing ellipse white

In the target trials, the verbal prompts included action/verb pairs of either "haben" *to have* or "besitzen" *to own* on the one hand and "wollen" *to want* or "brauchen" *to need* on the other. These stative verbs were chosen in order to make subjects and objects more comparable in terms of their semantic roles. According to Næss (23), semantic distinctness is measured in degrees of affectedness: the more affected the patient of a transitive action is, the more distinct are the participating agent/subject and patient/object roles. Thus, agent and patient roles are minimally distinct when the patient is *least* affected - as is the case with stative verbs. These four verbs were paired in opposites, such that the children had to choose between two verbs in each prompt, e.g. *He's got it or he needs it*. Of the resulting four verb pairs each also had a counterpart in which the verbs occurred in the reverse order. This yielded eight paired utterances, which were randomised across the two conditions (see Table 2).

Table 2. *Sentences prompts for both conditions*

SVO condition	**OVS condition**
Der hat die oder der braucht die. He$_{[SUBJ]}$ has it$_{[OBJ]}$ or he$_{[SUBJ]}$ needs it$_{[OBJ]}$. *He's got it or he needs it.*	Die braucht der oder die hat der. It$_{[OBJ]}$ needs he$_{[SUBJ]}$ or it$_{[OBJ]}$ has he$_{[SUBJ]}$. *This one he needs or this one he's got.*
Der besitzt die oder der braucht die. He$_{[SUBJ]}$ owns it$_{[OBJ]}$ or he$_{[SUBJ]}$ needs it$_{[OBJ]}$. *He owns it or he needs it.*	Die braucht der oder die besitzt der. It$_{[OBJ]}$ needs he$_{[SUBJ]}$ or it$_{[OBJ]}$ owns he$_{[SUBJ]}$ *This one he needs or this one he owns.*
Der will die oder der besitzt die. He$_{[SUBJ]}$ wants it$_{[OBJ]}$ or he$_{[SUBJ]}$ owns it$_{[OBJ]}$. *He wants it or he owns it.*	Die besitzt der oder die will der. It$_{[OBJ]}$ owns he$_{[SUBJ]}$ or it$_{[OBJ]}$ wants he$_{[SUBJ]}$ *This one he owns or this one he wants.*
Der hat die oder der will die. He$_{[SUBJ]}$ has it$_{[OBJ]}$ or he$_{[SUBJ]}$ wants it$_{[OBJ]}$. *He's got it or he wants it.*	Die hat der oder die will der. It$_{[OBJ]}$ has he$_{[SUBJ]}$ or it$_{[OBJ]}$ wants he$_{[SUBJ]}$ *This one he's got or this one he wants.*

These prompts were presented as statement pairs, not questions, thus the experimenter produced each prompt with a falling intonation curve. Tables 1. and 2. illustrate that the children were prompted with [*pronoun-verb-pronoun*] constructions, following full naming phrases. This *NP–then–pronoun* sequence accords with the pragmatics of the discourse and establishes subject and object as topic. Both subject and object arguments were always given, always animate, and always referred to with pronouns, in order to make both SVO and OVS constructions maximally comparable. The subject referent was the *elephant* in each case, referred to in its masculine demonstrative pronoun form "der" *he*. There were sixteen object referents: *cat, turtle, ant, lizard, duck, cow, spider, snake, bee, snail, mouse, giraffe, bat, fly, caterpillar* and *goat*. All these nouns are grammatically female in German, thus the corresponding female demonstrative pronoun form used was "die" *she*. Feminine pronouns are not morphologically marked for nominative or accusative case and the resulting ambiguity enables the use of the female pronoun in both transitive subject and object position in the exact same linguistic form.

2) Procedure

During the *pre-test*, the children were familiarised with the task. They were introduced to a 3rd person referent, *a toy elephant*. The experimenter explained that the elephant liked animal stickers and showed the child a box full of stickers that the elephant had brought along.

Then she showed the child the pre-test sheet with five double boxes, each containing either one sticker of an animal referent in one half of the double box or two identical animals in both halves of the double box (see Figures 1. and 2.). The experimenter further explained that the elephant collected stickers in the double boxes and that each animal sticker needed a sister sticker to form a pair. The experimenter explained that the elephant had already started putting in some of the sister stickers, but had not finished, and asked if the child was willing to help out. The child was first asked to name all the referents on the sheet (*lion, dog, rabbit, mole, bird*) and to show which of the referents already had sister stickers and which ones were missing. Then the experimenter drew a sticker from a box and the child checked the sheet to see if its sister referent was missing or not. When it was missing, the child placed the sticker in the empty box. The children took part in the subsequent test phase only if they successfully completed the pre-test. 5 of the youngest children did not pass the pre-test and were excluded.

During the test phase, the child was asked to repeat the task for the elephant's two sticker books. The experimenter placed a closed book on the table and placed the elephant on top of it. Then she laid out a row of stickers (corresponding to the stickers in the book) face down on the table. First, she drew the first sticker and asked the child what animal character it displayed. The experimenter either confirmed what the child said or corrected her in the few cases when a child named the animal incorrectly. Then she placed the sticker face up next to the sticker book.

In the *SVO condition*, the experimenter directed the child's attention to the subject, the elephant, in order to set it up as a topic and consequently motivate the SVO structure. She modelled the discourse context using a full NP + pronoun sequence twice in order to refer to the subject referent:

> "Schau mal hier beim Elefant, wie der sich freut über die Biene! Na sowas, der Elefant! Kannst du für ihn mal hier [in seinem Aufkleberheft] nachschauen?"
> (*Look here, the elephant, he's very excited about the bee* [sticker]*! Look here, the elephant. Can you check in here [his sticker book] for him?)*

Then she helped the child open the sticker book and provided the target prompt:

> "Der hat die oder der braucht die."
> *(He's got it or he needs it.)*

In the *OVS condition*, the experimenter directed the child's attention to the object, that is, to the particular sticker in question in order to set it up as a topic and consequently motivate the OVS structure. She modelled the discourse context using a full NP + pronoun sequence twice in order to refer to the object referent:

> "Oh, eine Biene! Schau mal, die hat zwei Fühler. Meinst Du, die kleine Biene gefällt dem Elefant? Schau doch gleich mal hier [in seinem Aufkleberheft] nach wegen der!"
> *(Oh, it's a bee! Look here, it's got two antennae. Do you think the elephant likes the little bee? Can you check here [in his sticker book] about it?)*

Then she helped the child open the sticker book and provided the target prompt:

> "Die hat der oder die will der."
> *(This one he has or this one he wants.)*

In both conditions, after the prompt, the children checked the respective page in the sticker book and then communicated what they saw, using (in virtually all cases) one of the statements they had just received as a prompt, albeit sometimes omitting arguments in their responses. For the filler trials without a prompt, the experimenter modelled the context according to condition, but left the child to her own devices to communicate if the sticker was needed or not. Occasionally, she would ask "*And?"* and look expectantly, if the child did not respond. In very few cases she asked *"Can you tell me?"* which then triggered a response from the child.

Between each of the eight sticker trials in each condition, the child was asked to turn the page and describe the geometric shape they saw, as well as to name the colour of the next page. The book then remained open at the coloured page and the new trial started. After the sticker book corresponding to the first experimental condition (SVO or OVS) was completed, the child and the experimenter did jigsaw puzzles together. After a little while, the experimenter asked the child if she could also help complete the elephant's other sticker book and proceeded to the remaining experimental condition.

3) Design, coding & analysis

There were two conditions according to the two different word orders, an SVO condition and an OVS condition. The four prompted trials per condition were included in the analysis. There were four measures across the conditions: 1) initial subject ([S]VO), 2) initial object ([O]VS), 3) final subject (OV[S]) and 4) final object (SV[O]). According to these measures we calculated the proportions of omission for each argument out of four trials per condition and child. We then compared the rates of initial argument omission with the rates of final argument omission (subjects and objects), and the rates of omission of subjects vs. objects in each position. Since the data were not normally distributed, we used non-parametric tests.

B. *Results*

Table 3 provides an overview of the mean omission rates of initial versus final arguments across all participating children.

Table 3. Omission rates

	initial	final
subject	.39	.08
object	.50	.07

First, we looked at the difference in omission rates of final and initial arguments. Overall, the children omitted initial arguments more often than final arguments. A Wilcoxon signed ranks test showed that this difference was significant, both for initial subjects (*Mdn* = .25) versus final subjects (*Mdn* = 0): $T = 6$, $p = .002$, r = -.06, *2-tailed* and for initial objects (*Mdn* = .58) versus final objects (*Mdn* = 1): $T = 0$, $p = .001$, r = -.7, *2-tailed*. Even though the children omitted slightly more initial objects (50%) than subjects (39%), this difference was not significant ($T = 28$, *ns*), nor was there any difference in the omission rates between final subjects and objects ($T = 5$, *ns*).

C. *Discussion of Study 1*

In study 1 we investigated whether the purported subject-object asymmetry arises from information structural properties such as word order and topicality as opposed to argument type (i.e., subject and object). Our results show that when subjects and objects assumed similar information structural roles, the subject-object asymmetry was neutralized. Three-and-a-half-year-old children omitted both types of arguments when they constituted given information and were placed in sentence initial position. In contrast, the children never omitted their verbs. Verbs were contrasted with one another in the current experiment, and under a discourse-pragmatic account, contrasted elements are more likely to be realized in the discourse [24]. This selective omission of elements in young children's language production has been dubbed the "Principle of Informativeness" by Greenfield and Smith [12]: young children are sensitive to the information structure of events and – during the early stages of language acquisition – tend to encode only the most informative aspects of an event. Those aspects that are presupposed – or given – in the situation are generally omitted, although adults would tend to realize these elements pronominally rather than omit them altogether.

We can confirm Hamann's [9] findings that initial objects are omitted more often than final objects, but our results show that this is extended to subjects as well. Thus, a subject-object asymmetry per se does not exist: Sentence-final subjects are omitted significantly less often than sentence-initial subjects even when factors such as giveness, animacy and pronominal realization were controlled for.

The data from study 1 showed that children omit subjects as well as objects at similar rates from utterance initial position in a language that allows for such pragmatically motivated *topic drop*.

III. STUDY 2: MODELLING ARGUMENT OMISSION

Yet, the question remains if and how the observed omission patterns are learnt and to what extent they are reflective of the input. Some accounts view argument omission as a performance limitation in production, such that arguments in initial position are omitted as a function of the length of an utterance [2]. An alternative account posits limitations of the learning mechanism and attributes argument omissions to a learning bias whereby children learn from the end of utterances [5], [6], and only gradually build up fully formed sentences.

Related to this is the role of the input a child receives. Do performance limitations in learning interact systematically with the frequencies of fragment sentences of the types [S]VO or [O]VS as they are observed in the input? There are different predictions for omission rates depending on the particular source of performance constraint, for example, production-based accounts would predict omission patterns somewhat independent of the input characteristics: initial arguments should be omitted to the same degree regardless of the frequencies of particular constructions. Learning-based accounts, however, might predict omission rates that are more dependent on the nature of the input, for example the frequency of omission within a particular construction: on the one hand, a mechanism based on learning holds that elements that are omitted in the input will also be omitted in child speech. On the other hand, this mechanism also means that children fail to learn the beginnings of long utterances that have low frequency, but are quicker to learn the beginnings of long utterances that are higher frequency. To address these possibilities, in study 2 we used a computational model that implements such a learning constraint to examine patterns of preverbal omission in the model's output as it learned from a corpus of child directed speech.

A. Input analysis

We first performed a caregiver-input analysis in order to determine the relative frequency of transitive SVO and OVS structures of the type used in Study 1, as well as their associated fragments, [S]VO and [O]VS. This allowed us to form a more complete understanding of the input to children that might relate to performance in Study 1. Based on these frequencies we were later able to compare the model's output to the input it received in order to see if output and input are related in a meaningful way as predicted by a performance account based on learning, and also, how the model output compares to the children's performance in study 1. We analysed the Leo corpus [25], a dense database of a German-speaking child that contains approximately 220,000 child-directed utterances, 65,000 of which are declaratives that contain a verb. In line with the materials used in the experimental study, the analysis was constrained to utterances with singular subjects, and modal utterances were excluded. Different to study 1, however, 1st and 2nd person contexts were included as well. This was done because 3rd person contexts were very rare in the corpus, a circumstance that may be particular to the samples of such caregiver-child interactions where 3rd persons are not present and thus, less often referred to. The total number of utterances in the corpus that fit these criteria is approximately 25,000. Transitive utterances were then extracted by selecting all utterances where a finite verb form was directly followed by a pronoun or a noun phrase as indicators of a transitive context. These utterances were then classified as either SVO (N = 2158) or OVS (N = 2341) on the basis of this pronoun or noun phrase. Selected utterances were then subdivided into items with an absent (N = 2111) or present (N = 2388) preverbal element. Utterances that were classified as lacking a preverbal element were subsequently hand-coded in order to verify that a preverbal element was actually missing. As we were interested in the frequency of initial subject and object omissions, this hand-coding excluded from the analysis any utterances that were imperatives or questions which resemble constructions with missing initial arguments by virtue of being formed by S or O inversion. The analysis yielded a total of 3453 utterance tokens. The distribution of SVO and OVS with absent and present preverbal elements is shown in Table 4.

Table 4: Distribution of SVO and OVS structures with present and absent preverbal elements in caregiver input

	present	absent	% omission
SVO tokens	1401	165	.11
OVS tokens	987	900	.48

As can be seen in Table 4, SVO tokens occurred a total of 1566 times in the input and OVS tokens occurred a total of 1886 times. Omission of preverbal elements is relatively rare in SVO structures (11%), but highly frequent in OVS structures (48%). Nearly half of all OVS tokens occur without the preverbal object. These numbers suggest that omission is more productive in OVS than SVO constructions. A logistic regression model run in R [26] with the outcome variable as preverbal argument provision/omission and the predictor variable as sentence structure (SVO/OVS) was significant ($\beta = -2.05$, $p < .001$), demonstrating that preverbal argument omissions were more common in OVS than SVO structures overall. These findings are in accordance with the Hamann [9] and Dittmar et al. [22] analyses and speak to a reversed asymmetry in German.

B. Simulating pre-verbal omission levels in MOSAIC

MOSAIC (Model Of Syntax Acquisition In Children) is a computational model of language acquisition that has successfully been applied to a number of phenomena in child language [6], [27], [28]. The version of MOSAIC used for these simulations is the one described by Freudenthal et al. [6], to which the reader is referred for additional detail. The basis of MOSAIC is an n-ary tree structure or discrimination net that is used to store (partial) utterances that have been shown to the model. MOSAIC is a simple distributional analyser that employs no built-in abstract linguistic knowledge. MOSAIC builds up its representation of the utterances it sees by starting at the right edge of the utterance and slowly working its way to the front, creating a simple discrimination net consisting of nodes and arcs that are used to store words and phrases that it has been exposed to. MOSAIC will only create a node encoding a phrase when everything that follows that phrase in the utterance has already been encoded in the network. Thus, if MOSAIC sees the utterance 'he goes home', it will first create a node encoding the word 'home'. Only when the node 'home' has been created will the model consider encoding the phrase 'goes home', and the node 'he goes home' will only be created after the phrase 'goes home' has been encoded. Initially, the probability of creating a node is low, and it increases as the model has seen more input. Output is generated from MOSAIC by traversing the network and generating all the phrases it has encoded. Output from MOSAIC thus consists of a corpus of (utterance-final) phrases that can be directly compared to corpora of adult, or child speech. Unlike corpora of human speech, MOSAIC output only contains utterance types (i.e., it contains no duplicate utterances). With increased training, MOSAIC encodes increasingly long utterances, and hence generates increasingly long output utterances, comparable to those of language-learning children.

C. The simulations

MOSAIC was trained on all (declarative) child-directed input (65,000 utterances) from the Leo corpus. The caregiver data from the Leo corpus was fed through MOSAIC repeatedly, for a total of 44 times, and output of increasing length was collected from MOSAIC after every exposure to the input. The output was then analysed with respect to the presence or absence of preverbal subjects and objects. In order to determine the influence of child-directed speech, the analysis of MOSAIC's output was restricted to full SVO or OVS utterances, and (O)VS and (S)VO utterances that contained a finite verb form and a post-verbal pronominal or nominal subject or object, that could be traced back to the hand-coded input utterances. This was done to ensure that where there were missing preverbal elements, these were indeed subjects or objects and not imperatives or items with missing question marks (that is, questions incorrectly coded as declaratives in the corpus). Thus, the restriction ensured that output utterances that were labelled as having the preverbal element missing were correctly classified as such.

The output from MOSAIC was analysed in 5 developmental stages ranging between 38 and 44 exposures to the declarative input corpus. The model output MLU in terms of target utterances ranged from 3.26 to 4.33, which broadly aligns with typically developing German children's MLUs at ages 3;3 to 3;8 (the age of participants in Study 1)[2]. For any given output, we determined the source utterance (that is, the input utterance something was learned from), in order to see if there was a preverbal element and then coded whether it was present in the output utterance. Using this method, two different analyses were carried out. In the first analysis, only output utterances that were learned from fully formed input utterances (i.e. utterances that included the preverbal element) were analysed. This analysis was aimed at determining if the relative frequencies and distributional statistics of (fully formed) SVO and OVS structures in the input differ to such an extent that the model shows differential rates of omission of preverbal elements. The model learns from the end of utterances, and the likelihood of it learning sequences of words is dependent on co-occurrence frequencies; thus omission rates for preverbal subjects and objects might differ as a result of differing frequencies of SVO and OVS in the input. In the second analysis output utterances that were learned from fragments, i.e., verb plus post-verbal argument ([S]VO or [O]VS), in the input were included in the analysis. A comparison between these analyses allows for an investigation of the role of fragments in determining omission rates in the model.

D. Results

As can be seen in Tables 5. and 6., omission rates decrease with increasing exposure to the input, though MOSAIC does not reach the low levels of omission displayed by the older children in the study. This latter finding is not surprising since MOSAIC's ability to 'unlearn' items is limited. That is, incomplete phrases that have been learnt early in development will continue to be produced once the longer phrases from which they have been learnt are encoded in (and produced by) the network.

Table 5: Preverbal omission rates in SVO and OVS structures learnt from fully formed utterances

model run	38	40	42	43	44
SVO omission rate (N)	.66 (108)	.51 (479)	.47 (1023)	.44 (1344)	.42 (1711)

[2] MLUs of typically developing children were calculated using language transcripts from seven German-speaking children at ages 3;3 to 3;8, available through the CHILDES database [29]. Observed MLUs ranged from 3.00 – 4.89 (M = 3.84) which fit broadly with the MLU range of the model's output generated at alternating runs from 38 to 44 (MLU 3.26 – 4.33, respectively).

OVS omission rate (N)	.66 (191)	.57 (409)	.49 (762)	.47 (966)	.44 (1196)

Table 6: Preverbal omission rates in SVO and OVS structures learned from fragments as well as fully formed utterances

model run	38	40	42	43	44
SVO omission rate (N)	.68 (168)	.54 (506)	.49 (1070)	.46 (1411)	.44 (1799)
OVS omission rate (N)	.75 (259)	.68 (552)	.61 (987)	.58 (1222)	.55 (1477)

It is also apparent from Table 5. that there is very little difference in subject and object omission rates for utterances that have been learned from fully formed structures. Omission rates from OVS structures are very slightly higher than from SVO structures, but there is very little difference in the distributional statistics of the two construction types. Both constructions are learned equally quickly.

As can be seen in Table 6., which lists omission rates for utterances learnt from all contexts, the inclusion of utterances learnt from fragments results in a different pattern. Preverbal elements from [O]VS structures are omitted at higher rates than from [S]VO structures in all developmental phases. The size of this difference ranges from 7 to 14%, and exceeds 10% in 4 out of 5 instances.

To test whether these observed differences were significant, a logistic regression model was fitted to the data in R [26] with the outcome variable as preverbal argument provision/omission and predictor variables of sentence structure (SVO/OVS), model run (characterised as a developmental series from one to five), and learning source (full vs. fragment). In this model, all predictors were significant, showing that preverbal argument omissions significantly decreased as the model was exposed to more input (β = -.15, p < .001), with SVO rather than OVS structures (β = -.18, p < .001), and for utterances learned from full utterances rather than fragments (β = -3.50, p < .001). There also appears to be an interaction between sentence structure and learning source such that OVS structures learned from fragments are particularly prone to preverbal omissions, however the lack of variance in this cell (100% omissions) makes the addition of an interaction term to the statistical model problematic.

The difference between omission rates from [S]VO and [O]VS contexts matches the difference displayed by the young children in the experimental study well. Taken together with the results from fully formed constructions, these results suggest that children's early high preverbal object omission rates in [O]VS constructions can be understood in terms of sensitivity to the frequency of [O]VS fragments in the input.

E. *Discussion of Study 2*

The findings from study 2 address several issues. First, the data from the caregiver input analysis lend support to Dittmar et al.'s [22] preliminary finding that (sentence-initial) objects are omitted more often than (sentence-initial) subjects in adult German. The corpus used in study 2 was considerably larger than the Dittmar et al. sample, yet the results are consistent: object omission in [O]VS is a more productive phenomenon than is subject omission in [S]VO.

The results of the computational model of syntax acquisition are consistent with a performance-based approach that places limitations on learning: A model implemented with a recency-bias (resulting in learning from the end of utterances), produces omission rates similar to those found in German child and adult language. Over and above its learning bias, however, the model is sensitive to omission patterns in the input: initial objects are omitted more often than initial subjects, and omissions of both argument types decrease "with age" (i.e., over iterations) – mirroring adult German.

It is important to note that unlike in study 1, givenness and topicality were not explicitly manipulated features in study 2. Due to the nature of caregiver-child corpora, many of the fragments stem from 1st / 2nd person contexts. 1st and 2nd persons are traditionally considered given. However, the assumption of omission indicating topicality/givenness in the corpus data is inherently circular: given referents are omitted and omitted referents are given. In the absence of fully observable or manipulatable discourse contexts, we cannot determine with certainty that the omissions in the corpus are reflective of givenness/topicality. Future work will have to address this circularity and offer approaches to modeling which can incorporate discourse context quantitatively. Regarding the manipulation of topicality, we will turn to this issue in the general discussion section.

IV. General discussion

The present studies investigated whether argument omission is a discourse phenomenon according to which sentence initial and given/topical arguments are omitted regardless of subject- or object-hood. The relatively free word order of German allowed for a more refined look at the subject-object asymmetry and our data suggest that the asymmetry might best be explained by an information structure account. In a canonical SVO construction, the roles of subjects and objects are maximally distinct as defined in terms of transitivity [23], and both arguments obey Preferred Argument Structure constraints [15]. That is, the mostly given subjects are frequently omitted, whereas objects, which contain new information, are predominantly expressed. In study 1, when objects assumed information structural characteristics similar to canonical subjects, they were omitted just as frequently as subjects. Study 2 showed that a learning bias interacts with the input such that a computational model produced initial object-less sentences at higher rates than subject-less sentences.

With regard to 1st/2nd vs. 3rd person contexts, the model and experimental data provide useful complements for one another. Discourse pragmatic accounts conventionally consider 1st and 2nd person contexts less important because “you” and “I” are always deictically present and highly accessible in the interlocutor’s knowledge state [30]. Thus, in these discourse contexts omission is very productive. 1st and 2nd person contexts were included in the caregiver input analysis, because 3rd person contexts were not frequent enough in the input sample. In study 1, however, we manipulated the discourse context in such a way as to make a 3rd person referent just as accessible, i.e., equally deictically present and activated in the interlocutor’s knowledge state. The omission rates in study 1 and study 2 were similar and lend support to the proposal that omission is a discourse phenomenon. That is to say, when referents were construed as given and topical, 3rd person arguments were omitted just like 1st/2nd person referents.

The overall omission rates of both argument types are similar across the two studies. Higher initial object omission rates were found in the caregiver input, and both the model and the children in study 1 omitted more initial objects than subjects. The finding that both performed similarly adds to the growing body of evidence that frequency patterns influence children’s early language.

A. The German object-first construction as a marker of topicality

Both SVO and OVS are productive word orders in German, however, SVO generally represents the more frequent construction, and is considered unmarked with regard to its information structure [31]. While there is a strong association between subjects and topics in canonical, unmarked SVO constructions, not all subjects are in fact topics. It follows that, although subject omission rates in adult and child German are considerable, a subject-first construction is not in and of itself a reliable cue for a topical – and thus omissible – subject. The results of our studies provide preliminary evidence: the children in study 1 omitted fewer subjects than objects (albeit not significantly so) and the model in study 2 produced fewer subject than object omission rates.

Similarly, the input analysis in study 2 showed that initial subjects were omitted at a rate of about 11% from the caregiver speech. Objects, on the other hand, were omitted at much higher rates (48%). This reversal of the traditional subject-asymmetry suggests that the information structure of the German object-first construction is marked: an object placed in sentence-initial position represents a strong topic cue and thus experiences higher omission rates.

The children in study 1 and the model in study 2 omitted arguments at similar rates. However, the model arguably produced these omissions without any sensitivity to information structural cues other than position. With regard to the model’s input, givenness/topicality was not explicitly manipulated. [S]VO and [O]VS fragments in the caregiver input are likely instances of topic-comment information structures due to the prevalence of 1st and 2nd person contexts of caregiver-child focused corpora. Put differently, in study 1, the discourse context was carefully constrained in order to achieve topicality experimentally while study 2 is based on a dense sample of speech produced in a communicative setting, and topicality is assumed to arise naturally.

The model’s output shows sensitivity to different omission rates in the input, such that it produces more omissions for objects than for subjects. Yet, while the model learns from different information structural constructions, it produces omissions partly as a function of its inbuilt learning bias, that is, independent of topicality. The learning bias, however, can be understood as reflective of the cognitive architecture in place for linguistic information management: Across languages, there is a statistical tendency for speakers to introduce given information mainly in subject position of transitives whereas new information tends to be realised in the object position of transitives [11]. Based on this preference, canonical information structures can be seen as providing the necessary "architecture for function" [15, p. 45]. Different constructions are supplying the speaker and the hearer with a "predictable locus for unpredictable work" [15, p. 46], that is, canonical constructions support the tracking of given and new information. The model's recency bias can then be understood as an implementation of speakers' tendency to place new information at the end of transitive utterances. This would affect both [S]VO and [O]VS structures equally. However, the data presented in these studies (in accordance with prior evidence) showed that omission in German OVS structures is more frequent, perhaps suggesting that German OVS constructions predictably signal topicality – and more so than canonical SVO constructions.

B. Postverbal drop and the syntax-discourse interface

Based on the fact that word order (SVO or OVS) signals topicality in German, Hamann [9] argued that topic drop is possible only in preverbal, but not postverbal position. Yet in study 1, some children omitted postverbal subjects and objects to some extent (15%). These postverbal arguments were also given in our study: both experimenter and child continued to make reference to what they had already shared with each other, a situation not unlike normal discourse contexts, which mainly revolve around given topics [15].

These instances of postverbal topic drop illustrate that argument omission cannot be explained by a purely positional account, even though topic and position strongly correlate. While postverbal elements are less likely to be omitted by virtue of their prototypical construal as the new element in the utterance, in our study they were not only given, but also deemphasized by placing the focus on the contrastive action. Similarly, Goldberg's [32] *Principle of omission under low discourse prominence* holds that objects of causative verbs can be omitted when they are deemphasized in favour of the action.

A study by Salomo and colleagues [33] also showed that discourse context influences postverbal omission. In an experiment targeting children's question-answering behaviour, the authors presented 2;3-year old children with video clips of a succession of transitive scenes and asked a predicate focus question, which targeted action and patient (What's the AGENT doing?). Whereas the given subject referent was dropped virtually every time, the children modeled their predicate-focus answers based on the changing element: When the patient was the new element, the children provided mainly action + patient answers. When the action was new, the children's answers mainly expressed just the action. Thus, they were effectively omitting the object/patient of a transitive action – in postverbal position. Questions provide for a very strong discourse context, and show clearly that discourse accounts for the structure of referring expressions.

V. CONCLUSION

The present studies provide evidence illustrating that argument omission is a discourse phenomenon, and as such it is sensitive to the information status of arguments rather than their grammatical category. Word order variability in German allows for both argument types to occur in sentence initial position when they are given/topical and therefore, can be dropped. Experimental and computational data suggest that while omission can be accounted for by performance limitations in learning, it mirrors the frequencies with which fragment structures occur in the input.

REFERENCES

[1] Hyams, N. (1987). The theory of parameters and syntactic development. In: Williams, E. & Roeper, T. (Ed.). *Parameter Setting*. Dordrecht: Reidel.

[2] Bloom, P. (1990). Subjectless sentences in child language. *Linguistic Inquiry*, *21*, 491-504.

[3] Wang, Q., Lillo-Martin, D., Best, C.T. & A. Levitt (1992). Null subject versus null object: Some evidence from the acquisition of Chinese and English. *Language Acquisition*, *2*, 221-254

[4] Bloom, L. (1970). Language Development. Form and function in emerging grammars. Cambridge, MA: MIT Press.

[5] Slobin, Dan I. (1973). Cognitive prerequisites for the development of grammar. In: Ferguson, C. A. & Slobin, D. I. (Ed.). *Studies of Child Language Development*. New York: Holt, Rinehart and Winston, 240-249.

[6] Freudenthal, D., Pine, J. M. & Gobet, F. (2007). Understanding the developmental dynamics of subject omission: the role of processing limitations in learning. *Journal of Child Language*, *34*, 83-110.

[7] Hyams, N. (1986). *Language acquisition and the theory of parameters*. Dordrecht: Reidel.

[8] Hyams, N. & Wexler, K. (1993). On the grammatical basis of null subjects in child language. *Linguistic Inquiry*, *24*, 421-459.

[9] Hamann, C. (1996). Null arguments in German child language. *Language Acquisition*, *5*, 155-208.

[10] O' Grady (1997). *Syntactic Development*. Chicago: University of Chicago Press.

[11] Du Bois, J.W (1987). The discourse basis of ergativity. Language 63(4): 805-855.

[12] Greenfield, P.M. & Smith, J.H. (1976). *The structure of communication in early language development*. New York, NY, Academic Press

[13] Lambrecht, K. (1994). *Information structure and sentence form*. Cambridge: Cambridge University Press.

[14] Allen, S.E.M. (1997). A discourse-pragmatic explanation for the subject-object asymmetry in early null arguments: The Principle of Informativeness revisited. In A. Sorace, C. Heycock & R. Shillcock (Eds.), *Proceedings of the GALA '97 Conference on Language Acquisition (pp. 10-15)*. Edinburgh: University of Edinburgh.

[15] Du Bois, J. W. (2003). Argument structure. Grammar in use. In Du Bois, Kumpf & Ashby (Eds.), *Preferred Argument Structure: Grammar as architecture for function* (pp. 11-60). Amsterdam: Benjamins.

[16] Clancy, P. M. (2003). The lexicon in interaction: Developmental origins of preferred argument structure in Korean. In Du Bois, Kumpf & Ashby (Eds.), *Preferred argument structure: grammar as architecture for function* (pp. 81-108). Amsterdam: John Benjamins.

[17] Jacobs, J. (2001). The dimensions of topic-comment. *Linguistics, 39*(4), 614-681.

[18] Frey, Werner (2004). The grammar pragmatics interface and the German prefield. *Sprache und Pragmatik, 52*, 1-39.

[19] Frey, W. (2007). Some contextual effects of aboutness topics in German. In Späth, A. (Ed). *Interfaces and interface conditions*. Berlin: de Gruyter, 329-348.

[20] Szagun, G. (2004). Learning by ear: On the acquisition of case and gender marking by German-speaking children with normal hearing and with cochlear implants. *Journal of Child Language, 31* (1), 1–30.

[21] Clahsen, H. (1991), Constraints on parameter setting. A grammatical analysis of some acquisition stages in German child language, *Language Acquisition, 1*, 361-391.

[22] Dittmar, M., Abbot-Smith, K., Lieven, E., & Tomasello, M. (2008). German children's comprehension of word order and case marking in causative sentences. *Child Development, 79*(4), 1152-1167

[23] Næss, Å. (2007). *Prototypical transitivity*. Amsterdam: John Benjamins.

[24] Allen, S.E.M. (2000). A discourse-pragmatic explanation for argument representation in child Inuktitut. *Linguistics*, *38*(3), 483-521.

[25] Behrens, H. (2006). The input-output relationship in first language acquisition. *Language and Cognitive Processes*, *21*, 2-24.

[26] R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

[27] Freudenthal, D., Pine, J.M. & Gobet, F. (2006). Modelling the development of children's use of Optional Infinitives in Dutch and English using MOSAIC. *Cognitive Science*, *30*, 277-310.

[28] Freudenthal, D., Pine, J.M., Aguado-Orea, J. & Gobet, F. (2007). Modelling the developmental patterning of finiteness marking in English, Dutch, German and Spanish using MOSAIC. *Cognitive Science*, *31*, 311-341.

[29] MacWhinney, B. (2000). The CHILDES Project: Tools for analyzing talk. Third Edition. Mahwah, NJ: Lawrence Erlbaum Associates

[30] Allen, S.E.M., Skarabela, B. & Hughes, M. (2008). Using corpora to examine discourse effects in syntax. In Behrens, H. (Ed). *Trends in corpus research: Finding structure in data.* Amsterdam: Benjamins*,* 99-137.

[31] Lenerz, J. (1977). *Zur Abfolge nominaler Satzglieder. Studien zu deutschen Grammatik 5*. Tübingen: TBL Verlag Gunter Narr.

[32] Goldberg, A. (2001). Patient arguments of causative verbs can be omitted: the role of information structure in argument distribution. *Language Sciences, 23*, 503-524.

[33] Salomo, D., Lieven E. & Tomasello, M. (2010). Young children's sensitivity to new and given information when answering predicate-focus questions. *Applied Psycholinguistics*, *31*, 101-115.