Universiteit Leiden
The Netherlands

Constructions emerging : a usage-based model of the acquisition of grammar

Beekhuizen, B.F.

Citation

Beekhuizen, B. F. (2015, September 22). *Constructions emerging : a usage-based model of the acquisition of grammar*. *LOT dissertation series*. LOT, Utrecht. Retrieved from https://hdl.handle.net/1887/35460

Version:	Corrected Publisher's Version
License:	Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden
Downloaded from:	https://hdl.handle.net/1887/35460

Note: To cite this publication please use the final published version (if applicable).

Cover Page

Universiteit Leiden

Leiden University Repository

The handle http://hdl.handle.net/1887/35460 holds various files of this Leiden University dissertation

Author: Beekhuizen, Barend
Title: Constructions emerging : a usage-based model of the acquisition of grammar
Issue Date: 2015-09-22

CHAPTER 4

Modeling the acquisition of meaning

In chapter 3, I presented a computational model of the acquisition of constructions. These constructions are incrementally learned from linguistic usage events, being pairings of an utterance and several situations, and are used to analyze novel linguistic usage events. An important question that remains is what these linguistic usage events consist of.

In this chapter, we will look at the way in which the conceptual side of the linguistic usage events (viz. the situational context) is represented in input items. What are the properties of these situational contexts? The motivation for studying this, is that computational models of symbol acquisition (word learning as well as constructional learning) often make strong assumptions about the nature of the set of communicated concepts at which the learner arrives independently of language. These assumptions, however, often do not rely on empirical accounts of how a learner constructs this set. The representations acquired by a computational model depend on what is in the input, and it is therefore equally important to provide the model with input items that are as realistic as possible.

This chapter sets out to provide such an account, looking primarily at the environmentally available information. The insights resulting from this investigation are then used to formulate a procedure for simulating realistic situational contexts in which utterances are produced. This procedure will then be used to provide the learning model with input.

4.1 Three problems in acquiring meaning

As I argued in the previous chapter, it is a logical necessity that the child has some coarse understanding of what an utterance refers to when she hears it (O'Grady's Interpretability Requirement). In studies on symbol acquisition it is often tacitly assumed that *all* of the meaning is correctly understood by the child, and moreover, that *only* the correct meaning is understood, i.e., there are no 'distracting', non-communicated concepts. Admittedly, the latter assumption is less frequently made, as most researchers recognize that 'distractors' are present in the space of candidate meanings (that is: the set of considered conceptualizations communicated with the utterance), and, in fact, this constitutes a learnability problem by itself (cf. Quine's (1960) Gavagai problem). Nonetheless, this assumption is still used as the starting point of many computational modeling studies, as we will see later. Let us, for future reference, call the 'all-and-only' assumption Assumption 1, with two corollaries, Assumption 1a and Assumption 1b:

- **Assumption 1**: The correct set of concepts to be mapped onto the utterance is active in the mind of the learner
  - **Assumption 1a**: All of the concepts to be mapped onto the utterance are active in the mind of the learner
  - **Assumption 1b**: Only the concepts to be mapped onto the phonological substrings of the utterance are active in the mind of the learner

When we do find the assumption in an explicit form, for example in O'Grady (1997, 260) or Wexler & Culicover (1980, 80), it is presented as a requirement for the acquisition of form-meaning pairings, but no supporting evidence for its veracity is provided. We can wonder, however, to what extent the assumption in its strong form holds. Even in a weaker form (most of the concepts are available, and there are few distracting ones), we would like to know the magnitude of the learning problem when the nature of the input deviates from Assumptions 1a and 1b.

We can quantify and conceptualize the deviation as follows. First, are all concepts the speaker wants to communicate with an utterance part of the candidate meanings? We will call this issue, corresponding with Assumption 1a, the question of *noise* (cf. Siskind 1996, 50). When we, in a simplifying manner, assume that the candidate meanings $M_{\text{candidate}}$ and the actually communicated meanings $M_{\text{communicated}}$ are sets of communicated elements (be they features, entities, or whole propositions), we can measure the noise as follows:

$$Noise = 1 - \frac{|M_{\text{candidate}} \cup M_{\text{communicated}}|}{|M_{\text{communicated}}|} \tag{4.1}$$

That is: what proportion of the set of communicated concepts are actually present in the set of candidate meanings? When $noise = 0$, all communicated

concepts are part of the set of candidate meanings, whereas no element of the communicated concepts is in the set of candidate meanings when $noise = 1$.

Second, to what extent are only the situations and objects the speaker wants to refer to present in the set of candidate meanings? How many concepts are there that are not referred to in the utterance, and thus increase the referential uncertainty? We will call this issue, corresponding with Assumption 1b the question of *uncertainty* (cf. Siskind 1996, 40).

$$Uncertainty = 1 - \frac{|M_{\text{candidate}} \cup M_{\text{communicated}}|}{|M_{\text{candidate}}|} \tag{4.2}$$

$uncertainty$ thus measures what proportion of the set of candidate meanings is not communicated by the utterance. $uncertainty = 1$ means that the candidate meaning $M_{\text{candidate}}$ consists fully of non-communicated concepts, whereas $uncertainty = 0$ means that $M_{\text{candidate}}$ is entirely made up of communicated concepts.

Uncertainty, like noise, can take place on many levels: conceptual features may be unavailable (*conceptual noise*), or superfluously available (*conceptual uncertainty*, but also entire entities (objects, events, each of which can be described with a number of conceptual features; *referential noise* and *referential uncertainty*), and even full propositions (*propositional noise* and *propositional uncertainty*). When operationalizing noise and uncertainty for specific cases, we have to specify on what level this noise takes place, but for the current purposes, the use of sets M generalizes over all three levels: it could refer to a set of conceptual features, entities, or full propositions.

Once we acknowledge that learners probably operate under non-zero uncertainty levels, another problem presents itself: is the non-target part of the space of candidate meanings (the concepts not referred to) independent from the target part of that space? If there are dependencies, this affects the ease of learning: if certain elements in the space of candidate meanings are often found together with other elements, the learner will have a harder time to use cross-situational statistics, to name one learning mechanism, in order to disentangle them (cf. Siskind 1996, 75). Examples of dependencies in the candidate meaning would be different conceptualizations of the same event (e.g., 'chase' and 'flee'), or meronymic relations (e.g., 'rabbit' and 'ears'), but also concepts that in principle engender different construals, but simply occur together often (e.g., 'sitting at the table' and 'eating', for the young child). Although the full extent of this problem is beyond the scope of this chapter, we will briefly touch upon the last kind of dependence, quantifying it and using the insights in our simulation procedure.

Independently researching the environmental and cognitive sources of the set of candidate meanings is relevant to the understanding of the cognitive mechanisms responsible for forming the symbolic mappings. Experimental work like Yu & Smith (2007) has demonstrated that learners can use the mechanism of keeping track of cross-situational co-occurance statistics in acquiring

symbolic pairings. However, in several simulations and experiments, Smith, Smith & Blythe (2011) and Blythe, Smith & Smith (2010) point out that, using varying amounts of referential uncertainty, there are different strategies that lead to optimal learning behavior: with higher levels of referential uncertainty, a more heuristic variant of cross-situational learning explains the subjects' performance in learning form-meaning mappings better than with lower ones. This means that, before we can determine (experimentally) what mechanisms underly the acquisition of symbolic pairings, we have to understand in what range the noise and uncertainty realistically fall.

This point becomes especially important in computational simulations of the symbol acquisition process. In these studies, a formal operationalization of a proposed cognitive mechanism is tested on data containing pairs of utterances with meaning representations, thought to reflect the set of candidate meanings. However, if amount of noise and uncertainty in the set of candidate meanings reflects the simplistic assumption, or the deviation from this assumption is not empirically grounded, then the mechanisms under scrutiny cannot be properly evaluated. Quantifying actual noise and uncertainty levels on the basis of empirical data, for instance spontaneous caregiver-child interaction, allows us to do so.

A note on terminology is in place here. The term *noise*, as borrowed from signal processing, is often used as a generic term concerning all undesirable modulations of the signal, including both noise in the narrow sense, as I defined it in this chapter, as well as uncertainty. Although ambiguity between the superordinate term and a subordinate is in principle undesirable in scientific discourse, and can lead to needless misunderstandings, it is at the same time not beneficial to introduce completely new terms. *Noise* is used in both the superordinate and subordinate sense in the literature and the value can be contextually determined (in pairs such as *noise and uncertainty*, it always means the absence of information in the signal, not both the absence and the superfluence).

4.2 The informativeness of the situation

4.2.1 Earlier research

Linguistic research on the informativeness of the situation

Studies discussing situational availability are rather scarce, and are typically framed on a propositional level, that is: does the utterance refer to a full situation in the here-and-now of the interactive setting. Moerk (1972) discusses the nature of the interaction between mothers and children, and remarks that "The mother [...] model[s] nearly continuously for the child the process of translating the structure of the objective environment and their own actions into verbal utterances", thus suggesting that little noise is to be expected in the

mothers' input. However, Moerk did not systematically investigate this, and focusses only on what could be seen as the lack of noise: whenever the caregiver talks to the child, the situation referred to is hardly absent. Cross (1977) presents features of child-directed language that are predictive for the child's vocabulary size at certain ages. She discusses in the appendix four features related to the referential nature of the mother's utterance, namely whether the utterance referred to 1) a child-controlled event, 2) a mother-controlled event, 3) other persons or objects present or 4) something outside of the here-and-now. She defines the here-and-now of the speech situation as the time span between the preceding and current conversational turn. Of the four features, the first is significantly negatively correlated with vocabulary size, meaning that mother will refer less to the child's actions the more sophisticated a language user the child is. Furthermore, the third is significantly positively correlated with vocabulary size, meaning that the mother will refer more to situations slightly more distal from the here-and-now the more advanced the child's language abilities are. As Cross provides no raw frequencies, we cannot determine the precise situational availability in her data. Again, in Cross' study, only the referential nature of the whole utterance is studied, and the question of uncertainty (how much of the current situation is not being referred to), is not addressed.

From the only literature explicitly discussing co-temporal situational presence in naturalistic settings, Gleitman (1990), we know that both Assumptions 1a and 1b are problematic, especially for relational concepts, such as events. Gleitman (1990, 20-22) discusses a paper by Beckwith, Tinkler & Bloom (1989), where the authors describe how in many cases, the event to which a verb refers is absent from the immediate context. This would constitute a case of referential noise. Gleitman further points to the imaginable plethora of cases where the learner does perceive an event, but the label is not used in the utterance, thus bringing about referential uncertainty.

With the scarcity of studies systematically addressing this issue on the basis of naturalistic data, it seems that we know very little about the extent to which the utterances in the input are co-temporally matched with the communicated concepts. It is striking that empirical investigations into the nature of the environmentally given information are so scant, whereas the Interpretability Requirement constitutes a central assumption in acquisitional research.

Modeling approaches deriving candidate meanings from the utterance

Most computational research on acquiring form-meaning pairings focuses on the cognitive mechanisms required to develop an inventory of symbols given an existing set of candidate meanings, rather than on the learner's understanding of the set of candidate meanings itself. Although computational studies on the mechanisms have greatly added to our knowledge of possible cognitive mechanisms, their evaluation remains problematic, as performance may depend to a large extent on the properties of the set of candidate mean-

ings. In this section, I will discuss computational studies of symbol acquisition and the assumptions concerning noise and uncertainty they make.

The first group of studies derives properties of the set of candidate meanings from the linguistic input. Corpora of child-directed speech are mostly not structurally annotated with the situations that co-occur with the utterances, let alone the child's likely mental representation of those. As a means of approximating the situation, several approaches, both in acquiring mappings between single words and their meanings, and in acquiring a grammar with meaningful rules, use the utterance itself to infer the situation it is paired with (Siskind 1996, Chang 2008, Alishahi & Stevenson 2010, Fazly et al. 2010). Taken by itself, this method would constitute a very strong instantiation of the assumption that all and only the correct meanings are present. Most, if not all, authors acknowledge the problematic nature of this assumption, and therefore introduce deviations from the 'all candidate meanings are present' assumption (by removing elements of the set of communicated concepts, thus adding noise) and the 'only the candidate meanings are present' assumption (by introducing additional elements into the set of candidate meanings, thus increasing the uncertainty) so as to make the experiments with the models of form-meaning pairing acquisition more realistic.

Older studies, like Regier (1992) and Bailey (1997) use toy examples with more complex meaning representations than many later studies. However, being toy examples, the input data is generated in such as way that the situation matches the word it is to be associated with. Because of that, we can also group them in the category of utterance-derived candidate meanings.

The addition of noise and uncertainty found in most models of the acquisition of form-meaning pairing is, by itself, a step in the right direction. By adding noise and uncertainty, the models are shown to be robust to noise and uncertainty (see table 4.1 below for some examples). However, few of the works mentioned discuss how the parameter setting for their noise and uncertainty values is motivated. That is: if we add noise, *how much* noise is realistic? And is the amount of noise the same for every conceptual type and every linguistic class? Are verb-to-event mappings noisier, for instance, than noun-to-object-class mappings? The same question can be asked for uncertainty. Crucially, as argued before, the evaluation of the explanatory value of the model depends on its ability to deal with realistic sets of candidate meanings: as long as we know little of what counts as realistic, the evaluations of the models are problematic.[1]

This is not to say that the method of generating situations on the basis of the utterances is useless. In fact, if one has an empirical grounding for the amounts and types of noise and uncertainty that one introduces in the model, this method may be currently the only way to obtain data sets large enough to train our models on, as long as we do not have fully symbolically annotated

[1]Interestingly, only Siskind (1996) explicitly tries to ground the amount of uncertainty and noise in acquisitional studies, citing Beckwith et al. (1989) and Snow (1977).

model	description	parameter settings
Regier (1992)	No noise or uncertainty is added	n.a.
Siskind (1996)	Propositional noise and uncertainty are added	Parametrized: between 0 and 20% of the utterances lacks the target candidate proposition completely and between 10 and 100 non-target candidate propositions are added.
Bailey (1997)	No noise or uncertainty is added	n.a.
Fazly et al. (2010)	Referential noise and uncertainty are added	In 20% of the utterances, one element of the meaning is discarded. Every other utterance's meaning is added as referential uncertainty.
Chang (2008)	No noise or referential uncertainty is added	n.a.
Frank et al. (2009)	Referential noise and uncertainty are as in video data	n.a.
Alishahi & Stevenson (2010)	Conceptual noise and referential uncertainty is added	In 20% of the utterances, one feature of the meaning is discarded. In another 20%, one feature is discarded and then inferred. The meaning may contain more referents than expressed in the utterance.

Table 4.1: The treatment of noise and uncertainty in several models of the acquisition of form-meaning pairings.

descriptions of the situations accompanying the child-directed utterances.

Approaches deriving candidate meanings from empirical sources

The second group of modeling approaches to the acquisition of form-meaning pairings explicitly addresses the issue of what can be gleaned from the situation accompanying the utterance by using videotaped caregiver-child interaction. Typically, this involves manual annotation of the candidate meanings, although early work on video data shows that a mapping between the raw visual input and the raw speech stream is possible too (Roy & Pentland 2002). Ambitious as this project is, it remains limited as a method of studying language acquisition, for two reasons. First of all, the data used by Roy & Pentland (2002) were not from natural dyadic interaction, let alone child-caregiver interaction, which makes the ecological validity of the discourse problematic. Secondly, the focus was on noun-to-object mappings only. Although this does constitute an important part of the acquisition process, we have to move beyond this to gain insight on a more general level. The main reason is that a narrow focus on, for instance, nouns artificially limits the hypothesis space of the learner: the event-like meanings form no uncertainty for the model learning nouns, whereas we expect some uncertainty to be present unless we assume that children start with attending only to objects and assuming that referring to those is the sole function of language.

More recent approaches using video data suffer from the same problem (Frank, Goodman & Tenenbaum 2008). Even if we assume that nouns are more easily learned, and even if knowledge of the noun-object mappings helps bootstrap other things, they artificially keep other kinds of candidate meanings (events, relations, properties) out of the hypothesis space. The contribution of these studies, however, is that they do show us, even for a narrow subset of candidate meanings, what is and what is not available to the learner (assuming that only the visual perception of spatiotemporally aligned objects leads to the availability in the set of candidate meanings). This provides us with the interesting opportunity of establishing empirically the levels of referential noise and (to some degree) referential uncertainty in caregiver-child interaction.

A final approach that is of interest is one in which the focus *is* on a broader class of candidate meanings than just object categories. Fleischman & Roy (2005) had subjects play a game in which one subject had to verbally guide the other subject through a video game world towards a certain goal. The language involved directive and descriptive utterances about the task of the other subject. The learning model received its input data from this experiment: the utterances of the one subject were paired with the actions and the overarching plans behind the actions (opening a door is an action towards the plan of entering a room) for the other subject. This represents a closer approximation of the breadth of candidate meanings than the studies on noun-object mapping acquisition. A point of criticism here could be the ecological validity, as

with Roy & Pentland (2002): the type of discourse is not the same as caregiver-child interacting, although it should be granted that the directive nature of the language and the fact that the subjects had a joint task approximate many situations of child-caregiver interaction relatively closely.

4.2.2 How available are the communicated concepts

What is the information in the actual environment in which children learn words? Narrowing this question down to the two corrolaries of the interpretability assumption, we have to ask what the noise and uncertainty is that children face when starting to develop a lexicon. In this section, I present research addressing these questions.[2]

Materials

Like the second group of modeling studies I discussed, we take videotaped interactions of caregivers and children to be the starting point of our information about the properties of the environment from which the set of candidate meanings is inferred. The interaction has to be relatively typical of the kind of interactions young children and their caregivers have. To this end, I used videotaped interactions of Dutch mothers and 16 month-old daughters playing a game of putting blocks in holes.[3] Games form an interesting setting, as they constitute a typical activity in which the child jointly attends the situation with the caregiver, and in which directive and descriptive language is used (Tomasello & Farrar 1986, 1457). From the 131 available dyads, I selected the first 32. The games were played for about five minutes per dyad, giving a videotaped corpus of 152 minutes (henceforth: the corpus).

Annotation

In the corpus, I transcribed all speech according to CHAT-guidelines,[4] and two assistants coded the video data for the objects, properties and relations in the situations. The transcriptions contained 7842 word tokens (480 types) in 2492 utterances. The language mostly refers to aspects of the game.

The situational coding was done according to guidelines described in Beekhuizen (2011). As the situation consists of just one type of activity (playing the game), the set of objects, properties and relations is relatively limited. The most common object categories are the BUCKET, LID, BLOCKs, HOLEs and

[2]Parts of the research reported in this section was previously published in Beekhuizen, Fazly, Nematzadeh & Stevenson (2013) and Beekhuizen, Bod & Verhagen (to appear)

[3]The data was courteously made available by Marinus van IJzendoorn and Marian Bakermans-Kranenburg of the department of Child Studies at Leiden University.

[4]Available at http://childes.psy.cmu.edu/manuals/CHAT.pdf

type	name	roles
action	GRAB,LETGO,HIT	Agent, Patient, (Instrument)
action	POINT,SHOW	Agent, Patient, Recipient, (Instrument)
action	MOVE,FORCE	Agent, Patient, Source, Goal, (Instrument)
action	POSITION	Agent, Patient, Ground, (Instrument)
spatial	IN,ON,OFF, OUT,AT,NEAR	Figure, Ground
spatial	MATCH,MISMATCH	Figure, Ground

Table 4.2: Coded relations. Parentheses denote optionality.

the two participants, MOTHER and CHILD.[5] The feature COLOR={RED, GREEN, YELLOW, BLUE} was coded for the blocks and the feature SHAPE={SQUARE, ROUND, TRIANGULAR, STAR} for blocks and holes. The relations and their roles can be found in table 4.2.

For every three-second interval of video, all coder-observed relations, the objects partaking in these relations, and their properties were coded using ELAN (Brugman & Russel 2004). The actions (first four rows of Table 4.2) denote simple manual behavior, which we assume children can recognize (Baillargeon & Wang 2002). The spatial relations reflect basic categories of containment and support (IN,ON) and their negation (OUT,OFF), as well as two relations denoting non-containment and non-support contact (AT) and nearness (NEAR). Understanding basic spatial relations precedes the onset of meaning acquisition and can thus be assumed to be in place (Needham & Baillargeon 1993, Hespos & Baillargeon 2001), although many specifics may be language-specific (Choi 2006).[6] The MATCH or MISMATCH with a hole was furthermore inferred from these relations. Spatial relations were deemed salient if a change in the relation occurred (e.g., if a BLOCK was the Figure of an IN-relation in the current interval, when it was not in the previous interval).

The coding procedure was evaluated for inter- and intracoder agreement (Carletta 1996) on a subset of the data: both coders coded three randomly selected dyads twice. All relations were coded reliably both within and between coders (Cohen's $\kappa > 0.8$), except POSITION (intercoder: $\kappa = 0.51$, intracoder: $\kappa = 0.47$). Closer inspection showed that there was some leakage from POSITION to MOVEMENT, which follows from the fact that the two predicates are

[5]In many cases, the complete description of a referent is a single feature. In those cases, only the single feature is given. If multiple features constitute the description of a referent, this is marked with curly brackets around the set of features making up a referent.

[6]Ideally, one would encode the range of construals of a situation, including 'tightness-of-fit'. As a first attempt at relational coding of situations, we opted for convenient, yet widely known notions like 'containment' and 'support'.

time	type	coding/transcription
0m0s	**situation**	<nothing happens>
	utterance	een. nou jij een.
	translation	one. now you one. "One. Now you try one."
0m3s	**situation**	position(mother, toy, on(toy, floor)) grab(child, b-ye-tr) move(child, b-ye-tr, on(b-ye-tr, floor), near(b-ye-tr, ho-ro)), mismatch(b-ye-tr, ho-ro)
	utterance	nee daar.
	translation	no there. "No, there."
0m6s	**situation**	point(mother, ho-tr, child) position(child, b-ye-tr, near(b-ye-tr, ho-ro)) mismatch(b-ye-tr, ho-ro)
	utterance	nee lieverd hier past ie niet.
	translation	no sweetie here fits he not. "No sweetie, it won't fit in here."
0m9s	**situation**:	point(mother, ho-tr, child) letgo(mother, lid) grab(mother, b-ye-tr) move(mother, b-ye-tr, near(b-ye-tr, ho-ro), near(b-ye-tr, ho-tr)) match(b-ye-tr, ho-tr) letgo(child, b-ye-tr) grab(child, b-bl-st) move(ch,b-bl-st,on(floor),in(air))
	utterance:	hier in. kijk e(en)s. een twee.
	translation:	here in. look once. one two. "In here. Look. One two."

Table 4.3: A sample of the dataset. The dash-separated abbreviations denote blocks and holes and their properties, where for blocks the order is **b**-{**re**d,**gr**een,**bl**ue,**ye**llow}-{**ro**und,**st**ar,**sq**uare,**tr**iangular}, and for holes **ho**-{**ro**und,**st**ar,**sq**uare,**tr**iangular}.

poles on the same scale (POSITION being motion in place, MOVE being motion from one place to another), and the demarcation point is in practice rather vague. When the coders disagreed, I decided the annotation. A sample of the resulting data is given in Table 4.3.

Evaluation

Using these data, we can get closer to an answer to the question what the environment is in which a learner acquires language. To do so, we first need to determine what features form the set of candidate meanings at the time of every utterance. As discussed earlier, we can do so at several levels of descrip-

RELATIONAL WORDS			NON-RELATIONAL WORDS		
verbs (V)			**nouns (N)**		
word	translation	meaning	word	translation	meaning
draaien	turn	POSITION	blok	block	BLOCK
duwen	push	FORCE	deksel	lid	LID
geven	give	GRAB,SHOW,LETGO	ding	thing	TOY | BLOCK
gooien	throw	FORCE | HIT | LETGO | MOVE	doos	box	BUCKET
halen	get	MOVE | POSITION,OFF | OUT	emmer	bucket	BUCKET
horen	belong	MATCH | MISMATCH	gat	hole	HOLE
kiepen	tilt	POSITION	grond	ground	FLOOR
pakken	grab	GRAB	insteekpuzzel	plug puzzle	TOY
passen	fit	MATCH | MISMATCH	pot	pot	BUCKET
schroeven	screw	POSITION	puzzel	puzzle	TOY
stoppen	put in	IN,MOVE	puzzelstuk	puzzle piece	BLOCK
zetten	put on	MOVE | POSITION,ON	spel	toy	TOY
			stuk	piece	BLOCK
			tafel	table	TABLE
			trommel	tin	BUCKET
prepositions/adverbs of space (P)			**adjectives (A)**		
word	translation	meaning	word	translation	meaning
af	off	OFF	rood	red	RED
in	in	IN	geel	yellow	YELLOW
op	on	ON	ster	star	STAR
uit	out	OUT	groen	green	GREEN
open	open	BUCKET,LID,OFF	vierkant	square	SQUARE
dicht	closed	BUCKET,LID,ON	rond	round	ROUND
			driehoek	triangle	TRIANGULAR
			blauw	blue	BLUE

Table 4.4: The lexicon of target words. Pipes denote that either of these features can apply.

tion. First, we can wonder what conceptual features are available (*conceptual noise/uncertainty*). Second, we can look at the availability of referents (entities and events) of linguistic items in the utterance (*referential noise/uncertainty*). Finally, we can look at the availability of entire situations to which utterances refer (*propositional noise/uncertainty*).

For this research, we focus on just the former two levels and collapse the distinction between conceptual and referential noise and uncertainty: as many events and objects were not coded with complex feature sets as representations, the single conceptual feature is identical to a description of the referent class. For some cases, however, words are intended to refer to an event that has to be described as a set of features. The verb *zetten*, for instance, means 'to put/position something on/onto something', so both POSITION and MOVE can be part of the valid referent of this verb, in addition to the presence of an ON feature. Other words refer to conceptual features of entities that do not constitute the complete description of the referent itself: *vierkant* 'square', means that the object is square-shaped, but the label can be applied to entities of different categories: both blocks and holes can be square-shaped.

We assume that for the list of content words in table 4.4, the correct meaning is the set of features given with it. Features that are separated with pipes mean that one of these features is part of the correct meaning of that word. We call this list the golden lexicon. Given this golden lexicon, we can investigate how much uncertainty and noise the learner would experience in acquiring that word. That is: we start from the words rather than from the sets of concepts (as we did in the initial definitions of $noise$ and $uncertainty$ in section 4.1). Let us for now assume that the set of candidate meanings consists of the set of features in the situation within the three-second interval in which the utterance was starting to be produced, thus leaving out any hierarchy or grouping in the annotation. Let us call the candidate meanings the situational context S, the utterance U, consisting of words w, and the set of meaning features to be associated with a word $Meaning(w)$ (which would, for a set of words constituting an utterance, be the set of communicated meanings of the utterance).

$$Noise(w) = 1 - \frac{\sum\limits_{f \in Meaning(w)} \frac{|U, S_{U,S:w \in U \wedge f \in S}|}{|U, S_{U,S:w \in U}|}}{|Meaning(w)|} \tag{4.3}$$

$$Uncertainty(w) = 1 - \frac{\sum\limits_{U,S:w \in U} |S \cup Meaning(w)|}{|U, S_{U,S:w \in U}|} \tag{4.4}$$

$noise$ is the proportion of Utterance-Situation pairs in which the manually assigned feature of the word was lacking (averaged over all features in $Meaning(w)$, in the case of multiple features). $uncertainty$, then, is the average number of features in the situation of the Utterance-Situation pair in

which a word occurs, that are not referred to by the word. In this operationalization, we do not formalize $uncertainty$ as a proportion, but rather give the average number of other situations. Again, the pipe-separated features applied when either of them was present (so that the $noise$ will not become higher when just one of them is present).

We calculate the levels of noise and uncertainty per word in the golden lexicon, but also per part-of-speech class. For these latter calculations, we take the average over the words contained in that class, weighted by the frequency of that word. These aggregate figures give us an insight in how noise and uncertainty values may differ between semantic/grammatical classes.

Noise in the input data

Figures 4.1 and 4.2 give the $noise$ scores per word in table 4.4 and per part-of-speech category respectively. We can see that the noise varies between 0.0 (everytime the word is uttered, the meaning is present in the set of candidate meanings) to 1.0 (the meaning is always absent when the word is uttered). For only 5 out of 41 words in the golden lexicon, the features to which the word refers are always found in the situational context accompanying that utterance. For another 21 out of the 41 words, the noise is lower than or equal to 50%.

Interestingly, when we look per part-of-speech category (figure 4.2), the category of adjectives (i.c., color and shape terms) has a substantially lower average $noise$ than the other categories. Furthermore remarkable is the lower average noise for verbs than for nouns and prepositions, meaning that verbal meanings (for the items listed in table 4.4) are less frequently absent from the immediate situation than the meanings of nouns and prepositions. The high values for nouns are striking; this is the class of words typically thought to be learnable by ostension, but the object referred to is not being manipulated in the immediate situational context in over 50% of all cases.

Uncertainty in the input data

Figures 4.3 and 4.4 give the $uncertainty$ scores per word in table 4.4 and per part-of-speech category respectively. For the uncertainty, we see far less variance between the words and different parts-of-speech: the majority of words seem to have an $uncertainty$ between 8 and 12. This does not come as a surprise: we can expect the amount of other events happening and object being present to remain approximately the same across different categories. In other words: most of the time, about the same amount of candidate meanings can be expected to be present.

Nevertheless, it is good to obtain this kind of information, because it provides us with insight in the amount of uncertainty per word, and shows how most simulation-based models actually do approximate realistic values for referential uncertainty. In Fazly et al.'s (2010) approach, for every sentence, an-

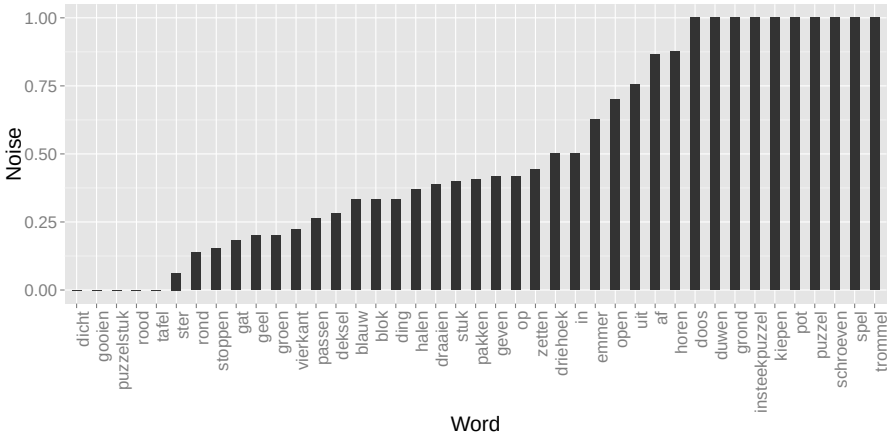

Figure 4.1: Noise per word.

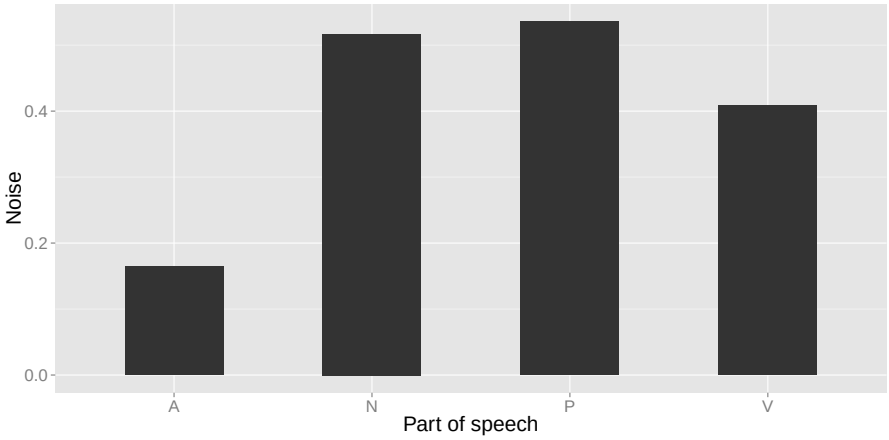

Figure 4.2: Noise averaged over the four part-of-speech categories.

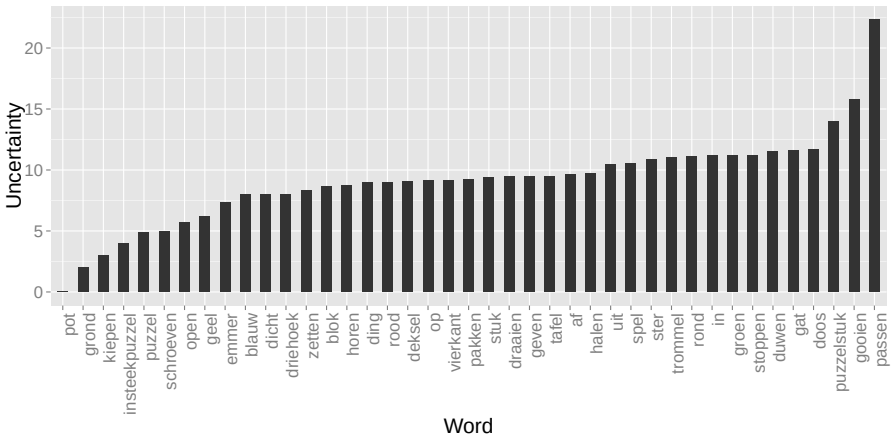

Figure 4.3: Referential uncertainty per word.

other sentence's situation is added to the current sentence as uncertainty: suppose we have sentences of five words, we will also have simulated situations of ten semantic features, which contains about the same amount of referential uncertainty as the empirical data discussed here, with for every feature 9 non-target meanings being present.

4.2.3 Noise-reduction through understanding intentionality

The values for $noise$ and $uncertainty$ obtained in the previous section have to be interpreted in the light of the assumption that the learner is only attending to the interval of three seconds in which the utterance was produced. This attentional scope is artificially narrow. However, if we want to make it wider, we need a principled way of doing so. In this section, we work out a principled extension of the attentional scope.

From behavioral experiments on word learning, we know that learners go well beyond the spatiotemporally contiguous situational context in creating a set of candidate meanings (Tomasello 1995, Sabbagh & Baldwin 2005). What these experiments show, on a conceptual level, is that the child uses other sources than the immediate environment to form the set of candidate meanings. Most of these sources require complex mental models: understanding that a word label applies to the object some person is looking for, but cannot find, requires the child to engage in a rather complex line of reasoning. Implementing these socio-cognitive mechanisms as computational models (or parts of symbol-learning models) would be an interesting research avenue, but for

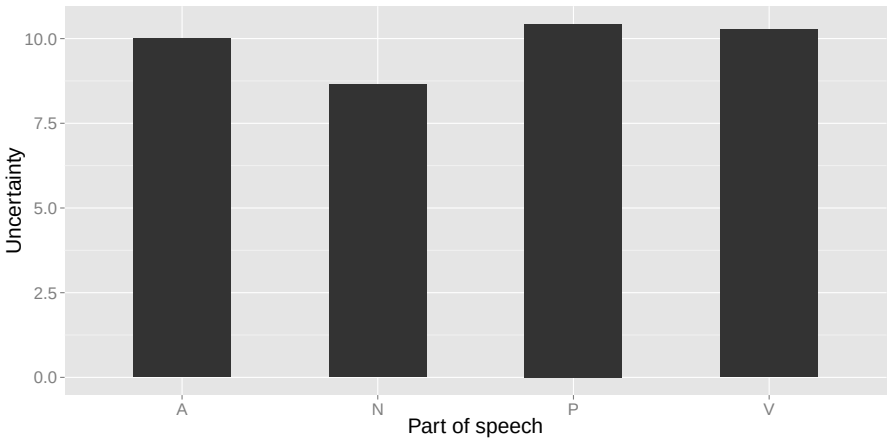

Figure 4.4: Referential uncertainty averaged over the four part-of-speech categories.

the current purposes we take a simpler approach.

Here we follow Cross's (1977) approach, viz. to take the situation between the previous and the subsequent utterance to constitute the attentional scope of the learner. This constraint can be motivated on socio-cognitive grounds. Tomasello (1995) showed how children acquire verb meaning more readily when the event follows the utterance than when it precedes the utterance, and preceding situations in turn allow children to learn the verb's meaning better than ongoing situations.

We extend Tomasello's (1995) insight to other categories as well, by generalizing that the child will attend to all *situations* in the context in close temporal proximity to the utterance. Once the child knows that the signal the caregiver is emitting is meaningful, that is, is intended to refer to something, the child can assume that some utterance U probably refers to something happening after the previous signal, and before the next one was emitted. After all, if another utterance U' intervenes at some time between the time of some situation S and the time of U, it is more likely that U' rather than U refers to S. Otherwise, the speaker would not have emitted a novel signal.

Operationalization

For every utterance U at time t, all situations are included in the set of candidate meanings that fall in the inclusive interval between the highest t' lower than t for which there is an utterance specified on the one hand, and the lowest t'' higher than t for which there is an utterance specified on the other. We

t	utterance	situational features	candidate meanings
1	*you grab ball!*	{}	{CHILD, GRAB, LETGO, DOLL}
2		{}	
3		{CHILD, GRAB, DOLL}	
4	*where's the ball?*	{CHILD, LETGO, DOLL}	{CHILD, GRAB, LETGO, DOLL, BALL, MOTHER, POINT, COOKIE}
5		{CHILD, GRAB, BALL}	
6	*good girl!*	{MOTHER, POINT, COOKIE}	{MOTHER, POINT, COOKIE, CHILD, GRAB, BALL, LETGO, DOLL}
	now this one.	{MOTHER, POINT, COOKIE}	{MOTHER, POINT, COOKIE, CHILD, GRAB}
7		{}	
8		{CHILD, GRAB, COOKIE}	

Table 4.5: A toy example of how the wide set of candidate meanings is formed.

use the same golden lexicon and evaluation metrics as in the previous section. Again, we describe the noise and uncertainty observed in these wider candidate meaning sets, and compare them with the noise and uncertainty observed in the narrower candidate meaning set, where the candidate meanings include only the features observed in the interval in which the utterance was starting to be produced.

Table 4.5 gives a toy example of the way the wide set of candidate meanings is constructed. For the utterance at $t = 1$, all features up to and including those at $t = 4$ (when the next utterance is produced) are included. Similarly, the utterance at $t = 4$ includes all features between $t = 1$ and $t = 6$ inclusive. At $t = 6$, two utterances are produced. The wider scope for the first thus is limited to the features in the interval $t = [4, 6]$, as at $t = 6$ the next utterance is already produced. For the second utterance, the interval for the candidate meaning is $t = [6, 8]$, because the previous utterance is produced at $t = 6$ and $t = 8$ is the endpoint of the fragment.

Noise given a wider attentional scope

As we can see in figure 4.7, the referential noise is lower for most words. This is a logical necessity: as the narrow set of candidate meanings is a subset of the wide set, anything present in the former is also present in the latter. For 13 out of the 41 words in the golden lexicon, about one third, there is no noise in the wide situational context, and for another 12 out of 41, the noise is lower than or equal to 25%. So, whatever the level of uncertainty, the features referred to by the words in the golden lexicon are often present in the situational context.

Interesting differences can be found between the different parts of speech. For three out of the four categories, viz. adjectives, prepositions and verbs, the noise is reduced on average with more than 50%, yielding noise levels for

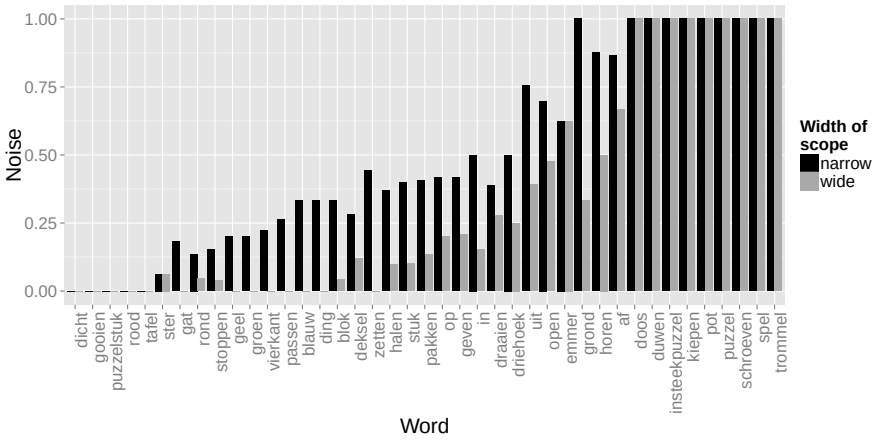

Figure 4.5: Noise per word, for both the narrow and wide set of candidate meanings.

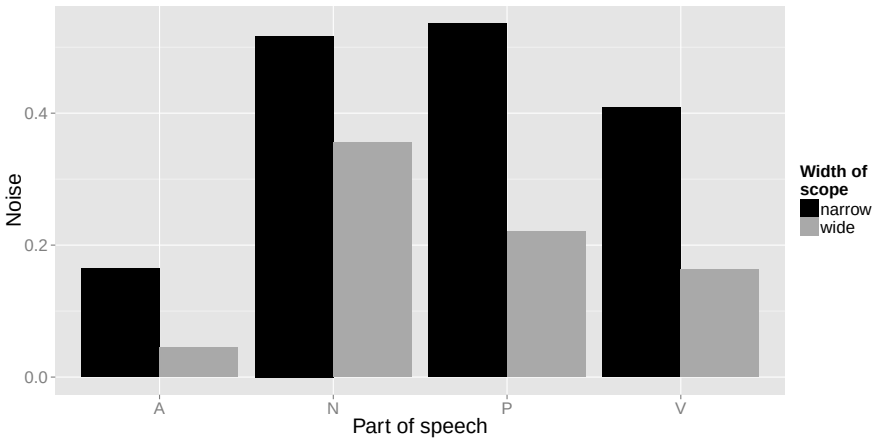

Figure 4.6: Noise averaged over the four part-of-speech categories, for both the narrow and wide set of candidate meanings.

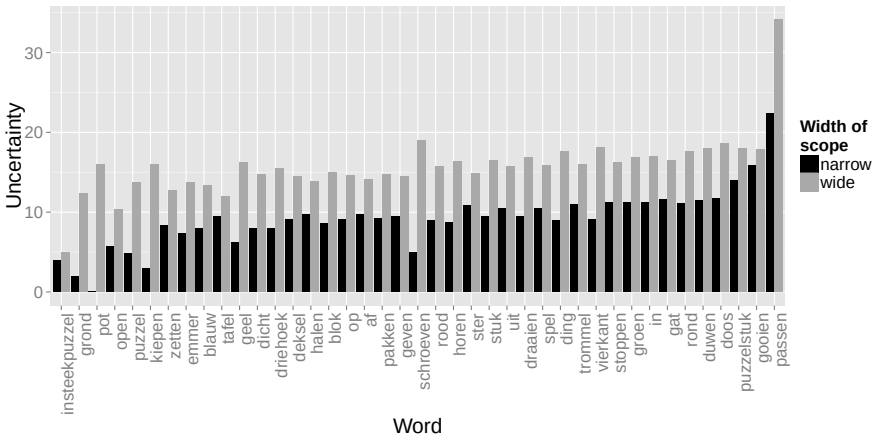

Figure 4.7: Uncertainty per word, for both the narrow and wide set of candidate meanings.

verbs and prepositions of around 20%. Nouns remain a category for which much noise is present: in about 33% of all cases, the object referred to by the noun is absent from the wide-scope set of candidate meanings. The low levels of noise for verbs and prepositions suggest that the absence of situational information may not be as problematic as Gleitman (1990) suggests, if we assign the language-learning child a slightly wider, but nonetheless temporally restricted scope of attention. The high levels of noise for nouns remain puzzling, as it is often thought that this category has a salience bias because of temporal stability (cf. Gentner & Boroditsky 2001) and can be learned through ostension. One caveat is that what are called adjectives in this model, are in fact most often expressions referring to objects (*de rooie*, 'the red (one)', *die vierkante* 'that square (one)'), so that the noise for all expressions referring to objects (either by using their class label, or some salient property), is not as high as that for nouns.

Uncertainty given a wider attentional scope

Increasing the scope of attention for the learner also logically increases the amount of uncertainty: if the narrow-scope set is a subset of the wide-scope one, all features present in the former are also present in the latter. The wide-scope set furthermore contains all features found within the narrow scope, so this set is always larger. As is shown in figures 4.7 and 4.8, most words now have somewhere between 12 and 18 non-target features present in the set of candidate meanings, again with little difference between the different parts of

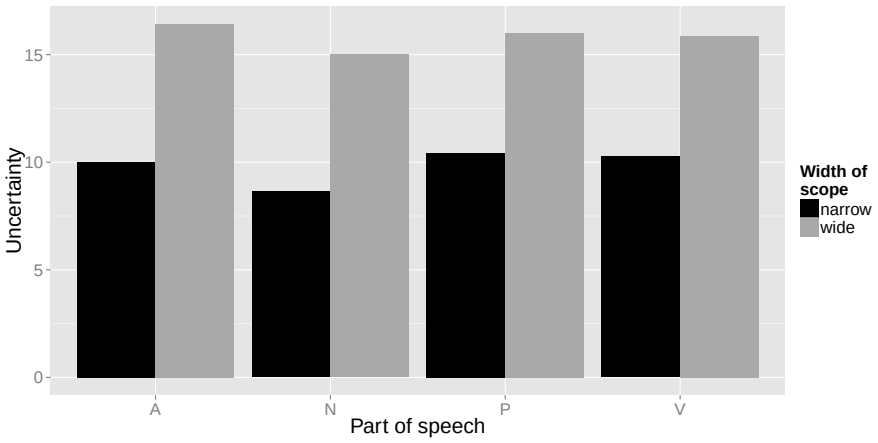

Figure 4.8: Uncertainty averaged over the four part-of-speech categories, for both the narrow and wide set of candidate meanings.

speech.

4.2.4 Interpretation and implications

What do these descriptive statistics imply for computational modeling? Firstly, the noise levels found in the annotated video data are higher than any of the authors suggest, even when applying a simple, motivated extension of the temporal width of the attentional scope of the learner. Nevertheless, the values, given the wider scope, are not much higher than with the methods of Siskind (1996), Fazly et al. (2010), and Alishahi & Stevenson (2010). What we do find, is a difference between parts of speech, with nouns displaying the most noise, followed by prepositions and other spatial relations, followed by verbs, and with adjectives displaying the least amount of noise.

Concerning uncertainty, we did not find any striking differences between the word classes. Given the narrow attentional scope, between 8 and 12 non-target features were present for every word, whereas given the wide scope, this figure rose to somewhere between 12 and 18. These numbers are hard to compare directly to the uncertainty parameters used by Siskind (1996) and Fazly et al. (2010), but show that their choice to use a relatively high amount of uncertainty is warranted.

Importantly, all of these results cannot be generalized without several caveats. First of all, the amount of noise and uncertainty depends upon the coding schema for the semantic features and the choice of features in the

golden lexicon. One can criticize these, and it is likely that the methods I used here can be improved. However, in formulating a method for measuring the noise and uncertainty, this research is among the first (together with, for instance, Matusevych, Alishahi & Vogt (2013); see section 4.2.5) to assess the level of noise and uncertainty in realistic situations of caregiver-child interaction.

Secondly, the setting in which this interaction is found is relatively narrow. We looked only at situations in which the caregiver and the child were playing a game of putting blocks in wholes. The setting of a game greatly influences the discourse, and other situational contexts may show different noise and uncertainty ratings.

Finally, the values may apply only to Dutch caregivers interacting with their children. Possible effects of cultural background are not included. Is it only the amount of verbal interaction that varies, or do we also find differences in how the utterances relate to the set of candidate meanings? I do not expect there to be any reason for the latter claim, but as long as this has not been investigated, it remains an assumption.

As for the simulation method, the amount of noise we incorporate has to be somewhat higher than the figures reported in table 4.1. With a stronger focus on uncertainty, I believe the problem of noise has been understudied and thus underestimated. Furthermore, a simulation method would have to approximate the noise-parameter differently for the different word classes. Although the sample I used is rather small and non-varied, we can assume the values for the different part-of-speech classes to hold until we have better information.

4.2.5 The issue of situational interdependence

Situational interdependence in earlier research

So far we have been making the assumption that the set of candidate meanings is an unordered set. However, the concepts can be structured into events, relations, their participants and their properties. This is information that can both be beneficial and detrimental to the learner. As Siskind (1996) notes, when a model recognizes that several parts of the utterance map to several parts of one situation out of the many possible ones, it can narrow down the space of candidate meanings for the non-mapped words of the utterance, because it can infer that these refer to (non-mapped) parts of that situation. On the other hand: events do not occur independently from each other (as noted by Siskind (1996) as well), so several different events and their participants may be highly similar to each other, which makes the task of identifying the correct one harder.

All models allowing for referential uncertainty incorporate this insight into their procedures for generating non-target elements in the conceptual space. Fazly et al. (2010) include the semantic representation of the previous utterance in the set of candidate meanings. The motivation for this procedure is

that contiguous utterances probably express related meanings (as the topics of discourse will more often stay the same than shift drastically), and that by adding these meanings, we have more realistic uncertainty than if we added the semantic representation of a random sentence.

Siskind (1996) does not use corpora of child-directed speech to simulate semantic representations and hence uses generation methods to obtain these representations. In his generation procedure, he acknowledges and addresses this issue of situational non-independence. His solution is to split up the space of candidate events (thus: the candidate meanings, as structured into events, represented as predicate-argument structures) into a number of clusters, each of some size k (in Siskind's case, $k = 5$). Within each cluster, the different situations are similar to each other. For each cluster, one event is first generated at random, after which it is copied to form the cluster $k - 1$ times, where in the copying elements of the event can be replaced with some probability, which he sets at 0.25. This results in the candidate meanings consisting of a number of internally similar clusters of events.

Siskind's method seems a good way to generate realistic uncertainty, capturing, among other things, Gentner's (1978) concern that there are many ways to conceptualize the same event or partition it into different sub-events (where in his method the different conceptualizations or partitions would form the different members of a cluster). However, we can again estimate the probability of a similar event happening on the basis of the annotated video data.

The inquiry into the dependence of situations on each other was pioneered by Matusevych et al. (2013), starting from similar concerns as the ones raised in this chapter, viz. providing more realistic simulated data to evaluate computational models of symbol acquisition on. Matusevych et al. (2013) used hand-coded video data of caregiver-child interaction in order to measure the overlap between different situations. Aspects of the situation were coded as atomic features, and every situation at some time consists of a set of such features. They then calculated the overlap between two subsequent situations by dividing the intersection of the two sets of features by the union of those sets:

$$Overlap(S_{t-1}, S_t) = \frac{|S_{t-1} \cap S_t|}{|S_{t-1} \cup S_t|} \tag{4.5}$$

Matusevych et al. (2013) measured the overlap between situations in natural interaction under two conditions. In the 'all' condition, all objects and situations that were present in the visual field were part of the situation, whereas in the 'active' condition, only the objects manipulated in actions performed by the caregiver or child, as well as those actions themselves, were part of the situation. Using the Overlap measure, they showed that the overlap between situations observed in natural interaction is significantly higher (0.436 for the 'active' condition, 0.912 for the 'all' condition) than when the situations are generated on the basis of the utterances (0.112 using Fazly et al.'s (2010) method).

feature type	features
objects	CHILD, MOTHER, TABLE, LID, BUCKET, HOLE, HANDLE, FLOOR, AIR, HAND OF CHILD, COOKIE, BLOCK
properties	RED, YELLOW, BLUE, GREEN, SQUARE, CIRCULAR, TRIANGULAR, STAR-SHAPED
relations	IN, ON, AT, NEAR, OFF, OUT, MATCH, MISMATCH
actions	REACH, GRAB, POINT, LET GO, HIT, FORCE, POSITION, MOVE, SHOW

Table 4.6: Feature types.

Obtaining continuation probabilities

Operationalization Apart from obtaining more general insight in the situational stability using Matusevych et al.'s (2013) $Overlap$ measure, we would also like to measure whether certain aspects of the situation are more stable over time. To do so, we can calculate the probability of a feature being present in the next situation given its presence in the previous situation. We call this measure the continuation probability, and we can calculate this per semantic feature. The continuation probability of a semantic feature thus is given as follows:

$$Continuation(f) = \frac{|S_{f \in S_t \wedge f \in S_{t+1}}|}{|S_{f \in S_t}|} \tag{4.6}$$

In other words: the continuation probability of a feature is given by the cardinality of the set of situations in which f occurs, as well as in the subsequent situation, divided by the cardinality of the set of situations for which f occurs.

We gain further insight in the continuation of certain types of features by grouping them according to the kinds of meanings they constitute. Table 4.6 presents the grouping into four categories: objects, properties, static relations and actions.

Results Matusevych et al.'s (2013) Overlap measure gives us a value of 0.429. This value is very close to the 0.436 reported for the 'active' condition in their study, which is the most similar to the coding method used with this data. The continuation probabilities per feature, for all features occurring more than 20 times in the data, are given in figure 4.9. We can see that there is quite some variation in the probability of a feature being found in the next situation, with the primary agents and patients of the situations (the mother, child and blocks) constituting the features for which it is most likely that they will be found

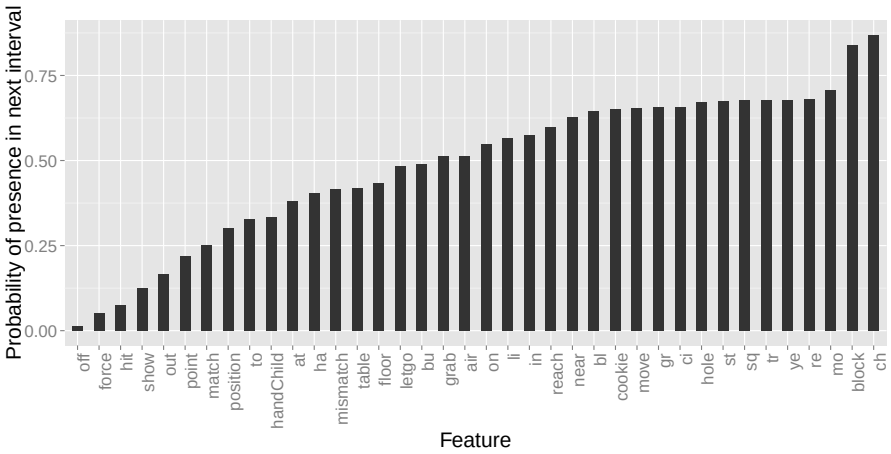

Figure 4.9: Probability of a feature being present in the next interval given presence in the current interval.

in the subsequent situation as well. When we look at the values for the semantic types (figure 4.10), we observe that objects (0.679) and their properties (0.661) have a higher probability of being found in the subsequent situation than actions (0.515) and static relations (0.493). For the last category, it should be remarked that it was only coded when a static relation came into being, assuming the relation would only be salient when it is novel. Obviously, this is a design choice that influences the continuation probability.

4.2.6 Discussion

In section 4.2, I reported several findings concerning the informativeness of the situation in which the child is trying to create symbolic pairs. One can have many doubts regarding the exact operationalization of the concepts and the method of studying these. The main point was, regardless of these specifics, to disentangle a set of concepts that influence the way we think about the acquisition of symbolic pairs. Recall that noise was the absence of conceptual material expressed with an utterance, uncertainty the superfluence of such material with respect to what the utterance conventionally conveys, and continuation the consecutivity of conceptual material. Each provides the learner with problems, and there may not be one learning mechanism to solve them all. These finer distinctions thus provide 'tools for thinking': one looks at the problem of the acquisition of symbolic pairings differently if one has to consider all three problems and they subdivide the bigger problem of learning

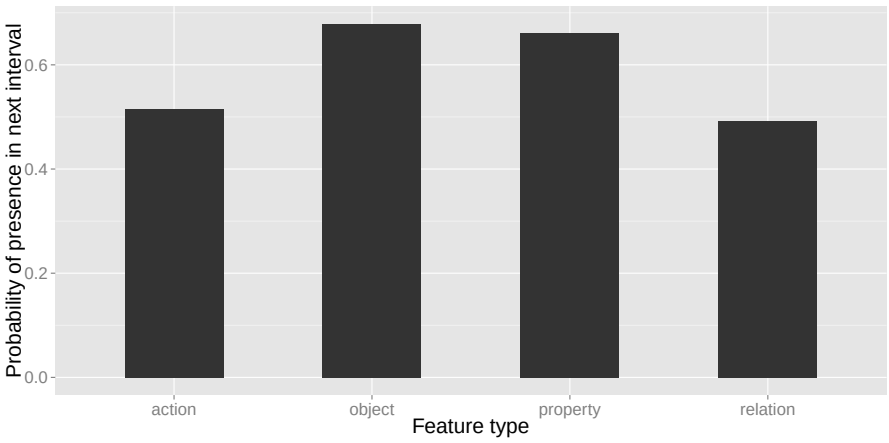

Figure 4.10: Probability of features being present in the next interval given presence in the current interval, averaged over feature types.

conventional symbolic pairings into conceptually coherent subproblems. One contribution of this chapter is to shape this conceptual toolbox.

The division into the levels of conceptual, referential and propositional noise can be seen as another step towards conceptual clarification within the domain. Here too, the ambiguity is not hurtful a priori, but the finer distinctions can help focus research on the informativeness of the situation. This distinction for instance allows us to consider the different sources underlying and mechanisms solving different kinds of noise and uncertainty: the absence of conceptual features at a sub-referential level, such as considering a ball as only being a round object, and not a toy, may point to misperception and cognitive biases towards certain regions of the conceptual space, whereas the absence of a referent or even an entire event is more likely due to simply not observing it, the former being more cognitive and the latter more perceptual. The superfluous presence of conceptual features may not be a problem at all (as long as the correct referents are identified, communication succeeds), but when too many entities and events are considered as referents, or when too many situations or propositions are considered to be expressed, we may investigate what mechanisms help the learner overcome this problem.

The reason why one would want to do an empirical exercise with such a toolbox, as I did in this chapter, is to recognize the different problems noise, uncertainty, and continuation cause and to evaluate the severity of these problems. This requires datasets such as the ones we used, and annotation programs such as ELAN. Although it requires prohibitively much effort to anno-

tate enough data manually to directly train computational models on, they do provide a source for further analysis of the concepts acquisitionists work with. In explorations such as these, we can see how technological and methodological innovation may direct further theoretical development.

4.3 Towards a realistic simulation procedure

For computational modeling studies, we need high quantities of training data. Because obtaining such amounts of data in the way described in this chapter is labor-intensive, the only way to proceed seems to be to use a method for artificially generating data, as the other models described have done. The properties of the data generated by his procedure have to be close to the parameter values for noise, uncertainty and continuation we have found in the empirical study presented in this chapter. In this section, such a method is presented, based on Alishahi & Stevenson's (2008) method, insights from the simulation method Matusevych et al. (2013) developed, as well as the findings of the study presented earlier in this chapter.

4.3.1 Earlier methods

Matusevych et al. (2013) investigate to what extent the noise, uncertainty and overlap (or: situation stability) values in naturally occurring caregiver-child interaction are similar to those found in methods where the features of the situation are based on the utterance, as in Fazly et al. (2010), and Alishahi & Stevenson (2010). Motivated by the big differences found on all three parameters, Matusevych et al. (2013) developed a simulation method for generating situation-utterance pairs whose noise, uncertainty and overlap is highly similar to the observed values.

The method Matusevych et al. (2013) propose generates situations, with actions and objects, as well as utterances, on the basis of the utterance and situation generated in the prior turn. The probabilities of the situation and the utterance at some time t thus depend (among other things) on the utterance and situation at $t - 1$. The data generated by this procedure have noise, uncertainty and overlap parameter values similar to the ones observed in the 'active' condition (see section 4.3.2 for a description of the conditions).

It is the insight of generating chains of events that we adopt from Matusevych et al. (2013). For the purposes of training a model of symbol acquisition that includes meaningful grammatical constructions, we need a semantic representation that goes beyond flat sets of features, as hierarchically structured grammatical representations correspond to hierarchically structured semantic representations. One generation framework that does so, is that of Alishahi & Stevenson (2010). The method described there generates utterances on the basis of the frequencies of a set of verbs, their argument structures, as well as their arguments in three subcorpora of child directed speech (the three chil-

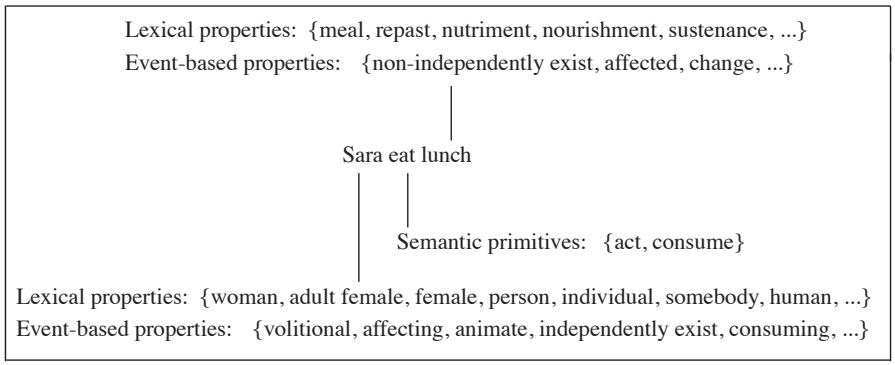

Figure 4.11: Semantic features extracted on the basis of the utterance in Alishahi & Stevenson (2010, 59).

dren in the Brown corpus; Brown (1973)). Only the intersection of the thirteen most frequent verbs in the child-directed speech in each subcorpus of the Brown corpus was used (i.e., the thirteen verbs *go*, *put*, *get*, *make*, *look*, *take*, *play*, *come*, *eat*, *fall*, *sit*, *see*, *give*). The frequencies of the argument structures was estimated by manually inspecting 100 instances of each verb, as were the frequencies of the arguments (nouns and pronouns) in these argument structures.

The verbs, arguments and prepositions marking several valency relations, as well as the valency relations themselves, are then used to determine the meaning of the utterance. To do so, several resources are used (Jackendoff's (1990) event features, Dowty's (1991) proto-roles, as well as event-specific roles such as 'eater' and 'moved entity', and WordNet hyperonym chains for objects (**?**)). Figure 4.11 gives an example of the sets of semantic features extracted on the basis of the utterance *Sarah eats lunch*.

Note that in this procedure the linguistic realization of arguments is not by necessity isomorphic to the conceptual argument structure of the event: it may be that the event has two participants, but only one is expressed linguistically as an argument of the verb. This is an important property of the input items, which we described as referential uncertainty, as linguistic descriptions of situations often leave out participants.

Alishahi & Stevenson's (2010) method includes a post-hoc procedure for adding noise to the data, viz. by removing or replacing features. Adding uncertainty and specifying amount of overlap is not something that can be done yet with the generation framework of Alishahi & Stevenson (2010). However, extending it to allow for the generation of a set of situations, with the appropriate amount of overlap, to be paired with an utterance, is a relatively small

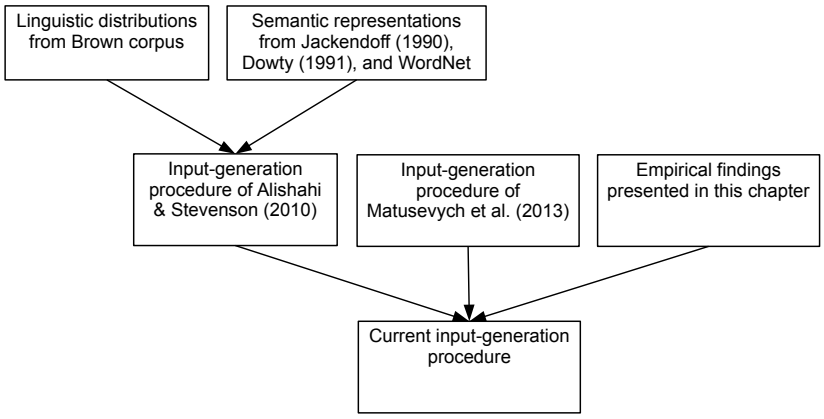

Figure 4.12: An overview of the components of the current input generation procedure.

extension.

4.3.2 Operationalization of the input generation procedure

Where do the input items for the model come from? It is not easy to just provide the learning model with a single source of input data; each method discussed in this chapter has pros and cons and the best option at this point seems to combine the best features of each. Figure 4.12 summarizes the components and main sources of inspiration for the procedure to be presented below.

Essentially, I extend Alishahi & Stevenson's (2010) procedure. This procedure generates pairings of a situational context and an utterance on the basis of a semantic ontology as well as the distribution of linguistic items in child-directed speech. As such, it provides us with utterances that are linguistically realistic in their distributional properties, and situational contexts or conceptual representations that are (arguably) cognitively realistic in their content (especially Jackendoff (1990) and Dowty (1991) claim so). The conceptual representations are, however, not realistic in their distribution, as the model operates under no uncertainty and as subsequent input items are generated independently of each other.

In order to resolve this, I extend Matusevych et al.'s (2013) line of reasoning: we generate input items as chains, where every subsequent input item is probabilistically constrained by the previous input item. Every input item

furthermore contains not just one, but a range of situations from which the learner then has to choose.

How does this work? As mentioned, the child often finds herself in the situation where multiple situations are likely candidates to be referred to, and we use the $uncertainty = [0, \infty]$ parameter to regulate the number of additional non-target situations in the input item. The $noise = [0, 1]$ parameter, on the other hand, regulates the probability of the absence of the target situation in the input item. In this procedure, I only operationalize $noise$ and $uncertainty$ at a propositional level, for convenience's sake. Future extensions of the procedure may involve operationalizing both parameters at the level of referents or features.

Recall that we defined the input of the model to consist of pairings of an utterance U and a number of situations S. How do we arrive at sets of situations that are grounded in what we know about the situational context in which the language-learning child picks up the symbols of her language? First, we create chains of U, s pairs. As we saw in paragraph 4.2.5, subsequent situations are not independent from each other. We therefore use the notion of the continuation probability to generate every situation at time t, or s^t on the basis of the situation at $t - 1$, or s^{t-1}. We define two continuation probabilities as parameters of the model: one for the objects or semantic arguments of the situation ($P_{\text{argument_continuation}}$), and one for the semantic predicate or event node of the situation ($P_{\text{event_continuation}}$). With these probabilities, we sample a set of nodes that should be present in s^t, or $node_constraints^t$.

Figure 4.13 gives an example. From the situation at $t - 1$, each object and the event is added to the set of $node_constraints^t$ with a probability of the continuation parameters $P_{\text{argument_continuation}}$ and $P_{\text{event_continuation}}$ respectively. In this case, say that the event node and the first argument node are sampled. They are then added to the set of node constraints. Using this set, we find all possible situations that fulfill all constraints, i.e., that have both nodes in their graphical representations. If we find this set to be non-empty, we sample one situation from it at random, for example the one on the right side of figure 4.13.

It is very likely that every now and then we will run into cases where the set of situations meeting all constraints is empty. In such cases, we back off and use the set of all possible situations meeting all but one constraints. If that set is empty too, be back off further to the set of all possible situations meeting all but two constraints, and so on until we have a non-empty subset. Globally, we could say that we sample from the subset of all possible situations maximally satisfying the node constraints. Given the subset of situations of which the members maximally meet the $node_constraints$, we sample similarly to Alishahi & Stevenson (2010), that is: on the basis of the corpus frequency of the verbs, argument structures and nouns expressing the situation ($P_{\text{situation}}$ in figure 4.14).

Furthermore, as chains of events in reality do not continue forever, we start sampling without an empty $node_constraints$ with a certain probabil-

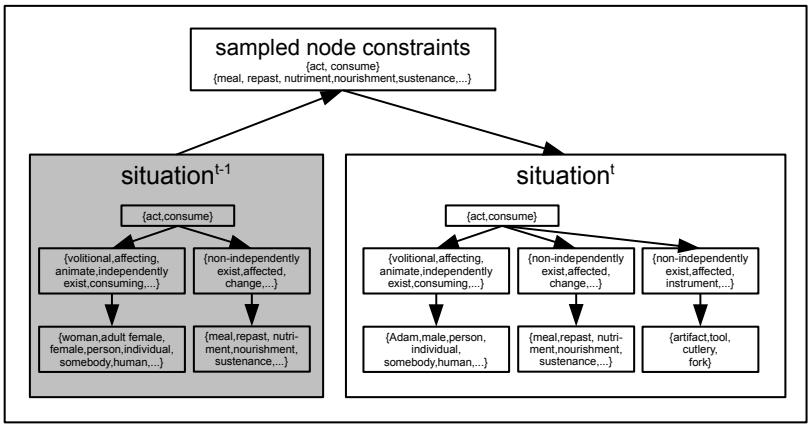

Figure 4.13: An example of sampled node constraints.

ity, called the reset probability P_{reset}. Figure 4.14 schematically represents the sampling procedure

This procedure yields a chain of U, s pairs. To get input items in the form of U, S pairs, we divide up the chain of U, s pairs into subchains. From each subchain, we then select one U, s pair to be the utterance and target situation s_{target}. All of the other situations in the subchain are then added to S. The target utterance U, as well as S constitute one input item. The division into subchains is thus the place in the generation procedure where we can parametrize uncertainty: the longer the subchain is, the more non-target situations there are in s, and the higher the uncertainty is.

We measure the uncertainty by the number of unique non-target nodes in S, similar to the way we did it in section 4.2.2. That is: given a target situation s_{target}, we take the cardinality of the set of all nodes in all non-target situations of S that are not part of s_{target}. The subchain is divided at the point where this cardinality exceeds the pre-set value for $uncertainty$, a non-negative value reflecting the maximum number of nodes in non-target situations in S. We do not differentiate between different referent types, as this would complicate the procedure too much. Figure 4.15 illustrates two chains, one with high uncertainty and one with low uncertainty.

This way of generating input data is in several ways similar to Siskind's (1996), the main difference being that his procedure selects several clusters of similar situations (see paragraph 4.2.5), whereas a subchain of situations in the proposed procedure is comparable to only one such cluster. Although several

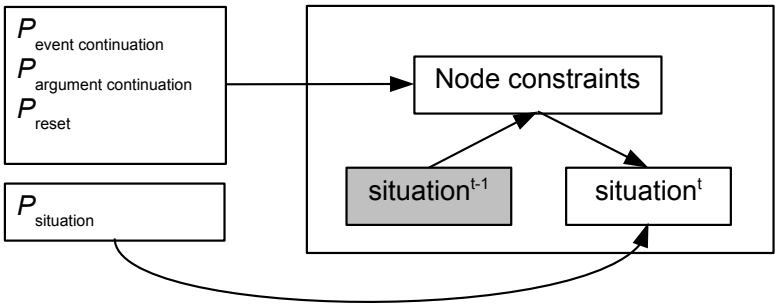

Figure 4.14: The procedure for sampling a situation.

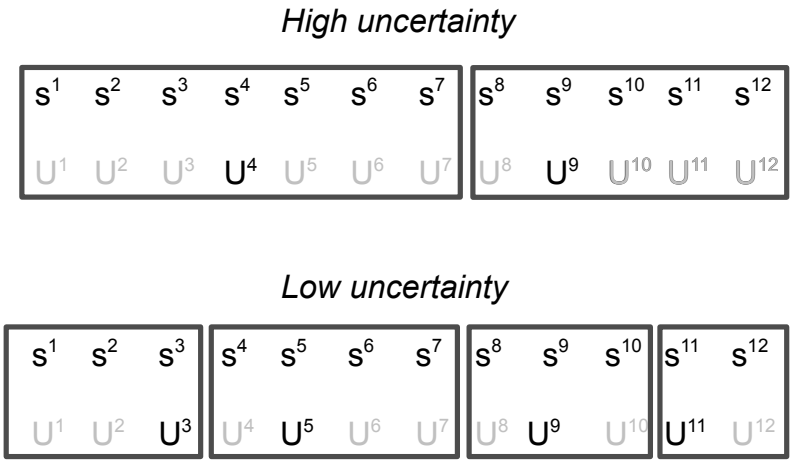

Figure 4.15: Two chains of situations, one subdivided with high uncertainty, the other with low uncertainty. 'U' denotes an utterance and 's' a situation. The grey utterances are non-selected. An input item consists of all black marked objects within one rectangle.

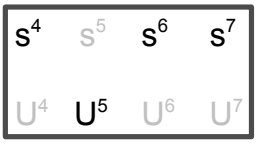

Figure 4.16: A noisy input item. 'U' denotes an utterance and 's' a situation. The situation corresponding to the selected utterance has been removed and is not part of the input item.

streams of events are likely to take place when the child is interacting with the caregiver, the child probably only attends to one such stream, namely the one that is in the joint focus of the caregiver and child.

Noise After creating an U, S pairing, we can add noise. We can do so on two levels. Similarly to Siskind (1996), we can remove the target situation from the set of situations S, so that the learner will always identify a non-target situation as the situation the speaker intended to refer to. This would constitute propositional noise. Conceptually, this means that the learner for some reason does not consider the target situation as a part of the set of candidate situations S. This may be because she did not observe it, or because she thought it to be communicatively irrelevant. The parameter that determines the amount of situations with propositional noise is called $P_{\text{propositional_noise}}$. Another approach would be to change the feature sets for some parts of the representation. This would constitute conceptual noise, and it corresponds to the situation in which the learner misperceives aspects of the situation. The parameter that controls the probability of replacing the feature set of a node in the target situation for another is called $P_{\text{conceptual_noise}}$. Figure 4.16 provides an example.

Parameter settings Using the parameters of this input generation procedure, we can generate data that fits realistic parameter settings. One major caveat is to what extent the results from the video data can be extended to apply more broadly. After all, they are derived from a very limited pragmatic setting. We can apply them directly, which would give the values in table 4.7, but in the following chapters, we will also evaluate the model presented in those chapters with other values as well, to see under what conditions the model performs well.

Note that several findings are not reflected in the parameters, e.g., the dif-

parameters	value	motivation
$P_{\text{argument_continuation}}$	0.7	$Continuation$ for objects in 4.2.5
$P_{\text{object_continuation}}$	0.5	$Continuation$ for actions and relations in 4.2.5
P_{reset}	0.05	None
$uncertainty$	15	Given the average of 15 non-target referents under the wide condition in section 4.2.3
$P_{\text{propositional_noise}}$	0.1	High estimate on the basis of the different values for referential noise in the wide scope condition (section 4.2.3)

Table 4.7: Parameters of the generation procedure and values obtained from the video data.

ference between various parts-of-speech in the parameters settings of $noise$ and $uncertainty$. This would require us to operationalize these parameters at the level of semantic referents (entities and events), which turns out to be problematic given the current definition of the model, and is therefore left for future research.

4.4 Directions for modeling symbol acquisition

The experiments on the annotated video data described in this chapter provide a very simple first approach to empirically grounding the assumptions concerning the availability of meaning independently of language. To this end, we made some simplifying assumptions. We were only concerned with features that were actively being attended to, following research on joint attention (Tomasello 2003), and we assigned hardly any socio-cognitive skills to the learner, beyond assuming that whatever situations are present between the previous utterance and the subsequent one constitute the set of candidate meanings for the current utterance.

Furthermore, we assumed that the features were independent within a situation, thereby making no difference between bundles of features occurring together (properties always being the property of an object, events always having participants). This inherent structure of the situations may provide valuable cues for the learner. We will exploit this structure in the modeling work described in the later chapters.

Finally, starting from a set of semantic primitives is problematic. Although one can argue for a universal set of features underlying the semantics of all natural languages (Jackendoff 1990), typological research shows that such a

set at least has to be very flexible to accomodate the distinctions made in different languages. Conceptualizing the space of potential meanings in terms of continuous scales rather than discrete features may prove to be a more insightful starting point (Bowerman 1993, Levinson, Meira, & the Language and Cognition Group 2003, Majid, Boster & Bowerman 2008) for describing language-specific categories. Beekhuizen, Fazly & Stevenson (2014) describe how we can use these continuous spaces to study semantic error patterns in language acquisition, showing how overgeneralizations can be predicted on the basis of continuous spaces and the insight that groupings of situations with one linguistic marker that are cross-linguistically more common, are probably also easier to acquire than groupings that are cross-linguistically less common.

One can always push realism further. I believe, however, that the current proposal at least provides more realism than input generation procedures hitherto proposed. With a computational model satisfying many constraints or desiderata imposed by usage-based theorizing and a realistic input generation procedure, we can now see how the model behaves and what kinds of representations it acquires. These issues will be addressed in the subsequent three chapters.