Constructions Emerging

A Usage-Based Model of the Acquisition of Grammar

The research reported here was supported by NWO grant 322.70.001

Published by

LOT
Trans 10
3512 JK Utrecht
The Netherlands

phone: +31 30 253 5775
e-mail: lot@uu.nl
http://www.lotschool.nl

Cover illustration: #9f4c62

ISBN: 978-94-6093-183-3
NUR: 616

Constructions Emerging

A Usage-Based Model of the Acquisition of Grammar

Proefschrift

ter verkrijging van
de graad van Doctor aan de Universiteit Leiden,
op gezag van Rector Magnificus prof. mr. C.J.J.M Stolker,
volgens besluit van het College voor Promoties
te verdedigen op dinsdag 22 september 2015
klokke 13.45 uur

door

Barend F. Beekhuizen

geboren 7 januari 1987
te 's Gravenhage, Nederland

Promotores:	Prof. Dr. Arie Verhagen
	Prof. Dr. Rens Bod (University of Amsterdam)
Promotiecommissie:	Prof. Dr. Roberta D'Alessandro
	Dr. Afra Alishahi (Tilburg University)
	Prof. Dr. Ewa Dąbrowska (Northumbria University)

There is a crack in everything
That's how the light gets in.
Leonard Cohen, "Anthem"

Little feet take small steps
Bloom (1991, 11)

Contents

Acknowledgements

Throughout the (almost) five years that I have been working on the project constituting this dissertation, I have had the pleasure of being surrounded by many great people who deserve to be acknowledged.

The first words of gratefulness definitely go out to my supervisors, Arie Verhagen and Rens Bod. With each of them packing a vast breadth and depth of knowledge, combined with their highly original and thorough ways of thinking, I could not have wished for a more nurturing home for my ideas.

Second, I would like to thank all other teachers and mentors that have fulfilled important exemplary roles in various phases of my academic development, both before and during my doctoral research. I would like to thank, in a somewhat chronological order, Gé Vaartjes, Leo van Santen, Ronny Boogaart, Cor van Bree, Ariane van Santen, Marijke van der Wal, Ton van der Wouden, Felix Ameka, Egbert Fortuin, Sandy Thompson, Pat Clancy, Jack DuBois, Stef Grondelaers, Remko Scha, Jelle Zuidema, Melissa Bowerman, Afsaneh Fazly, and Suzanne Stevenson for sharing their insights and stimulating my academic growth.

Important parts of this research have been discussed with Libby Barak, Gideon Borensztajn, Ailis Cournane, Max van Duijn, Stewart McCauley, Aida Nematzadeh, and Gareth O'Neill, whose criticisms helped further my thinking and writing and whose contributions I would like to acknowledge. The comments of the members of the committee, Afra Alishahi, Roberta D'Alessandro, and Ewa Dąbrowska, stimulated me to dot the i's and cross the t's, for which I would like to thank them.

I am very grateful for the academic environments I have spent time in over the past years. Many thanks go out to the participants of the CLC-lab and the LaCo-group at the ILLC, University of Amsterdam, colleagues at the LUCL, and the members of the SuzGroup and, wider, the CL-group, during my stay at the University of Toronto.

The research reported in chapter 4 could not have been carried out without the help of my two research assistants, Eva Rieborn and Marten van der

Meulen, who laboriously and conscientiously coded the caregiver-child interactions. Those caregiver-child interactions would not have been there at all if it weren't for Marinus van IJzendoorn and Marian Bakermans-Kranenburg of the department of Child Studies at Leiden University. Their generosity, in allowing me to use their data as well as occupying a desk in the department to digitize the material, is gratefully acknowledged. The input-generation procedure described in the same chapter is an extension of that of Afra Alishahi, who generously allowed me to use her code.

In the academic years 2013-2014 and 2014-2015, I lectured at the department of Dutch Language. I would like to thank my colleagues there, Ronny Boogaart, Gijsbert Rutten, Ariane van Santen, Tanja Simons, Arie Verhagen, and Marijke van der Wal for helping me out, as an inexperienced teacher, and providing me with an inspiring environment.

The research reported in this dissertation could not have been carried out without the generous *Promoties in de Geesteswetenschappen* grant of the NWO, whose support I would like to acknowledge.

On a personal level, there are too many to be thanked. I am grateful for my family, friends, and (former) partners for all the support, wisdom, joy, and love I have experienced. A warm thank you to all of you.

One person definitely deserves the 'last-but-not-least': my dear friend Folgert Karsdorp. During times of academic and personal misery his support has made the difference. Folgert's contributions at various stages of my dissertation research have been numerous and our conversations shaped chapters 5-7 to a large extent.

All remaining imperfection is solely my own.

CHAPTER 1

Introduction

When I utter the sentence *John kissed Mary*, anyone having a sufficient command of English will understand that I make an assertion, namely that an event took place in which some person whom we both know, named *John*, engaged in the act of kissing another person, again known to both of us, named *Mary*. In comprehension, language users connect an observable signal, in this case the string of sounds produced when uttering *John kissed Mary*, to an unobservable conceptualization of the situation and the speaker's communicative intent. In production, the reverse process takes place: given a conceptualization of a situation and a communicative intention, the speaker tries to figure out which observable signals to use in order for the hearer to arrive at the desired conception of the situation and communicative intent.

At the heart of linguistics is the question how observable signals, such as speech or sign, are connected to conceptualization. Various theoretical frameworks have been developed to account for the connection. Whereas there is widespread agreement between various recent frameworks concerning the question how words work (everyone harks back to de Saussure's (1916) idea of a word being a symbol, i.e., a conventional pairing of a signifying form and a signified meaning), the paths separate when it comes to the question how words are combined into larger units, such as phrases (*the red ball*, *deeply enlightened*) and whole clauses (*The red ball seems deeply enlightened*). Generative grammar, from the 1950s (Chomsky 1957) up to its various present-day incarnations (Chomsky 1993), argues that at the core of the human ability to form complex linguistic representations is a cognitive mechanism for structure-building that is autonomous, i.e., whose properties cannot be de-

rived from other cognitive domains. Construction grammar, the theoretical perspective constituting the starting point of this dissertation, provides a different perspective: the cognitive representations responsible for comprehending and producing complex utterances are **constructions** – like words, conventional pairings of an (observable) signal and a signified meaning (e.g. Goldberg 1995). Because the content and structure of these representations fully comes from other cognitive domains, a language user's grammar is not an autonomous cognitive system. Furthermore, if the representations responsible for building complex linguistic structure are pairings of an observable signal and an inferred meaning, they are qualitatively the same as regular words, and as such, according to construction grammar, all linguistic representations can be regarded as constructions.

The magnificent task faced by an infant being born into a community of speakers, is to figure out how the connections between the observable signals and the meanings work. Again, various theoretical frameworks differ in how they conceive of this task: in the Generative tradition, rooted in rationalist thought, the language learner's task is to deduce which properties from among a finite set of possibilities the grammar of her community's language has (Baker 2001). The constructivist approach, on the other hand, argues that children build up their inventory of linguistic representations in a bottom-up way, and without any preconceptions concerning grammar-wide regularities.

In the literature, one often finds a comparison of the frameworks, where empirical data is presented as being suggestive for the truth of the one and the falsehood of the other framework. To my mind, this approach is unwarranted given the state of any current theoretical framework. Because of their informal nature, any datapoint presented by an adherent of one framework can be brought into accordance with another framework or simply be dismissed as non-data (because it is not part of the 'core' of language, or because it is a 'theory-internal matter'). Of course, this possibility exists even with highly explicit and formalized theories in other fields of research, but it seems that challenges to any linguistic theory can be resolved too easily. This is not to lament the state of linguistics: independently, various theories have internally developed themselves to an enormous extent.

A more productive strategy would be to devote one's energy to the maturation of the theoretical framework by scrutinizing the theoretical construct on which it rests. This dissertation should be read as an attempt to do so. In it, I employ formalization and computer simulation as tools for achieving a deeper understanding of the theoretical framework of construction grammar. Results from the modeling work presented in this dissertation that confirm the line of reasoning of the theoretical vantage point are interesting and provide a basic sanity check of the relation between the formalized model and the informal theory. More insightful, however, are the cases where we fail to simulate a phenomenon. In those cases, it is either the model that is not faithful to the theory, or there are gaps in the theory. It is in the recognition of these gaps that theoretical progress can be made.

1.1 Early grammar

Before we turn to the more theoretical issues, let us have a brief look at what kinds of phenomena this dissertation will be occupied with. Children's early linguistic productions differ in several ways from the utterances adults produce. Their utterances are typically shorter, and markers of grammatical categories such as tense and number are mostly omitted (Brown 1973, 98-99). The main phenomenon of interest in this research, however, is the realization of argument-structure patterns over development. The event expressed by a verb has certain roles, which are linguistically realized as the arguments of that verb. Some examples of deviations from the language of adults are shown below.

(1) open drawer (Kathryn 2;0[1], opening a drawer, Bloom, Lightbown & Hood (1975))

(2) I made (Eric, 1;1 1, just reassembled a train, Bloom et al. (1975))

(3) put truck window (Adam 2;3, Brown (1973))

(4) pick up (.) puzzle up (Adam 2;6, wants to pick the puzzle up, Brown (1973))

(5) Adam fall toy (Adam 2;3, dropped a toy, Brown (1973))

(6) eat Benny now (Ben, between 1;7 and 2;6, wants to eat Sadock (1982), cited in O'Grady (1997, 61))

(7) the bridge knock down (Aran 2;4, knocked the bridge down; Manchester corpus, cited in Marcotte (2005))

Examples (1)-(4) show how elements of the argument-structure patterns that are grammatically obligatory for adult speakers of English are omitted. We find subjects and objects being left unexpressed, for instance in examples (1) and (2), but also obligatory prepositions, as in example (3). As Bloom et al. (1975) note, children often produce patterns expressing different aspects of the event they want to express in a sequential way. Example (4) is such a case: to express an event that would be expressed by an adult as *I want to pick up the puzzle*, or *I want to pick the puzzle up*, the child uses two structures, separated by a short pause, seemingly because he was not able to integrate both under a single syntactic constellation. Some basic findings with regards to the argument omissions are that more arguments are expressed over time (Tomasello 1992, 244) and that subjects more frequently omitted than objects (Bloom et al. 1975).

In examples like (5)-(7), we see cases of children's productions where not only are elements omitted, but also rules applied that are not in line with the adult usage, so-called errors of commission. In a case like example (5), the

[1] I adopt the conventional <year>;<month> notation here for the child's age.

verb *drop* would be used in the transitive frame, instead of *fall*, which cannot be used in a transitive syntactic frame. In example (6), the child's wish would be expressed with a pre-verbal subject and *want to*, as in *Benny wants to eat now*, but we find the child changing the word order of the subject and the verb. From this piece of behavior it is hard to glean what the underlying representation might have been: is the child applying a grammatical rule at all, and if so, which one? Example (7) is the mirror image of example (5): here, the noun expressing the patient role is expressed pre-verbally, in the position where subjects are typically found. Some verbs, such as *roll*, allow for the alternation whereby the semantic patient role is expressed as the syntactic subject, but *knock down* is not among them. Again the question is: which representations underlie this production?

From a developmental perspective, the occurrence of both errors of omission and commission in the same time span (roughly, Roger Brown's Stage I) is interesting, because the underlying representations and mechanisms produce both of them. As acquisitionists, we thus face the puzzle of how the learner both under- and overshoots the target. Despite decades of work on these early productions, a comprehensive account of the representations and cognitive mechanisms leading to these productions, and their development over time, has not been satisfactorily given.

1.2 Theoretical background

The constructivist theory of language acquisition, briefly introduced earlier, constitutes the starting point for our understanding of the phenomena I just outlined. Construction grammar centrally makes a representational claim: all linguistic representations are pairings of signifying elements, prototypically phonological form, and signified meaning or conceptualization. This holds for words, but also for grammatical patterns. A complementary addition to the representational claims of construction grammar is the usage-based perspective. This perpective holds that linguistic representations are qualitatively and quantitatively grounded in language use (Langacker 1988). The qualitative grounding is taken to mean that the representations consist of the cognitive reflections of usage events. Importantly, only representational content derived from the processed usage events can be part of the linguistic representations. That is to say: the language learning child has no preconception of contentive elements that are expected to be part of linguistic representations. An important consequence is that there is no place in a usage-based constructivist theory of language for universal syntactic categories, such as 'noun' and 'verb'.[2] At the core of linguistic representations should be elements of the observable signal (sound structure for spoken languages) and elements of the

[2]The rejection of universal distributional or syntactic categories follows not only from Langacker's (1988) ideas about the grounding of linguistic representation in usage. The most explicit rejection is given by Croft (2001) on the basis of language-typological arguments.

conceptualization of the world that these signaling elements are assumed to refer to.

Given this perspective, the language-acquiring child faces the task of building up an inventory of representations allowing her to communicate successfully with the other members of her community. To do so, she has several, highly general mechanisms at her disposal. Primarily, these involve mechanisms of understanding other people's intentionality, which emerge around the first birthday, and a set of abilities to recognize patterns (Tomasello 2003). The former allow the child to recognize that the speaker has communicative intentions with the speech signal he is producing and roughly what this communicative intention encompasses. The latter enable the learner to discover patterns of regularity in the signal and inferred communicative intent. These patterns are discovered, furthermore, in a bottom-up and gradual way. If we assume that the child has no preconceptions of the content of the linguistic representations, any abstraction over the processed usage events can be expected to arrive through a comparison of the structure of various usage events. Abstraction in the inventory of linguistic representation therefore only emerges after multiple comparable usage events have been observed.

1.3 Computational cognitive modeling

The past decade has seen a large number of publications on computational models of the acquisition of grammar from a usage-based constructivist perspective. We find many learning models taking semantics into account, in line with the constructivist tenet that linguistic representations consist of pairings of form and meaning throughout. This does not only include work explicitly being framed as being constructivist, such as the dissertation of Chang (2008) and Alishahi & Stevenson's (2008) clustering approach to argument-structure constructions, but also modeling research in other frameworks such as Combinatorial Categorial Grammar (e.g. Kwiatkowski 2011). Distributional learners, taking only distributional properties of the phonological form into account, have shown how multi-word units, which are assumed to play a large role in language acquisition on the usage-based perspective, are extracted (McCauley & Christiansen 2014*a*), how the integration of the constructivist view and a connectionist, sub-symbolic representation can be achieved (Borensztajn 2011), how grammatical patterns assumed to be unlearnable can be learned from usage data (Bod 2009), and how early patterns of language production can be modeled (Freudenthal, Pine & Gobet 2010). Finally, highly interesting work on the appropriate analysis of corpus data of children's early production have sparked interesting, and perhaps even productive back-and-forths between adherents of the usage-based perspective and generativist researchers (Lieven, Salomo & Tomasello 2009, Yang 2011).

The development of computational models is interesting for several reasons. First of all, in developing a computational model, the researcher is forced

to operationalize the concepts of a linguistic theory at a level of precision that allows a computer to run it as a piece of software. This means that the modeler has to make design choices in the representations and algorithms of the model that are not specified in the theory. The observed consequences of these design choices can then be 'fed back' into the theory. For this reason, it are not only the success stories that are interesting in modeling. In fact, the rhetorics of success can be misleading: when we regard modeling as an extension of theorizing, a model's failure is essentially the exclusion of a logically possible variant of the theory. Through successive failures, we can gradually delimit the range of potentially successful theoretical options. Unfortunately, modeling failures are rarely shown.

Secondly, modeling allows us to observe the interaction of various components of the model. Rather than isolating the phenomenon of interest, as is often done in experimental studies, modeling allows us to observe what the effects are if multiple components of a theory interact in processing the data (cf. Beekhuizen, Bod & Verhagen 2014). This does not mean that all models do so, or should do so, but it is a possibility of the method that I believe is worth exploring. A further consequence of this approach is that we can work towards comprehensive models, that is, models being able to fully comprehend and produce utterances.

Finally, modeling allows us an evaluation of the theory both on a wide level and a narrow level. We can model both how an operationalization of the theory behaves across the board (e.g., in understanding or producing novel utterances), but also how the operationalized theory behaves in a simulation of a particular experimental set-up or in the case of a rare event. Both lines of inquiry are important: with linguistic theory often focussing on rare events (crossing dependencies, sentences like *Pat sneezed the foam off the capuchino*), the more global behavior of the theory is often left out of consideration.

1.4 Goals of this research

It is to the background of the modeling inquiries presented in section 1.3 that the line of research reported on in this dissertation starts. At the time when I started it, there were several issues that I thought to be insufficiently addressed in existing work. Four of them constitute the central theoretical issues of this dissertation: a call for greater **comprehensiveness** of usage-based computational models, a scrutiny of the **conception of learning**, a plea for more **naturalism** when the acquisition of meaning is concerned, and the reassessment of the **starting-small** position.

As such, the research presented in this dissertation should be regarded primarily as a theoretical exercise. It involves reinterpreting, scrutinizing, synthesizing, operationalizing, and, finally, evaluating ideas from the usage-based approach to grammar. The conceptual work, to my mind, is an important part of the endeavour of computational cognitive modeling. On an epistemological

note: this does not suggest that the conceptual work has any kind of primacy over the more empirically oriented work: both operate in a heuristic loop, informing and enhancing each other.

1.4.1 Providing a comprehensive model

The usage-based approach to language is a non-modular approach, meaning that what in componential theories are called 'the lexicon' and 'the grammar' are not distinct entities, both conceptually and cognitively. Because of this, it is not possible to fully isolate the acquisition of lexical constructions ('words') from that of grammatical constructions. At the very least, the mechanisms involved in them operate in lockstep: it is unthinkable (from any theoretical perspective, really) that the child waits until she has acquired an adult-like lexicon before she starts to figure out the grammatical rules. Starting from a usage-based perspective, even the weaker form (the child waits until it has a lexicon of some size before it starts learning the grammar) is not satisfying. As both are symbolic representations, roughly the same set of mechanisms have to be involved in acquiring them from situated utterances, and as such it has to be possible for a learner to pick them up at the same time. Except for a model that is, interestingly enough, not framed as a usage-based model (Kwiatkowski 2011), none of the models listed in section 1.3 does this. Models that do incorporate meaning start with many lexical form-meaning pairings already acquired (Chang 2008, Alishahi & Stevenson 2008). The first goal of this dissertation therefore, is to develop a computational model of language acquisition that starts with no representational symbolic content and that acquires lexical and grammatical constructions at the same time.

The comprehensiveness of a model bears on another issue too. If language users are able to form utterances on the basis of some conceptualization and a linguistic inventory and understand language on the basis of the utterance and the same linguistic inventory, we have to model the task of language use on both ends. That is to say: we want a model to be able to produce utterances from a conceptualization of some situation and to comprehend utterances with its inventory of linguistic representations to arrive at an interpretation. Again, many computational models model only part of this task and do not show how they can account for comprehension on the basis of an utterance, and production on the basis of only a meaning to be expressed.

1.4.2 The conception of learning

When modeling the acquisition of grammar, cognitive modelers often find inspiration in learning algorithms developed in computational linguistics or artificial intelligence. Even if they turn out to be descriptively adequate, it remains important to reflect upon the conception of learning they encompass. Usage-based theorizing adheres to a general empiricist approach to learning, in which categories are induced from the input rather than deduced given

a pre-existing hypothesis space, as rationalist approaches have it. This does not mean that there is no hypothesis space: given the nature of our cognition, certain representations are possible, whereas other, logically possible ones, are not. This 'design space', however, is not very informative, and it is mainly the learner's exploration of the space through language use that is of interest for usage-based theorists. The learning algorithms used in usage-based models typically reflect this (e.g., unsupervised clustering in Alishahi & Stevenson (2008) and Bayesian Model Merging in Chang (2008)).

Another question is whether learning involves a rational decision making process. Linguistic production and comprehension arguably involves such processes: the speaker selects from his inventory a number of representations that allow him to produce of comprehend an utterance. Whether the learning process is also affected by (subconscious) decision making, is another issue. If we follow Langacker (1988) in the argument that learning is merely the effect of processing linguistic usage events, are there any cognitive mechanisms that only affect the learning taking place after the processing of the usage events but not the processing itself? This is not a question that I believe has been answered yet, but for the research presented here, I believe it to be the best null hypothesis to start from the idea that all learning takes place in the processing of usage events and the reinforcement of representations used in processing.

1.4.3 Starting small

One of the central tenets of construction grammar, as well as the usage-based perspective, is the idea that language users compute language with symbolic representations that are 'bigger' than a single word. This focus has led to the view in language acquisition that children often first acquire larger, unstructured wholes, 'chunks' of linguistic material, which they subsequently start breaking down into their component parts.

There is definitely a lot to be said for this 'starting-big' approach. However, the wholes-to-parts account of language acquisition may be overemphasized in current usage-based approaches. Especially when we look at early production, it seems that, besides the use of chunk-like structures, there is also a gradual build-up of used representations taking place. If children produce increasingly more arguments with a verb, for instance, I believe it is likely that they are extending existing representations with additional valency roles, rather than learning the larger unit first as a chunk, then breaking it down, and only then recognizing its similarity with earlier acquired representations.

1.4.4 Naturalism in meaning

A final issue that is central to this thesis is naturalism in the acquisition of any linguistic structure involving conceptual structure, or meaning. Many modeling approaches, but psycholinguistic studies as well, make simplifying assumptions concerning the question how available the conceptualization of the

situation is to the child. If children are able to understand the communicative intention of the speaker, they must derive the content of this intention from some source. This source is often thought to be the situational context, but if we assume the situational context to provide the learner with a source for her meanings, we face a number of problems. The situational context, after all, is a confusing place if you want to learn meaning. Even granted that the child has joint-attentional skills that narrow down the set of conceptualizations that the speaker can possibly be trying to convey, there are still often many conceptualizations of the same situation possible, and often many situations co-present with the utterance that are likely to be expressed. On the other hand, some situations may simply be absent from the consideration of the learner. If we want to simulate the acquisition of meaningful structures, we have to make a realistic assessment of the overwhelmingness of other possible conceptualizations and the frequency of the absence of the situation which the speaker is trying to express from the learner's consideration.

1.5 A note on notation

In this thesis, I adopt Langacker's (1987) Cognitive Grammar shorthand formalism for describing constructions. I briefly describe this notation at the outset, because I will use it throughout the dissertation. The formalism works as follows. Conceptual structure is represented in small capitals and phonological structure in italics. As a conceptual structure in the case of the model to be developed often has many features (e.g., the feature set behind the meaning of *daddy* being {ENTITY, ANIMATE, MALE, PERSON, FAMILY-MEMBER, FATHER}), I use only the most concrete, or otherwise most recognizable feature to represent the set of features (so *daddy* would signify FATHER). A **unit**, or construction, is a linguistic representation of a number of constituents where each constituent is given between square brackets.

If, in a constituent, a phonological and a conceptual structure are present, the phonological structure can be said to **symbolize** the conceptual structure, which is represented with a slash sign. The whole of constituents is delimited by square brackets as well. If the unit signifies conceptual structure beyond the meaning in the constituents, this meaning will be represented after the formal structure, marked with a pipe '|'.[3] In many cases the non-compositional meaning is not directly relevant, and will be omitted for the sake of space. Examples are given in (8) and (9).

(8) [BALL / *ball*]

(9) [[HUMAN] [GRAB / *grab*]] | GRAB(GRABBER(HUMAN))

Using this formalism, we can also describe complex analyses or constructs. Suppose the slot of the construction in (9) is filled (by means of what Lan-

[3] Here, my notation differs from Langacker's.

gacker calls 'elaboration') with a construction [FATHER / *daddy*]. In that case we denote this slot-filling operation with an arrow after the slot, followed by the unit filling the slot, as in example (10).

(10) [[HUMAN]→[FATHER / *daddy*] [GRAB / *grab*]] |
GRAB(GRABBER(FATHER))

Another feature of Langacker's notation is the use of parentheses for extensions of known units. I will employ this notation in several cases. To give an example: the model to be presented allows for the concatenation of two partial analyses of an utterance without any construction governing both of them. We find a case of this concatenation operation in (11), where the partial analysis in (10) is concatenated with the unit given in (8).

(11) ([[HUMAN]→[FATHER / *daddy*] [GRAB / *grab*]] [BALL / *ball*]) |
GRAB(GRABBER(FATHER),GRABBEE(BALL))

Similarly, the notation in parentheses will be used for the so-called 'bootstrapping' of unknown words into open slots of constructions. When a word is bootstrapped into a schematic position of a construction, a novel representation is created that links the hypothesized meaning of that word to its form. An example of the notation of the bootstrapping process is given in (12), where the unknown word *epidemiologist* is bootstrapped into the slot of the construction we saw before in examples (9) and (10).

(12) [[HUMAN]→(EPIDEMIOLOGIST / *epidemiologist*) [GRAB / *grab*]] |
GRAB(GRABBER(EPIDEMIOLOGIST))

1.6 Overview of the dissertation

The core of this dissertation is the model, the Syntagmatic-Paradigmatic Learner (SPL), presented in chapter 3. I set out the theoretical and empirical issues which I believe are worth studying using a computational model in chapter 2. After this theoretical core, we first look at the issue of the acquisition of meaning in chapter 4. This chapter does not constitute a modeling experiment, but rather an observational study into the observable sources of meaning. The results of this study are used in the subsequent three chapters that report several aspects of simulation experiments performed with the model. In chapter 5, first, we look at the model's ability to comprehend utterances in a noisy situational context. Chapter 6 presents a 'look under the hood': what is the nature of the representations acquired over time by the model, and what learning mechanisms are involved in doing so. In chapter 7, finally, I discuss the results of several production studies: what happens if we give the model a situation and ask it to express it.

CHAPTER 2

A usage-based conception of language acquisition

The research reported in this dissertation consists of a computational model that aims to operationalize the key concepts of a usage-based theory of language acquisition in order to investigate their validity. In this chapter, I present the usage-based perspective in some more detail (section 2.1) and criticize several of its proposals (section 2.2). This theoretical discussion leads to a set of theoretical desiderata which I believe a computational model should meet (section 2.3). These desiderata are not so much empirical constraints, but rather theoretical constraints on the kind of computational model to be built. The class of algorithms and representations accounting for the linguistic behavior of the child may be vast, but as such this class is not very interesting. What we want to know, echoing, but non-trivially reinterpreting, Chomsky's (1965) distinction between descriptive adequacy and explanatory adequacy,[1]

[1] Reinterpreting, because Chomsky's notion is theory-laden, as the programmatic formulations in the following quotes in Chomsky (1962) suggest:

- "In short, *I think* [emphasis mine, BB] that the development of the theory of grammar [as, a.o.t., a class of potential grammars BB], and intensive application of this theory is a necessary prerequisite to any serious study of the problem of language acquisition." (p. 534)
- "*[I]t seems to me* [emphasis mine BB] that the scope and effectiveness of heuristic, inductive procedures has been greatly exaggerated. [...] the task remaining to heuristic procedures is obviously lightened as we make the specification of the form of grammars increasingly narrow and restrictive" (p. 536).

I consider these claims, in conjunction, to contain a central programmatic commitment of the generative approach to look for restrictions on representations to explain the structure of language. This commitment can and has be contested, but as such it is not an empirical claim (just as the competing commitment to an explanation of the limitations on the representational struc-

is to what extent a learning theory t, as operationalized in a computational model, is a better (more principled) cognitive theory than a competing theory t'.

Next, I discuss how the usage-based perspective accounts for several broad-scale phenomena of child language acquisition (section 2.4). The work discussed in this section gives us a set of empirical explananda that a computational model starting from the usage-based vantage point has to meet as well. In section 2.5, I will present several usage-based computational cognitive models of language acquisition and discuss how they fare against the desiderata and explananda.

2.1 Usage-based linguistics and language acquisition

Over the recent history of linguistics, the focus of attention has shifted from a conception of language as a structural, abstract entity, to a more fine-grained conception of language as a cognitive and social phenomenon (cf. Chomsky's (1986) 'I-language' and 'E-language', but also de Saussure's (1916) 'langue' and 'parole'). In this dissertation, I focus on the cognitive side of the medal, without denying the reality of language as a social phenomenon. When we take the perspective of language as a cognitive phenomenon, several positions concerning the characterization of the cognitive representation of language are possible. The generative approach, broadly speaking, characterizes linguistic knowledge, in particular grammatical knowledge, as a modular cognitive system, that maintains interfaces with other cognitive modules (such as the sensory-motoric system for producing and processing sound, and the conceptual-intentional system, where the conceptualization underlying meaning resides). The core of grammatical knowledge is not only modular in the way it can be analyzed and described, but also ontologically: it is cognitively independent from either the two systems it interfaces with, as well as from language use. The usage-based and constructivist view can best be understood, at least historically, as denying the modularity of the grammatical system. Starting with Langacker (1988), who coined the term, the usage-based perspective holds that linguistic representations are grounded in experiences of language use. The cognitive processes involved in processing linguistic structures are furthermore not specific to the domain of language, but are shared with other cognitive domains, such as planning and inference.

This conception has several theoretical consequences. First of all, the representational system is tightly linked to the processing of language in com-

tures and contents used from restrictions on the individual (the 'heuristic procedures') and social mechanisms, as is found in most usage-based approaches, cf. infra, is not an empirical claim, but a research program). Abstracting over this and other programmatic parts, what is left of the notion of explanatory adequacy is a predictive theory that provides a principled choice between alternative theories that may cover the data equally well.

prehension and production. This means that the representations a language user employs in producing and understanding utterances are instructions for linking sound (or any other kind of observable signal) to a conceptualization, or: meaning. Construction grammar (Fillmore, Kay & O'Connor 1988, Goldberg 1995) gives representational hands and feet to this idea by arguing that all linguistic knowledge, both 'words' and 'grammatical rules' constist of pairings of form and meaning.

The language learning child, having no preconception of what the system may look like, gradually finds the ways of linking sound to meaning that are conventional in her community. Langacker (1988) characterizes the system of constructions that thus emerges as non-reductive, bottom-up and maximalist. Assuming that the mind cares little about economy of storage, the set of constructions may display redundancy. More general or abstract patterns only come into existence after the language user has found evidence for the abstraction in the overlap between multiple more concrete constructions. However, these patterns are not separate cognitive entities, 'extracted' from more concrete instances and stored elsewhere, but are rather 'immanent' in the maximally concrete representations of the usage-events themselves (Langacker 2000, 7).

Importantly, if this abstraction consists of the mere reinforcement of shared elements of form and function, abstraction per se can be thought to occur early. This seems to contradict the findings (by the same usage-based researchers) that children are conservative in generalizing abstract patterns to novel usage events (see section 2.4.1 below), but essentially does not do so. Although abstractions may arise as soon as some shared structure between two constructions is detected, the representational strength of this overlap may initially be too weak for the shared structure to be used. Through the recognition, use, and hence reinforcement, of this shared structure, it may grow stronger in representational strength, and become increasingly likely to be applied anew. This perspective can be found in Langacker (2000), and, from the perspective of the apparent opposition between exemplar-based and prototype-based views, in Abbot-Smith & Tomasello (2006).

On the usage-based view, child language acquisition is furthermore not something 'special' in the sense of being qualitatively distinct from what adults do: language-acquiring children try to interact with their caregivers and siblings, and in the process of doing so, acquire an inventory of constructions that facilitate that interaction. The mechanisms whereby the inventory emerges, as a by-product of language processing rather than as a goal in itself, are still active in adults. Nonetheless, early linguistic development is the phase in which some of these processes become most evident, and the study of the mechanisms whereby representations are acquired has been most comprehensive for children.

2.1.1 Constructions and the constructicon

Constructions

If the processing of linguistic experiences leaves traces in the mind, and if the representations are not separate entities from these traces, it can be expected that linguistic representations consist of elements of the situated linguistic experience and no elements from domains outside of it. This means that conceptual structure and (for spoken languages) phonological structure are the main building blocks of linguistic representations.[2] Words, as the prototypical form-meaning pairings, can be easily explained from this vantage point: a word is a recurrent experience of some phonological structure and some conceptual structure which the language users assume to be conventional. The resulting representation is an instruction about inference, if we follow the constructivist tenet that constructions are signs in the Saussurean sense (cf. Fillmore et al. 1988, Goldberg 1995): if a hearer hears a particular phonological string, he should make the inference that the speaker has a certain conceptualization in mind that she wants to convey. 'Big words', such as fixed idioms (e.g., *how are you?*, *top of the morning*) can be accounted for similarly. But according to construction grammar, all sorts of ways of productively combining smaller word-like constructions into larger, structured, wholes, that is: those things traditionally considered to be the grammar, including morphology, are symbolic units as well (Langacker 1987, Goldberg 1995).

Goldberg's (1995) case of argument-structure patterns constitutes an insightful example. For a sentence like *he gave Pat a book*, the 'transfer' sense and the distribution of conceptual roles over the nominal arguments can be said to be part of the lexical representation of the verb *give*. Such an account becomes problematic for *she smiled herself an upgrade*, where it is less sensible to assume that the 'resultative' sense and the interpretation of the nominal arguments ('she caused herself to be in the state of receiving an upgrade by means of smiling') is part of the lexical representation of *smile*. Rather, Goldberg argues, language users have an inventory of grammatical constructions that contribute elements of meaning themselves to the composite conceptualization. This means that the instructions for combining elements into larger wholes should be considered signs as well.

When we assume that constructions only consist of phonological and conceptual structure, grammatical constructions pose a problem. If a sentence such as *she smiled herself an upgrade* is generated with a 'resultative construction', what is the form of this construction? For words, again, the question what the form is, is easily answered: it is the phonological structure. However, in many construction-grammatical descriptions, no phonological form is specified, and we can find descriptions of a resultative constructions such as:

(13) [**form**: NP_i V NP_j NP_k | **meaning** 'i causes j to be in a state of k']

[2] In fact, it is likely that language users store more information besides phonological and conceptual structure, e.g., social information (Geeraerts 2010).

This description involves representational entities that are not phonological or conceptual structure, such as 'noun phrase' (NP) and 'verb' (V). Traditionally, notions such as 'subject', 'noun phrase', and 'verb' have been called grammatical categories or relations, but their role as primitives in a usage-based theory is doubtful. As Croft (2001) argues, their universality can be doubted on methodological and empirical grounds. His solution is to reverse the line of reasoning and consider the constructions primary and the paradigms they define (e.g., the four positions in the resultative construction above) as derived. Multiple paradigms across constructions may display a certain overlap, creating (the appearance of) a category. However, even if such high-level cross-paradigmatic distributional categories exist, they are construction-specific, and, more importantly, language-specific. Croft's proposal, then, is that the slots suggested by the paradigms are in the end not defined in distributional terms, but in conceptual ones, thereby reducing the notion of grammatical form. Nonetheless, Croft does assume that a level of 'morphosyntactic structure' is part of the form side of a form-meaning pairing (Croft 2001, 62)

Langacker (2005) reaches a slightly different conclusion. In his theory of Cognitive Grammar (cf. Langacker 1987), a sign consists of a semantic (conceptual) and a phonological pole. There is no room for grammatical categories and, as Langacker argues, these can be reduced to conceptual structure. This reduction constitutes a more parsimonious position and is for that reason worth pursuing. Langacker therefore proposes that the 'form'-side of constructions only consist of phonological structure. Verhagen (2009) points out that this view is problematic as well, as a completely unspecified phonological structure provides no constraints on what can fit in a slot and thus any phonological string can be used to make the hearer recognize that part of a construction in processing language. Verhagen argues that it is fruitful to distinguish the notion of 'form' as referring to phonological structure from 'form' as referring to the signifier of a construction (i.e., "what triggers the inference of something unobservable", p. 136). When we consider a construction to be a sign, there is nothing that restricts us from considering conceptual structure to signify, or: trigger the inference of a more encompassing conceptual structure, as well.[3] This means that we still need only phonological and conceptual structure, but that these building blocks are (partially) orthogonal to the roles they fulfill in constructions: phonological as well as conceptual structure may signify, but only conceptual content can be signified. Membership of a paradigm of a construction, or strong conceptual and/or phonological resemblance to units that are members of that paradigm can thus be said to signify as well. In this research, I follow Verhagen's critique of Langacker (2005) and consider as the form side (or rather: the signifying side) of a construction anything that can trigger the inference of a more complex, unobserved conceptual structure. This means that both phonological and conceptual structure can constitute the

[3]Interestingly, as Verhagen notes, this grounds the processing of grammatical construction in the domain-general skill of understanding part-whole relations metonymically.

signifying side of a construction.

I therefore take it to be crucial to the notion of constructions (be they lexical or grammatical) as signs to consider them as conventional instructions for making an inference. Conventionality, in Lewis's (1969) sense, means that constructions are mutually shared solutions to a coordination problem. This aspect becomes important in the acquisition of the constructions, as it allows the learner to make use of additional pragmatically-defined notions such as contrast.

The constructicon

In order to produce and understand utterances, a speaker needs many constructions, differing in size and shape. This inventory, often called the **constructicon**, is not an unordered list, but is typically conceived as a network in which constructions bear different kinds of relationships to each other. A commonly assumed relationship is that of ancestry, meaning that if one construction is another construction's parent, the other construction has all features the one construction has, and more, given the complete inheritance position (cf. Croft & Cruse 2004, 270). Besides complete inheritance, normal inheritance has also been hypothesized to be a possible ancestral link between constructions in the network (for a discussion, see Croft & Cruse 2004, 275-276). Normal inheritance is the situation whereby certain features from a parent construction are inherited by a construction, but others are not inherited, for instance because they conflict with another parent of that construction. In this research, inheritance plays no role. In fact, one can consider inheritance to be a superfluous aspect of the constructivist theory if a usage-based perspective is taken. If abstraction is immanent in the more concrete representations it is derived from, the notion of inheritance emphasizes the misguiding metaphor of abstract constructions being separate cognitive entities, which makes the point of deciding between complete and default inheritance a moot one. Perhaps inheritance has mainly a descriptive function, but I fail to see an important role for it in a theory of linguistic cognition.

As the constructicon comes into being through experience with language, the representational strength of the various constructions can be thought to reflect the amount of experience with them. Bybee (2006) discusses two ways in which the amount of experience affects the representational strength. On the one hand, there are several effects of a construction having a high token frequency, that is: the amount of times that particular construction has been processed. A high token frequency leads to the automatization of the unit: the more frequently a unit is processed, the more readily it will be used in the future as a whole. It will lose its internal structure and possibly be phonologically reduced. The other effect Bybee discusses is that of a high type frequency over an element of a construction. The more different units are found filling the slot of a construction, the more the language user will expect that slot to be extended to be filled with even other units. That is to say: a language user

expects a slot to be more productive the more different items are used in it. This expectation is constrained by the (functional) generalization that can be made over the items filling the slot. The acceptability of novel items, then, is (co-)determined by semantic fit with the slot (cf. Ambridge 2013).

2.1.2 Producing and understanding an utterance

Despite the adherence to the idea of language use as the central factor in the formation of linguistic representations, most usage-based work still focuses on the representational properties of the constructions rather than how they are used. Nevertheless, we find several ideas about the use of constructions in the literature.

Langacker (2000) describes the operation of combining representational units into larger, complex, wholes as the **composition** of linguistic units, or constructions. In order for language users to compose multiple units, they first have to recognize them, either by identifying a part of the linguistic usage event (the utterance and the conceptual context) with them, or by extending them to fit a particular part of the linguistic usage event (p.12)

Complex expressions, then, are assemblies of recognized symbolic structures. Importantly, composition in a situated context always involves an element of non-compositionality. That is: there are always aspects of the joint meaning of two units that go beyond the contributions of the two items themselves. Langacker gives the example of novel noun-noun compounds in English: although we may have a schematized representation allowing us to assign a generic meaning to two juxtaposed nouns, we always understand the conceptual value or meaning of the composition in a context. This conceptual value is not just the situated pragmatic resolution of a more abstract nominal-compound meaning, as it can become part of the conventional meaning of that particular nominal compound over historical time and the lexicalized meaning of that compound for a language user. The identification of the conceptual value with the nominal compound (say: *beer belly*) is, on the first encounter, an elaboration of the schematized meaning of nominal compounds (something like example (14), with the analysis in example (15)). Because a language user stores all conceptual detail present along with the specific form *beer belly* (in the form of a neural co-activation pattern), he can, upon future encounters, identify part of the potentially intended meaning by means of identification of the more concrete representation in example (16), rather than the on-the-fly elaboration of the pattern in (14).

(14) [[ENTITY$_i$] [ENTITY$_j$]] | ENTITY$_j$ STANDS IN A CERTAIN RELATION TO ENTITY$_i$

(15) [[ENTITY$_i$]→[BEER / *beer*] [ENTITY$_j$]→[BELLY / *belly*] | BELLY$_j$ IN A PARTICULAR STATE THROUGH THE CONSUMPTION OF BEER$_i$

(16) [[BEER$_i$ / *beer*] [BELLY$_j$ / *belly*]] | BELLY$_j$ IN A PARTICULAR STATE THROUGH THE CONSUMPTION OF BEER$_i$

This example also serves to illustrate another concept, viz. the primacy of low-level units over more schematic or abstract ones in Langacker's conception of language use. Because a low-level schema shares more features with the target of identification (i.e., the conceptual space and the phonological structure of the speech situation), it is more readily activated and hence a better candidate for being selected as the activated unit. More generally, a language user will often have multiple units at his disposal for interpreting or producing an utterance. These units then compete with each other. Langacker describes this as a more general categorization problem: out of an 'activation set' of potentially applicable units, one has to be selected as the 'active structure'. Both the degree of entrenchment of the units and their fit with the conceptual and phonological structure play a role in deciding which unit wins the competition.

Langacker's notion of composition involves all sorts of operations. Although he does not explicitly describe them, we find in the examples two types: slot-filler operations, whereby one unit is used as the constituent of another unit and the juxtaposition of two units. These units are similar to the ones Dąbrowska (2014) describes, namely **superimposition** and **juxtaposition**. In the former, two constructions are combined in such a way that the corresponding elements are 'fused' (p. 623). This can be through regular slot-filling, but also through overlaying two constructions (e.g., the hypothetical [[HEARER / *you*] [ACTION] [OBJECT / *it*]] and the [[ENTITY$_i$] [GET / *get*] [ENTITY$_j$]] constructions). Juxtaposition, on the other hand, merely involves taking two constructions and listing them in some order. In production, multiple verbalizations are possible given the same conceptualization, but the one that is retrieved first (i.e., the one with the most highly-entrenched constructions) will be selected, unless the speaker rejects it, for instance because of a low fit with the communicative intent.

Dąbrowska furthermore makes a difference between a holistic and an analytic mode. In the former, language users use their highly concrete schemas to arrive at an utterance. In the analytic mode, language users use the more abstract schemas. As Bod (2009) argued, and in line with Langacker's (2000) perspective, we can also regard the maximally concrete and maximally abstract schemas to be the end points of a scale. Following that line of reasoning, there may not be something like an analytic mode as opposed to a holistic mode: language users will try to stay as close as possible to what they know about the conventions of the language (i.e., the maximally concrete schemas), while sometimes experiencing the need to rely on more abstract units when novel meanings need to be expressed.

2.1.3 Acquiring a grammar

On the usage-based account, acquiring a language is a process that is grounded in the processing of linguistic usage events with symbolic constructions, which is in principle identical for adults and infants alike. This continuity is impor-

tant, as discontinuity would be a less parsimonious explanation in want of an additional explanation. Whereas in Pinker's (1984) version of the continuity assumption, the contents of the representations are equal over time, in Tomasello's (2003) version, the contents may vary over time, but the mechanisms with which language is processed, intentions are understood, and patterns are learned remain the same over time. However, perhaps a slightly stronger claim to content continuity can be made as well: if constructions consist of phonological and conceptual structure throughout development, a usage-based account can also claim content continuity on a qualitative level: constructions are built up out of phonological and conceptual components and this property is stable over time.

Tomasello (2003) assumes that two sets of domain-general cognitive capacities are central to answering the question how infants acquire an inventory of linguistic symbols allowing them to be proficient communicative agents in their communities, viz. intention-reading skills and cognitive pattern-finding mechanisms. Importantly, the same sets are used to acquire both words and grammatical patterns. The former allow a child to understand that other people are mental agents, with (communicative) intentions and belief states, like herself. On the basis of this understanding, the child can engage in cultural learning (cf. Tomasello 1999), that is: the reverse engineering of behavioral solutions to repeated coordination problems. Directing someone's attention, manipulating someone's behavior or knowledge state, or engaging in joint projects constitute some of these problems, a subset we could call 'social coordination'. The language of a community can be regarded as the set of solutions of that community to these coordination problems (cf. Lewis 1969). The task of the language-acquiring child then, is to use her intention-reading and pattern-finding mechanisms to find out what these solutions are.

In this research I will focus on the latter set, the pattern-finding mechanisms, as it is more evident how a formal operationalization of these mechanisms may work and may help shed light on the hypothesized processes. As pattern-finding mechanisms, Tomasello lists such things as the ability to build up perceptual and conceptual categories, the ability to form sensory-motor schemas, performing distributional analysis over perceptual and behavioral sequences, and being able to analogize over larger structures, finding the commonalities and differences (Tomasello 2003, 4). All of these skills are available to the child before she starts to speak, but it is only when social understanding starts to develop around the child's first birthday that the child will substantially put them to work.

Using these mechanisms, the language-learning child will gradually progress from a state of knowing holophrases (single-word utterances referring to a complete situation) and chunks (unanalyzed multi-word utterances), via semi-abstract patterns, in which few constituents are not lexically specific and may vary, to the more abstract patterns we typically assign to adults. Underlying this behavioral development, Tomasello (2003, 295-305) hypothesizes a number of specific pattern-finding operations. Schematization, first off, is the

process whereby the child observes that, over multiple utterances, some elements are identical while others vary, and that at the same time, some parts of the communicative intent are identical, while other parts vary. The varying elements are then abstracted over and a construction with a functionally defined slot emerges. Elements common across these slots in different patterns may give rise, through functionally based distributional analysis, to highly abstract categories, such as 'noun' and 'noun phrase'. However, as Tomasello argues, these more abstract categories are still grounded in the function these elements fulfill communicatively. Secondly, processes of entrenchment (routinization, automatization) cause certain symbols to be processed more readily than others. A pattern becomes more entrenched the more it is processed. Entrenchment plays a central role in the competition between similar patterns, as we have seen in section 2.1.2.

In Tomasello's older work, we find a slightly different exposition of the learning mechanisms (Tomasello 1992, 234; 250-253). Two processes of 'constructional integration' he discusses there are the expansion of paradigms (the widening of the initially narrow paradigmatic categories) and the addition of syntagmatic terms. In the former case, the allowed complexity of elements filling a paradigmatic position of a rule changes over time such that more complex linguistic structures are allowed in there. It is not clear whether Tomasello also includes the widening of the semantic scope of such a paradigmatic position. The addition of syntagmatic terms simply means that another semantic dependent is added to some semantic head. In the case of argument structure constructions, this means that another argument or adjunct of the verb can be expressed.

In addition to the parts-to-whole line of learning proposed by Tomasello, usage-based theorists often also argue that early representations may not even have an internal syntagmatic structure, but consist of unanalyzed chunks, in which gradually 'slots' over varying elements are learned (Bannard, Lieven & Tomasello 2009, Arnon 2010). That is to say, we can conceptualize a gradual build-up of the grammar both through bottom-up procedures, going from the parts to the wholes, as in Tomasello's explanation, or through top-down procedures, going from the wholes to the parts, as in Arnon's (2010) explanation. The latter has been emphasized in usage-based studies, for historical reasons, but there are good reasons to also consider parts-to-whole learning to be a central mechanism in a usage-based account, as I will argue in section 2.2.3.

Two final processes of learning in Tomasello's (1992) account are single-verb coordination and two-verb coordination. The former happens when two constructions with the same head but different dependency valencies are used at the same time. In the case of the latter, the complexity of the production is increased by having a complex structure (with a head and a dependent) as a dependent of another construction. The development of the paradigmatic positions and the syntagmatic relations is an instance of general categorization according to Tomasello, where the emergent paradigms are an "organizational outgrowth of the process of constructing syntagmatic structures", i.e., a by-

product of the attempt to construct complex messages using the syntagmatic relations the child has picked up by then.

2.2 Theoretical issues with the usage-based perspective

In the previous section, I have given a brief account of the usage-based perspective on language. The presentation of the mechanisms and representations serves as a starting point for developing a computational model of the acquisition of grammar. Several issues, however, are in want of further clarification or elaboration.

2.2.1 Representational metaphors: blocks and streams

Both in theoretical linguistic (e.g. Tomasello 2003, Goldberg 2006, Dąbrowska 2014) and computational modeling work (e.g. Chang 2008), we find descriptions of the process of language use involving discrete building blocks that are combined into composite constructs by filling the slot of one construction with another construction. However, Langacker (2000, 8) argues that the conception of language use as stacking together building blocks is incomplete at best. Given the dynamic perspective on representation (units or constructions are always in development and any abstraction is immanent in the memories of the concrete usage events), the building-block metaphor unduly emphasizes a static nature of the units. True as this may be, I believe that this should not stop us from cautiously pursuing the metaphor of 'building blocks', as it does yield great explanatory value. What I mean with explanatory value is the following. When operationalizing a theory in computational terms, many researchers take recourse to static structures and are able to explain what happens in the model in the process of composing the building blocks. Such models thus use a metaphor, which may foreground certain aspects of our conception of linguistic representations and the way they form composite structures, while backgrounding other (nonetheless important) ones. The fact that it is relatively easy to 'look under the hood' and see what the model does in analyzing novel utterances, gives it the power to corroborate linguistic theories with relative ease.[4]

When one adopts a more classic parsing approach, as is taken in this research, the explanation of what is happening is relatively straightforward, as will be seen in later chapters. Acknowledging the inherently metaphorical nature of the formalization (just as Langacker does when he draws discrete box-diagrammatic representations and illustrates their combination by using mul-

[4]Of course, it is conceivable to radically rethink language use in more dynamic terms. Connectionist models such as McClelland & Kawamoto (1986) do exactly this. The problem with such models, as noted in section 2.3, desideratum D7, is that interpreting them becomes difficult: what happens in a neural network is to a large extent a black box.

tiple such representations) is then a cautionary note which has to be kept in mind, but which should not stop us from using that metaphor for explanatory purposes.

Nonetheless, there are approaches which do emphasize a more dynamic conception of linguistic structure. Worth acknowledging here is Borensztajn's (2011) model, in which a parse is a path through a self-organizing space. This space corresponds to a continuous version of distributional categories, akin to part-of-speech and other grammatical categories. By using this continuous space, Borensztajn does away with the often faulty metaphor of discreteness of categories while keeping the resulting parses relatively interpretable. However, one reason why I believe the parses in Borensztajn's model are interpretable, is that the parses themselves do constitute well-defined discrete graphical representations (i.e., trees). If we were to give up on the discrete nature of the composition, as one should when taking Langacker's (2000) ideas to their extreme, I believe the potential for the linguist's interpretation of the resulting analyses would be seriously impeded.

2.2.2 Mechanisms operating on early representations

The parsing approach taken in this research does not imply that we have to use the traditional operations defined for parsing in the computational-linguistic literature. Most linguistic theories regard as their central operator a single operator that allows a language user to combine structure recursively in order to build a composite hierarchical structure. An interesting question following from this conception of adult linguistic competence is whether language-acquiring children use hierarchical structure building mechanisms from the start as well. But it may be the case that early linguistic perception and production is to a large extent guided by non-hierarchical rules, and that the use of hierarchical rules only emerges later. To take this argument a little further, it may be the case that some (or even a lot) of adult linguistic processing is governed by processing mechanisms that are simpler and cognitively cheaper than building hierarchical structure. The burden of proof, however, is on those arguing for a multitude of mechanisms, as it is a less parsimonious account.

Tomasello's (1992, 2003) perspective is of interest here. Tomasello (2003, 226) argues that when very early multiword utterances are not rote learned, they are not by necessity 'grammatical' in the sense that we say an adult's production is, but rather instances of mere concatenation of linguistic items (even if the order of the concatenation is copied from the input language). For something like word order to be 'grammatical', in Tomasello's (1992) view, it has to be used contrastively with another word order.[5] Importantly, Tomasello

[5]There is, however, a conceptual problem with defining 'grammatical' (in general) in terms of existing in a system of oppositions. Certain word orders (e.g., determiner-noun in English) are non-contrastive (noun-determiner does not mean something different, it rather is ungrammatical), yet I would be reluctant to call this element of linguistic knowledge non-grammatical. This is to say that some grammatical rules may be conventions and known as such without them be-

(1992, 259) argues that concatenation, the mere stringing together of linguistic elements, is a possible mechanism that operates before the use of mechanisms giving rise to (hierarchical) structure. A related take on early production is the 'groping patterns' approach, which I will discuss in greater detail in section 2.4.3.

The idea that children may initially operate with a mechanism that linearly concatenates linguistic material is to be dispreferred on prior grounds, namely as a violation of the continuity assumption. However, the continuity assumption rests on the preconception that adults only use a single mechanism of hierarchical structure building. In several recent works outside language acquisition, we find support for the position that different cognitive structure-building mechanisms are at work in language production in the linguistic literature. Jackendoff (2002), and more recently Jackendoff & Wittenberg (2014) argue that, although cognitively all methods of composition (stringing, combining into hierarchies) are available to all humans, different languages employ different levels of syntactic complexity to different extents. Jackendoff & Wittenberg (2014) discern levels such as 'word concatenation grammars', 'simple (non-recursive) phrase grammars' and 'recursive phrase grammars' (with increasing complexity). Simpler mechanisms, Jackendoff & Wittenberg (2014, 1) argue, "put more responsibility for comprehension on pragmatics and discourse context", as the syntax does not restrict the interpretation in structural ways. On a biologically, as well as culturally phylogenetic timescale, Jackendoff & Wittenberg (2014, 16-17) suggest, the simpler mechanisms probably precede the more complex ones, although they admit that this is speculative, and at most plausible. A similar perspective is put forth in Frank, Bod & Christiansen (2012), who argue that linear processing (most akin to Jackendoff & Wittenberg's (2014) simple phrase grammar) may be what language users rely on most in processing and producing language, relating it to psycholinguistic processing studies (Frank & Bod 2011) rather than to structural analysis, as Jackendoff & Wittenberg (2014) do.

Interestingly, Jackendoff & Wittenberg (2014) claim that the early grammatical production of a given language may rely more on the simpler mechanisms than the adult's grammatical production of that language, thus allowing for some degree of quantitative discontinuity in the mechanisms used by the language-learning infant and the adult: "As the child's grammar acquires more grammatical devices, it provides more resources for making complex thoughts explicit, reducing the workload on the hearer" (Jackendoff & Wittenberg 2014, 2). In other words: the increase of complexity of the acquired grammatical structure allows the child to verbalize more complex thoughts.

Concluding, both structural linguistic and psycholinguistic studies provide evidence that language users employ grammatical representations at different levels of complexity in linguistic processing. Tomasello's perspective

ing contrastive. Furthermore, when a child uses *cup fell* and *fell cup* interchangably, it does not necessarily mean that the child is just stringing together words (even though this is the most parsimonious analysis), an issue discussed further in section 2.4.3.

that early combinatorial productivity may not be guided by rules building hierarchical structure may be on the one hand less parsimonious, but may also reflect the cognitive mechanisms underlying (early) language production and understanding better. In the model to be developed, I will start off from this perspective, arguing that concatenation-like mechanism may play a central role in the development of a hierarchy-building mechanism but are not used in production.

2.2.3 Gradualism and simultaneity in learning

Whereas the usage-based conception has a relatively clear explanation of the development of linguistic behavior once some knowledge is in place, the initial emergence of that knowledge remains somewhat obscure. Two questions require some more attention. First, how do the initial holophrastic, chunk-like units and later lexical constructions come into being? Second: how do grammatical constructions develop from the initial lexical units?

Acquiring lexical constructions

The acquisition of lexical units is a process that is ongoing during the whole life of a language user. Given Tomasello's notion of developmental continuity, we should expect the same mechanisms to be available to infants learning their first words, and an adult language user learning a new word, and in fact, they do (Golinkoff, Hirsh-Pasek, Bailey & Wenger 1992, Landau, Smith & Jones 1992). However, als Hollich, Hirsh-Pasek & Golinkoff (2000) suggest, the relative weight of different mechanisms may vary over time, and, contrary to the null-hypothesis of developmental continuity, different mechanisms may emerge at various points during ontogeny.

The varying weight of different mechanisms is easily explained from a usage-based perspective. Assume that a language user always has the capacity to identify the meaning of a phonological string in a top-down way, that is by looking at the syntactic context (e.g., *that's a WORD* vs *Look! He's WORDing*; (cf. Brown 1957)). To do so, a learner first needs to know grammatical constructions and the paradigmatic distribution over the words on the position of the novel word. This requires an inventory of linguistic units to be in place already, so this way of learning cannot be used at the very start of language acquisition.

At the very beginning, there must be some more naïve form of association, one that is less guided by knowledge of the rest of the linguistic system. This mode of learning should be available throughout ontogeny, but can be expected to lose ground to the more structure-dependent modes of learning lexical units, such as the one described above, as they are far more powerful, allow for immediate inference about a word's meaning, and allow the language user to acquire terms whose meaning is not easily identified in the situational context (see also Gleitman, Cassidy, Nappa, Papafragou & Trueswell 2005).

Acquiring grammatical constructions

Children's initial productions are limited in length and, according to the usage-based view, in the amount of combinatoriality involved. Presumably because usage-based research argues against a nativist view in which abstract grammatical categories are available to the child before the acquisition commences, the latter has attracted more attention than the former. However, the developmental pathway leading from *daddy give* via *daddy give it* to *daddy should give it to me* remains unexplained under a strict 'starting-big' perspective.

One mechanism that is often invoked is the break-down of larger chunks. The learner does the 'blame assignment' of the parts of the chunks by function-based distributional analysis. This is a whole-to-parts strategy. Tomasello (2003, 39) acknowledges that a parts-to-whole strategy is likely to be employed as well, but does not go into the question how this works, nor do we find accounts of this procedure elsewhere. Nonetheless, I believe understanding how the parts-to-whole acquisition of grammatical units works is crucial for a full understanding because whole-to-parts learning gives us an incomplete understanding for two reasons.

First of all, it is doubtful that early learners are able to process the full phonological and conceptual structure without having any linguistic units to analyze them with. The problem is similar to recalling meaningless strings of numbers: the string 07011987 is, as such, hard to memorize. Once one regards it as a date, January 7, 1987, the string of numbers itself can be memorized more easily as well. Finally, if one happens to be the author of this dissertation, the date gestalt becomes even more meaningful, as it is his date of birth. For linguistic processing, I expect, we find the same: more of the conceptual and phonological structure in the speech situation can be processed if we have linguistic gestalts (i.e., constructions) to analyze the speech situation with. An early learner will thus not be able to process as much of the speech situation as an adult, leading to processed experiences that are of a lower granularity and level of detail than those processed by adults.

The effect of this is that parts-to-whole learning quite naturally follows from Roger Brown's law of cumulative complexity (Brown 1973, 186). A grammatical phenomenon f is cumulatively more complex than a phenomenon f' if f involves everything that f' does plus something else. The developmental law associated with this is that we expect, ceteris paribus, f' to emerge in behavior after f. As a cognitive law, we could formulate this as follows: a representation r is cumulatively more complex than r' if r involves everything r' does and more. Developmentally, we expect r to arise before r'.

Note that this covers both parts-to-whole and whole-to-parts learning. Adding more parts to, say, a verb-argument construction, should happen incrementally whereby verb-argument constructions with fewer arguments should precede ones with more. On the whole-to-parts side this means that finding out what parts of a thitherto unanalyzed chunk play certain roles in the chunk should be an incremental process.

The second reason why more attention to parts-to-whole learning is desirable is an empirical one. Children do use apparently unanalyzed multi-word units from very early on, but large parts of early production also consists of single-word utterances and two-word utterances that are not very chunk-like in nature. Tomasello (2003, 39) admits that little is known about why children begin with one-item units instead of larger productions, but dismisses purely working-memory-based accounts (p. 312), a suggestion backed up by the study of Berk & Lillo-Martin (2012), who argue that it is a development specific to language.[6] If chunk-learning is the primary mode, why do children have such early control over well-segmented single words?

After this phase, it seems that not all that is produced is chunk-like in nature either. When an 18-month old says *daddy get ball*, do they have only an unanalyzed chunk or do they know what (at least) *daddy*, (possibly) *ball*, and (maybe also) *get* mean? This is an empirical issue, but I find it harder to believe that *daddy get ball* is an unanalyzed chunk than I find it to believe that *where's the ball* is one.

That is: different units may have been built up in different ways. In the case of *daddy get ball*, the child has possibly processed something like *daddy will get the ball for you*, with the child knowing the nouns *daddy* and *ball* in advance, thereby 'bootstrapping' the meaning of *get* and leaving out the phonologically weak items *will*, *the* and *for you*. This way, the child has done the blame assignment in the initial processing of the pattern: no undersegmentation (in the sense of Peters 1983) takes place.

For *where's the ball*, it may be different: the child knows what *ball* means, and stores the whole substring *where's the* along with the communicative intent of the speaker (she is looking for something), almost as a kind of interrogative-locational prefix. Only later, *where's the* is broken down, when the child encounters utterances such as *who's the oldest man*, *where was the book*, *what's a beer belly* and so on.

Finally, if we adopt the view that a parts-to-whole learning process plays an important role as well, a desirable consequence is that we achieve a further integration of the theoretical apparatus of the usage-based conception. Recall that both Langacker (2000) and Dąbrowska (2014) allow for the juxtaposition of linguistic units. The processing of the juxtaposition of two units should leave a trace in memory that involves more than the mere union of the two units. As Langacker (2000, 4) formulates it: "When motor routines [i.c., linguistic units, BB] are chained together into a complex action, their coordination entails that no component routine is manifested in precisely the form it would have in isolation". This trace can then form the starting point of the further entrenchment of the juxtaposition as a linguistic unit. I believe

[6] From which they conclude that this goes against the usage-based view, which in my opinion is an unwarranted conclusion: the amount of experience with language shapes the representations, and it is therefore completely expected that a 6-year old with limited exposure to language will go through a two-word phase like an 18-month old does. Again, it is simply a matter of cumulative complexity.

this pathway to be crucial in language acquisition, and, as we will see in later chapters, it will play a central role in the model I develop.

Simultaneity

As I argued in the first paragraph of this section, different mechanisms for processing and thereby learning have to be available throughout development, unless we have strong evidence to the contrary. This means that we can assume that cognitive operations for (among other things) identification, composition by juxtaposition and superimposition, and schematization are 'waiting to process relevant input' from the start. From this, cognitive cumulative complexity naturally follows. In a first stage, some lexical units are extracted using naïve associative mechanisms. The other mechanisms are available, but as there is nothing to apply them to, they remain unused. In a subsequent stage, the lexical units are both broken down by analogical reasoning and juxtaposed, thereby leaving a trace of the juxtaposed units. Third, something like schematization can only operate on structures that are already partially blame-assigned, that is: the results of wholes-to-part or parts-to-whole learning. With those schemas, finally, new lexical units can be bootstrapped by extending a constituent of a schema to fit a phonological structure not seen before.

Thus, I expect that, given a set of available learning mechanisms and operations on use, the frequencies of these mechanisms and operations will vary over time in such a way that the learner becomes more and more reliant on the knowledge of the language in trying to find out what the unknown parts are. Although this is much in line with Hollich et al.'s (2000) take on word learning, I do not believe it requires its own 'model'[7] as it follows from an interacting set of operations that take each other's output as their input.

2.3 Desiderata for a usage-based model of language acquisition

If we want to develop a computational model of language acquisition from a usage-based perspective, what are the central ideas from the usage-based theory of language acquistion that we want to see instantiated in such a model? McCauley & Christiansen (2014*b*), Beekhuizen, Bod & Zuidema (2013), and Beekhuizen, Bod & Verhagen (2014) discuss several desiderata for usage-based models of language development, that, together with the previous discussion of the theory, constitute the starting point for this section.

[7] Except for those cases where maturation clearly plays a role, such as the development of the Theory of Mind (see for instance de Villiers & de Villiers 2000).

2.3.1 D1: Explicitness

McCauley & Christiansen argue that models should make their simplifying assumptions clear and explicit, and motivate them with developmental data. Their desideratum states that these assumptions should not only be explicit, but also grounded in our knowledge of what is available to the child cognitively and perceptually. Studying meaning thus involves obtaining naturalistic input of available conceptualizations of the learner (a topic to be discussed in chapter 4). I formulate this desideratum as follows:

D1 *The simplifying assumptions of a computational model should be clear and explicit.*

The explicitness of simplifying assumptions is a general point that holds for any model. Both in the mechanisms and representations, any model of language acquisition makes simplifying assumptions (after all, it is a *model* of something else).

2.3.2 D2: Comprehensiveness

McCauley & Christiansen argue furthermore that, if language use is held to be central, a working model of language acquisition should be able to model the processes of language use, in both production and comprehension. I believe this desideratum can be made unconditional from the assumption that language use is central. Even for a theory in which the use of the linguistic system is regarded as both logically and ontologically distinct from the representational system, it has to be shown how the linguistic system interacts with processes of use such that it can account for linguistic comprehension and production. Usage processes form a bridging hypothesis between the representational theory and linguistic behavior in that case, but one that has to be shown to work if we want to link the representational system to linguistic behavior.[8] This leads to the formulation of a second general desideratum:

D2 *The model should be able to produce an utterance on the basis of a conceptualization and a conceptualization on the basis of an utterance.*

2.3.3 D3: Simultaneity

In the previous section, I argued that with a usage-based conception, the same set of mechanisms should be able to account for the acquisition of both lexical and grammatical constructions. Ideally, a usage-based model of language acquisition performs both tasks at the same time. More specifically, it should

[8]The important question of what is considered to be data arises here. Although a fuller treatment of this issue is outside the scope of the present work, I believe that even with a modular perspective on the linguistic system, any cognitive operation on the system should be regarded as behavioral (including introspection), and therefore in want of an auxiliary bridging hypothesis.

not be the case that the model has to await the formation of a set of lexical constructions in order to start building up more abstract grammatical constructions.

D3 *A model should have the mechanisms to learn both lexical and grammatical constructions at the same time.*

2.3.4 D4: Cognitive realism in representations

Another constraint discussed by McCauley & Christiansen (2014*b*) is that computational models should reflect realistic conceptions of how (we think that) the mind works. For usage-based linguists, this is not only a constraint on computational models, but on all linguistic work, known as the 'cognitive commitment' (Lakoff 1990). We can, somewhat artificially, separate the idea cognitive realism into a set of desiderata on the representations, a set of desiderata on processing, and a set of desiderata on ontogenetic development. I will discuss these in the sections 2.3.4-2.3.6.

Concerning representational realism we can formulate the following, wide, desideratum:

D4 *The model should adhere to psychologically plausible constraints on representation.*

A consequence of language use being central in the usage-based account is that the representations should reflect properties of language use (which is why the separation is somewhat artificial). Qualitatively, this means that the contents of the representations should be derived from the usage events. That is: they should contain only phonological and conceptual structure, and distributional knowledge about these, as long as it is built up in a bottom-up way. On the quantitative side, grounding the representations in the usage events means that the grammar should "encode best what people do most" (cf. Du Bois 1987) and that the representational strength of the various representations should reflect the frequency of use of the representations (cf. the notion of the grammar as 'probabilistic' in Beekhuizen, Bod & Zuidema 2013).This brings us to the next two desiderata:

D4-1 *The content of the representations the model employs should contain only aspects of the usage events from which they are learned.*

D4-2 *The representations should reflect the frequency of their use.*

Following Langacker's (1988) notion of immanence, the abstract representations of the model should not be ontologically distinct from the representations they were derived from. That is: they should not be seperate entities in our conception of their cognitive status. I do not take this to mean that, for instance, the abstractions cannot be represented in the computational model

separately from the constructions they were derived from (cf. the discussion in section 2.3.7). However, the content of the abstractions should reflect the content of the more concrete representations in which they are immanent.

D4-3 *The more abstract representations of a model should be immanent in the more concrete representations.*

Importantly, the concept of immanence implies that abstractions are not abstract-ed, meaning that they are not novel entities created from other entities. If we, nonetheless, represent grammar, for explanatory purposes, as a set of discrete structures (where abstractions are separate members of this set, or nodes in the network), it is inevitable that the grammar will display redundancy (cf. Beekhuizen, Bod & Zuidema 2013), as all possible overlaps between the usage events is explicitly represented. These overlapping patterns resemble each other to a large extent, and the resulting grammar can thus be considered to be redundant. However, this is only an effect of the reification of abstraction, which also, reversely, means that the often-made a priori argument for the parsimony of storage need not bother us here, as the redundancy exists on a linguistic-analytical, rather than a cognitive-ontological level.

2.3.5 D5: Cognitive realism in processes

The counterpart to the desiderata concerning representational realism are the desiderata concerning processing realism:

D5 *The model should adhere to psychologically plausible constraints on processes of comprehension and production.*

Whereas most computational models assume a single combination operator, we want to allow our models to be less restricted. Especially a simple mechanism like concatenation should be part of the models potential. This creates a less parsimonious explanation on the theoretical level, but mechanisms beyond combination are necessary to get grammar learning of the ground (as I will argue in chapter 3). Furthermore, even if we label them differently, the difference between a concatenation and a composition is not that big: both create graphical objects in which the meaning structures are unified. In concatenation, this novel object has no semantic top node (both concatenated objects are hierarchically equal), whereas in composition, the novel object has a semantic top node.

D5-1 *In language use, the model should be able to employ a variety of structure-building mechanisms, ideally involving slot-filling, concatenation, and proper superimposition.*

Furthermore, language processing does not involve a query for the analysis that is optimized over the whole utterance that is being processed. Whereas

many computational linguistic approaches conceive of the task of parsing an utterance as finding the best analysis, a cognitive model has to be more constrained. We find evidence that processing takes place linearly and without utterance-wide optimization in the experiments done by Ferreira & Patson (2007). So-called garden-path sentences, of the type *While Mary bathed the baby played in the crib*, involve an initial misinterpretation of an utterance (with *bathed* being interpreted as a transitive verb and *the baby* as its direct object), exactly because a language user does not keep track of all possible analyses and is processing the utterance linearly. We find a similar take in Langacker's (1988) notion of activated unit, where only a single analysis is arrived at (cf. section 2.1.2). Desideratum D5-2 can be formulated as follows:

D5-2 *In language use, the model should not perform utterance-wide optimization, but arrive at an analysis while linearly processing the utterance, keeping track of only the most likely analyses.*

2.3.6 D6: Cognitive realism in ontogeny

We do not only want a model to be realistic at the time scale of the processing of utterances, but also at the time scale of ontogenetic development. Generally stated:

D6 *The model should adhere to psychologically plausible constraints on ontogenetic processes.*

A first constraint on development comes from Brown's (1973) law of cumulative complexity. We do not want a model to allow for more complex representations before simpler ones are acquired, and we want it to find its evidence in its set of simpler representations for a more complex one. This holds for both syntagmatic and paradigmatic aspects of the representations (i.e., shorter constructions should precede longer ones, and more concrete constructions should precede more abstract ones). The desideratum can be formulated as follows:

D6-1 *A model should not allow for novel representations of greater complexity (abstraction, length) than it has evidence for given its then current representations.*

If we furthermore conceive of language learning as a blind effect or trace of processing, learning operations should not constitute a separate process in the theory. This does not mean that the processes are done in the computational model with the same mechanism: there may be a methodological or analytical separation of learning and processing, as long as it can be interpreted as ontologically reflecting a unified process. Foreshadowing desideratum D7, the separation may in fact provide us with more insight in the exact nature of what learning involves.

Crucially, the learning mechanisms should not involve any decision-making processes after an exemplar has been processed. This would, after all, constitute a case of learning being ontologically distinct from language use. Of course, future evidence may point out that there is something akin to reorganization going on in language learning, but this is, in the first place, a less parsimonious hypothesis, as it involves reinforcement *and* reorganization instead of only reinforcement, and secondly less coherent with the current usage-based conception of language learning. Until such evidence is presented, we can state the following as a desideratum:

D6-2 *The 'learning' of a model from an exemplar should not involve a decision-making process between what is learned and what is not learned but rather be a blind effect or trace of the processing of that exemplar.*

Thirdly, I argued how, despite the emphasis of whole-to-parts learning, learning from parts-to-wholes should also be expected to play an important role in the acquisition of grammatical constructions. Although the emphasis on whole-to-parts learning is historically understandable, a usage-based theory of language acquisition should involve both types of learning, and so should a usage-based computational model:

D6-3 *A model should allow for both parts-to-whole and wholes-to-part learning.*

Finally, the roles of the various mechanisms involved may shift over time, but they should be available to the learner throughout. Concatenation, for instance, may be useful for the early language learner and then be hardly of any use later on. Nevertheless, we would not want to say that the potential for using concatenative operators goes away. Rather, the demand for the operator in usage decreases, as the learner has ways of building up structure that allow for more semantic integration, and thus more interpretability of the utterance. Nonetheless, it is crucial that the mechanism itself remains available:

D6-4 *A model should adhere to the idea of developmental continuity.*

This desideratum foreshadows an important theoretical issue that will come back in the discussion of the results of the model I will present in the next chapter, namely that, even in a usage-based approach, it is important to distinguish between a (usage-based notion of) linguistic competence and linguistic performance.

2.3.7 D7: Explanation

Although I fully agree with the spirit of the endeavor to ground computational models in our conception of cognition, pushing the quest for cognitive realism can conflict with the explanatory power of a model. To take an example: in line with Langacker (1988), we can assume abstractions to be immanent in the

more concrete representations that instantiate them. Thus, we take the immanence of abstractions to be a psychological constraint. We can even model this immanent potential without the abstractions being reified in the model. This happens in so-called lazy learners: models that use analogical reasoning (and thereby some form of abstraction) on the fly to generalize from an exemplar to a novel, unseen exemplar (Daelemans & Van den Bosch 2005). However, there is also some analytical insight gained by making the abstractions explicit and discretely separated from the more concrete units, namely that doing so allows us to see explicitly what kind of potential for abstraction the model has at some point in time. That is: implementing abstractions as separate entities gives us an analytical handle on the internal states of the model. Skousen's (1989) Analogical Modeling does exactly this for categorization: abstractions are reified as nodes in a network of feature combinations, but the behavior of the model consists of analogical reasoning over exemplars (thus assuming abstraction not to be ontologically real). Crucially, we can glean easily what happens in the model if we give it an input item, and, because of the explicitness of the abstraction, we can use the model to investigate the level of abstraction needed for optimal (linguistic) categorization behavior (cf. Beekhuizen 2010).

The general message is this. In a computational model, we may methodologically and analytically separate what we believe, ontologically, to be a single thing. The reason to do so, is that we may inspect properties of the model that are otherwise harder to glean from the learned representations. If everything is latently present, as often in connectionist models (e.g. McClelland & Kawamoto 1986), but also in analogy-driven lazy learners such as Daelemans & Van den Bosch (2005), the interpretive step between the model and the linguistic or cognitive-scientific conception of how language (i.c., abstraction) works becomes hard because of the size and massively interactional nature of the neural network. It is exactly the linguistic or cognitive-scientific conception and its bridging hypotheses to the model that provide the researcher with an intersubjective method of explaining the data.[9]

Minimizing the interpretive step between the computational model and the theoretical conception it is argued to instantiate, constitutes the sixth desideratum:

D7 *The interpretive step between the computational model and the theoretical conception it instantiates should be minimal and maximally intersubjective*

An aspect of usage-based theorizing on which progress could be made is the unification of mechanisms hypothesized to be involved in the process

[9]This does not mean that computational models that are harder or even impossible to interpret are of no value; if a computational model that is hard to interpret turns out to predict a developmental phenomenon really well, there must be 'something to it'. However, unless we arrive at a deeper level of understanding of the phenomena through the explicit connection with a comprehensive theoretical conception, a model that is hard to interpret remains a mere promissory note.

of language use. Especially in the explanations of behavioral patterns in language acquisition, we find many mechanisms that are invoked to explain certain phenomena. Ambridge, Pine & Rowland (2012), for instance, in their study of the overgeneralization of verb-argument structure, explain the phenomena they find with statistical pre-emption, the entrenchment of patterns, and the semantic fit of the verbs in the argument-structure patterns (see section 2.4.3 for a fuller discussion). The linking of model and theory discussed in desideratum D7 may involve linking a single aspect (representation type, mechanism) of the model with a number of aspects (representation types, mechanisms) of the theoretical conception. This is a desirable feature, as we simplify our conception.

At the same time, the *methodological* virtue of searching for unifying explanations should not stop us from proposing a multitude of mechanisms that we assume to be at work in language. If language phylogenetically emerged as a cultural phenomenon that employs all sorts of pre-existing cognitive mechanisms, as Tomasello (2003, 2008) argues, it is well conceivable that there is not a single, overarching mechanism doing all the work is involved. The metaphor Gigerenzer & Brighton (2009) employ is that of the use of old tools for all sorts of purposes, many of which these tools were not intended for (or 'selected for' in evolutionary terms). Human cognition, and the cognition underlying language and other (presumably) phylogenetically recent phenomena can be expected to involve the use of a number of old tools for new problems.[10] As a final sub-desideratum to D7, we can formulate these ideas as follows:

D7-1 *The more aspects of a theoretical conception can be linked to a single aspect of the model, the better.*

2.4 Core developmental phenomena

Within four years, the language-learning child moves from saying nothing to producing utterances very similar to those produced by adults. A viable theory of language acquisition not only accounts for a possible way of arriving at adult-like behavior, but also for the various waypoints, i.e., the linguistic phenomena typical for linguistic development over developmental time. Instead of looking at detailed case-studies, I take three phenomena that apply across the board to be crucial explananda of a theory of language acquisition and discuss what the usage-based account has to say about them.

2.4.1 The abstractness of early representations

A central question in the study of language acquisition is how abstract the representations underlying children's early productions are. As this is a ques-

[10] Inversely, as researchers like Kirby (1999) have it, the nature of these old tools does shape the range of solutions to the new problem.

tion about the representations, which cannot be directly observed, one must reason from the utterances a child produces or her behavior elicited in experimental set-ups to the most likely level of abstraction in the representational potential of the child.

Limited scope grammars

Braine (1963, 1976) analyzed children's early productions in terms of combinations of a fixed element, a pivot, and a variable, or 'open' element and argued that most of young children's productions can be understood in terms of pivot schemas. Examples of pivot schemas are X + *gone* (*ball gone*, *daddy gone*) and *more* + X (*more juice*, *more play*). Braine arrived at this conception by a counting method over corpora of children's productions. The categories employed in the pivot schemas are, importantly, not the same as those an adult language user uses, Braine argues, as evidenced by errors such as *more outside*, where the child allows for elements to be combined with *more* in ways an adult would not.

Regarding abstraction, Braine's account thus is more specific and more abstract at the same time. As anything can fit the open position of a pivot-open schema, this position underspecifies the constraints on combinatoriality found in adults. The pivots, on the other hand, are more specific than adults; they do not form a paradigm of interchangeable items with each other on Braine's account.

Importantly, the pivot-schema conception is not so much a cognitive account of learning, but rather a method of reasoning from behavior to a hypothesized cognitive state. This is by itself not a problem, but because of this, Braine provides no account of the developmental course of abstraction in the schemas: questions like 'how does the learner generalize over various pivots?' and 'how are longer utterances built using the pivot schemas?' remain unanswered.

The semantic grammar approach

A similar idea about the nature of early representations was put forward by Schlesinger (1971). On his account, children will initially learn from the linguistic and situational input simple 'realization rules' such as [[ACTION] [OBJECT]], which underly early productions like *grab ball* and *want cookie*. The content of these linguistic representations is thus purely semantic.

Schlesinger's account relies on the assumption that children initially take the semantic realization rules to consist of prototypical event-structures. In that sense, the early representations are of a highly abstract nature. This level of abstractness of the semantic content of early representations has been challenged from a usage-based perspective. In Tomasello's (1992) analysis of early grammatical productions, he shows that notions like AGENT are not applied across all verbs at the same time but rather emerge over the course of linguis-

tic development. The semantic relations between a verb and its arguments, on Tomasello's account, are initially formulated in highly action-specific ways, so that that AGENT of a hitting action is a HITTER. Only over time do children come to understand that HITTERs and KICKERs belong to the same superordinate category of AGENTs. In Tomasello's line of reasoning, if a child were to understand a general notion like AGENT very early on, she would display more grammatical productivity in her behavior, generalizing the notion of AGENT across all ACTION-predicates. Furthermore, as Ambridge & Lieven (2011, 202) argue, children do not seem to give preference to the prototypical transitive schemas with AGENTs ACTing on PATIENTs over less prototypical ones.

This argument is furthermore supported by cross-linguistic analyses of semantic roles. Languages vary in where they draw the boundary in the expression of different semantic roles, thus casting doubt on the usefulness of abstract universal primitive categories of conceptualization such as AGENT in general (Bowerman 1990). As the semantic conflation of different micro-roles (HITTERs,KICKERs, and others) under one formal marker (ordering position or case ending) is language-specific, the child has to be open to different conflation patterns (possibly with universal biases in them, cf. Gentner & Bowerman 2009). Having universal abstract semantic roles thus creates a linking problem, rather than a solution to the bootstrapping problem, as the learner will have to link the observed conflation pattern to a prior abstract semantic role (cf. Beekhuizen, Bod & Verhagen 2014, fn. 1).

The early abstraction account

Pinker's problem with Braine's limited-scope accounts is that it may cover the data well, but that, at the same time, there are "ambiguous gaps in the space of possibilities in a corpus" (Pinker 1984, 140), that is to say: instantiations of rules that are not found in a corpus can be either absent from the learner's inventory of linguistic knowledge *or* be present, but not produced *in that sample* for other reasons. Pinker therefore concludes that the child's productions do not contradict an account on which they are generated by a grammar that is as abstract as the adult's, and favors this account as it allows for more (quantitative) continuity between the child state and the adult state.

Moreover, the child's representational state may be more abstract than the adult's on Pinker's account: given the underspecified innate rules for bootstrapping the syntactic categories from the semantic information, the child may have formed categories initially that are more abstract than the adult's, to be constrained later with narrow rules specific to the target language that govern the exact distributions in a more specific fashion (Pinker 1989). Pinker's account is further discussed in section 2.4.3.

Valian (2009) states the 'more abstract' position most clearly, arguing that "the child's set of theoretical categories does not differ form the adult's in kind, only in degree: infants' categories are underspecified phonetically, mor-

phologically, and syntactically." Valian argues that from the very beginning, children have access to the categories specified by a universal grammar. Studying the acquisition of determiners, she cites the fact that children hardly make any errors in determiner placement (e.g., the fact that children never produce **red the book* instead of *the red book*) as evidence for the view that they have an adult-like syntactic representation of determiners and determiner phrases (DPs), including their combinatorial properties. Furthermore, the fact that determiners facilitate the recognition of nouns long before children use determiners productively in their own language (around 0;11, see papers cited in Valian (2009)) is taken as evidence that they have a syntactic category from the onset of (observable) linguistic development.

Conservatism and lexical specificity

Contrary to Pinker's account, usage-based approaches such as Tomasello (2003) argue that early representations are more concrete than later ones, and may develop in relative isolation from each other early on, only to be linked later on in ontogeny. Proponents of this view defend this position with the analysis that early on in production, lexical items are used in a more restricted way than adults would use them. Tomasello (1992) crucially argues that early utterances are structured around verbs, with the argument roles they project being verb-specific, both semantically and distributionally, which he calls the verb-island hypothesis.

Through processes of analogical reasoning, the verb-specific restrictions become weaker over time, and general argument-structure patterns emerge. For the argument structure constructions, this means that verbs are combined with increasingly many argument-structure constructions over development, as noted by Tomasello (1992, 241). Similarly, the restrictions on the combinatoriality of the verb-island patterns (what elements can fit their slots) become weaker over development as well.

As Tomasello (1992) only studied one child, McClure, Pine & Lieven (2006) tested Tomasello's hypotheses concerning the item-based, lexically-specific nature of early representations against a larger corpus of children. Similarly, Theakston, Maslen, Lieven & Tomasello (2012) analyzed a longitudinal densely sampled corpus of child speech from a single child. These hypotheses, and the outcomes of the two corpus studies were the following. First, few verbs will first appear in multi-argument structures. Most verbs start out in one-argument constructions. This is, of course, also due to the kinds of verbs these children are learning: verbs that only occur in intransitive patterns will simply not occur with multiple arguments. However, in both studies it was found that also for the group of transitive-only verbs, the early cases occurred more in single-argument patterns than the later ones. Secondly, utterances with verbs that were learned early will later on be found in more complex structures than 'newer' verbs. Finally, it was found that utterances with newly learned verbs are generally as complex earlier in development as

utterances with newly learned verbs later in development (i.e., rather simple), in line with Tomasello's (1992) idea that argument structure constructions develop in a verb-specific manner. This is to say that, although there are representations licensing highly general SVO patterns in the children's utterances, they somehow do not apply massively to newly learned verbs. Both studies, however, did also find newly learned verbs that were directly used in more complex argument-structure patterns later in development. They explain this with reference to the idea that argument-frame constructions such as [[SPEAKER / *I*] [ACTION] [OBJECT / *it*]] might also play a role at this stage (cf. Dodson & Tomasello 1998). Another more general developmental observation, found first by Tomasello (1992, 233-234), and established later with the so-called 'traceback method' in Lieven, Behrens, Speares & Tomasello (2003), Lieven et al. (2009) and Bannard et al. (2009), is that children's novel productions can in many cases be explained as minimally different reproductions of earlier productions: often only a single substitution is necessary to make a novel utterance identical to a previously uttered one (e.g., *you give me book* on the basis of an earlier production *you give me ball*).

Centrally, these findings point to a low level of generalization early on, which becomes higher as the inventory of verb-specific patterns grows. This suggests that the representations gradually become more abstract over time in a lexically specific way, rather than through generalizations across the board.

Most studies cited here describe the patterns on an observational level, reasoning from the child's behavior to likely cognitive representations in the child and providing only a rough account of the mechanisms operating on the input such that these patterns emerge. Important questions are what the representations underlying these behavioral patterns are and how they develop. The various studies seem to argue for a 'what you see is what you get' approach: if the child produces a VO pattern, there is no SVO representation underlying it – the differences between the early productions of VO versus contemporaneous SVO patterns support this position. This does not mean that the child is not conceptualizing the full event she wants to express, only that the child, at that point in development, finds the more restricted representation optimal. This can be because the child does not have a general enough more complex representation to produce the full pattern, or because the shorter representation is more entrenched than the fuller one.

In defense of the early-abstraction view

Recent usage-based approaches claim that children's spontaneous and elicited production provides evidence for a linguistic system that is initially organized around specific linguistic items, and that only gradually becomes more abstract. In response to this claim, several researchers have defended the early-abstraction view, mainly by criticizing aspects of the method. The two main lines of criticism on the usage-based view are foreshadowed by Pinker's (1984) critique of earlier work characterizing the grammatical knowledge of children

in terms of pivot schemas (Braine 1976). As Pinker (1984, 100; 127-133; 142-143) argues, there are two main reasons why we find less productivity in samples of early linguistic development. The first concerns the fact that the lack of productivity is exaggerated, either because the sample is too narrow, or because we would expect the amount of productivity to be that low on statistical grounds. The second concerns the idea that language-external factors (such as memory constraints, linear biases, and the limited set of referents) may be the reason why only a subspace of the grammatical possibilities is used.

An early representative of this response to the usage-based line of reasoning is Fisher (2002). Without denying the role item-based constructions play in early language acquisition, Fisher argues that there is also evidence that abstract representations exist from early on. Fisher argues that the apparent limited generalizability of argument-structure patterns may be due to the gradual development of the lexical representations. It thus is not a matter of *grammar* but of the *lexicon*. Furthermore, Fisher argues, language-external reasons such as processing may play a role. Finally, there is the methodological issue that we do not know how children interpret the meanings of verbs presented in isolation in experiments, and the subsequent lack of willingness to extend these verbs can thus be the result of many things over which an experimenter has little control.

In response to Fisher's line of reasoning, Abbot-Smith, Lieven & Tomasello (2008) argue that the limited abstraction that emerges is not just an artefact of full competence plus memory limitations or developing lexical preferences. They provide evidence for this on the basis of a cross-linguistic elicited production study. Importantly, the cues for semantic roles in German are stronger than in English (word order and case in the former, and only word order in the latter). In a repetition task, Abbot-Smith et al. found that German children at age 2;0 more often corrected a novel verb presented with mismatching cues (wrong case and/or wrong word order) to the prototypical constellation of cues than the English children. Given that the performance constraints (assumed to inhibit the child from correcting the grammatical error) are expected to be identical for German and English children, the difference must be, according to Abbot-Smith et al. (2008) attributed to differences in the representations of German and English children, where the German representations are stronger than the English ones. Abbot-Smith et al. (p. 50) conclude that "representations are graded in strength, with only strong representations allowing clean signaling to other parts of the cognitive system."

Yang (2011) makes the important point that we have to consider what the most sensible baseline pattern of expectation is. Usage-based approaches such as Lieven, Pine & Baldwin (1997) argue that children's knowledge of the language is item-specific, as the overlap of different members of a paradigm to the items they combine with is low. Lieven et al. argue, for instance, that children's knowledge of determiner-noun combinations is item-specific initially, given that most nouns occur only with one determiner (from the set *a* and *the*). Given that many linguistic items are distributed in their frequencies in

adult language proportional to the inverse of their frequency ranks (the most frequent item occurring twice as often as the second, and thrice as often as the third; known as Zipf's law after Zipf (1935)), Yang argues that this fact is not unexpected, and thus not contradicting a view on which children operate with fully abstract rules.

Another important work in this respect is Naigles, Hoff & Vear (2009), who had caregivers keep track of the first ten instances of 34 selected common verbs their child produced. The reason for choosing a diary study is that many sampling methods where the researchers records the child's spontaneous production at regular intervals, have a situational bias. That is to say: they typically record only utterances produced in one or a few situational settings (free play, eating), whereas the interactional world of the caregiver and the child extends far beyond these.

Using this method, Naigles et al. (ch. 5) found that the syntactic flexibility of children went beyond that reported in many usage-based studies (Tomasello 1992, McClure et al. 2006, Theakston et al. 2012). In their sample, the 8 children produced on average 66% of the verbs in more than one syntactic frame within their first ten instances of use. Most changes were due to the addition or deletion of a single noun. Although this latter finding is in line with results from the traceback method discussed in section 2.4.1, Naigles and colleagues disagree with Lieven and colleagues on the interpretation: given that a child only produces very short utterances anyway, it seems that changing a single word is almost the only opportunity for a child to produce a different syntactic frame.

Analyzing these results, Naigles et al. found that the difference with Tomasello's and Lieven et al.'s results was to a large extent due to the differences in coding decisions. Whereas Tomasello (p. 40) counted all one-argument structures (whether that argument occurs pre-verbal or post-verbal) as instances of one structure, Naigles et al. counted them as two (one for patterns with pre-verbal arguments and one for patterns with post-verbal arguments). Furthermore, Naigles et al. counted as different syntactic frames not only different argument-structure patterns, but also the presence of negation and the morphological marking of the progressive (*-ing*).

A final interesting finding in Naigles et al.'s (2009) study was that children would display more syntactic flexibility in the number of different syntactic frames they select per verb, than lexical flexibility, in the number of different arguments they select per verb. Naigles and colleagues interpret this as meaning that the child's production is not syntactically limited, and that children are 'avid generalizers'.

Discussion

Deciding at what level of abstraction young children operate is not a trivial matter. A lot seems to depend on the method of counting, the sample size, and the baseline expectations. As these are complex methodological issues, it

seems that at this point, we can only await the outcomes of the ongoing discussions. That is: the level of abstraction of early representations (and the nature of these abstractions – as weak schemas or something different) is an open issue and cannot form an empirical constraint on a usage-based computational model.

There is one empirical phenomenon that seems unchallenged, namely that simpler structures typically precede more complex structures, up until the developmental point where learners have pronoun frames or otherwise more abstract representations that allow novel verbs to be used in more complex representations from the moment they are learned.

2.4.2 Argument omission in early production

In children's Stage I productions, we often find lexical material not being expressed, despite the language not allowing such 'omissions'.[11] We find productions such as *put up dere* (Adam, 2;3) or *I put truck* (Adam, 2;4), where arguments (subjects, objects, locatives) are omitted. The observation is that, over time, the child will produce more and more arguments (as we have already seen in the previous section about the abstractness of early representations), and that subjects are more often left out than other arguments.

On the explanatory side, the central question is: what causes children to do so? Are the representations that Stage-I children use different from the representations used by adults or are they the same with there being extra-linguistic reasons for these omissions? And if there are extra-linguistic reasons, how do they interact with the child's linguistic knowledge at that point in time?

Predicting the omissions

Starting from the perspective that there is strong continuity between the child's linguistic knowledge and the adult's, the answer to these questions has often been that children have a fuller understanding of the grammar, but that there are extra-linguistic factors that cause the child to not produce certain linguistically obligatory material. Before we get into the discussion of the relation between representations and extra-linguistic factors, let us first have a look at factors known to influence argument omission.

Bloom et al. (1975) found that in early productions, the omission of arguments is not random, but rather follows a systematic pattern. In a corpus study of four children (ages 1;10 - 2;30, four or five sample moments each, discarding imperatives and incomplete utterances), Bloom et al. looked at four categories of verb relations, given in table 2.1 below.

A basic finding was that children started producing utterances with fewer of the main constituents (i.e., the verb and its obligatory dependents) before

[11] The term 'omission' should be understood here at a descriptive level; whether children actually 'omit' something that is present in their linguistic representation is exactly the issue under scrutiny.

verb relation	main constituents (ordered)	example
actions	agent, verb, object	*You open it*
agent-locatives	agent, verb, object, location	*You put it on the table*
mover-locatives	mover, verb, location	*I sit in the chair*
patient-locatives	object, verb, location	*The ball goes in the box*

Table 2.1: The four verb relation categories studied in Bloom et al. (1975).

producing ones with more constituents for each category, gradually moving towards the complete expression of the main constituents. There was furthermore a pattern in the constituents that were produced: agents and movers were most frequently left unexpressed. Moreover, and this being the central point of their study, Bloom et al. compared certain properties of utterances with two constituents (from the ones in table 2.1) with those of utterances with three constituents (henceforth: 2Cs and 3Cs) produced in the same sample, and found that the number of arguments expressed covaried with several of those properties.

On many measures, 2Cs were grammatically more complex than the children's contemporaneous 3Cs, for instance containing more modified arguments (with the word *another*, an attributive adjective or a possessive) and more cases of negation. Interestingly, on other measures of complexity, 2Cs were as complex as contemporaneous 3Cs. This mainly concerns morphemes with less semantic content than the ones named above, such as inflectional forms for nouns and verbs, the presence of determiners.

On a lexical level, 2Cs are also more complex than contemporaneous 3Cs. A larger number of verbs occurred in 2Cs compared to contemporaneous 3Cs. 2Cs furthermore attracted more new verbs (verbs not used by that child in the previous sample) than 3Cs (cf. the findings of McClure et al. 2006, Theakston et al. 2012). Furthermore, Bloom et al. reported that they found in an earlier study that two of the children preferred pronoun forms in early productions whereas the other two preferred nouns. The preferred argument form was found to covary with the presence of arguments as well: for three out of the four children, 2Cs had significantly more dispreferred arguments (nouns if the child preferred pronouns, and pronouns if the child preferred nouns) than 3Cs. Bloom et al. speculatively attribute these covariations to memory limitations.

On a discourse-pragmatic level, we find covariation as well. Graf, Theakston, Lieven & Tomasello (2015) studied the effects of certain discourse properties on the elements that were omitted in the productions of young children (age 3;2 to 4;2) with an elicitation study in which children were exposed to different discourse-pragmatic conditions of contrast. A pragmatic focus on con-

trast (one action versus another, or one object versus another) predicted the selective realization of linguistic elements well: when contrasting one element in a transitive utterance and keeping the other parts identical (e.g., with a situation in which one puppet acts on an object, and another in which another putter acts in the same way on the same object), children will produce the contrastive element when asked to describe a scene, but are likely to leave out the 'given' or non-contrasting linguistic material.

An interesting piece of evidence comes from a study by Berk & Lillo-Martin (2012), who looked at two normally-intelligent, deaf children who were only exposed to a language accessible to them (American Sign Language) at around age 6. If general performance constraints are the sole cause of the two-word stage around age 2, it is not expected that otherwise normally-developed six-year-olds display such linguistic behavior. Nonetheless, Berk & Lillo-Martin found that the two children they studied did follow a developmental trajectory similar to two-year-olds. From this, they concluded that general performance constraints cannot be the driving factor behind the existence of a two-word phase. Similarly, vocabulary size and biological maturation cannot be the factors, as these children had a larger vocabulary than expected for a Stage-I child, and they should have passed the state of maturation held responsible for the two-word phase. The two children did use short utterances but did so with a wide range of semantic relations (ones one would not expect in 2-year-olds) and lexemes. Berk & Lillo-Martin therefore conclude that the two-word stage as we see it in young children consists of two components: a linguistically-specific one and a general-cognitive one. For the two children they studied, only the linguistically-specific one applied, as they were otherwise cognitively similar to their age peers.

Accounts of the limited length of early production

Studies looking at the factors leading to omission are insightful, but provide us, in principle, only with relatively loose constraints on the kinds of representations children have. In his discussion of ideas about the representations underlying the truncated utterances, Pinker (1984) discerns four classes of hypotheses that explain why children produce such utterances:

- The deletion-rules account (e.g., Bloom 1970) states that children have fuller linguistic representations, but that there are deletion operations that cause certain elements not to be expressed.
- Under the incomplete-rules account (e.g., Braine 1976), children simply have rules that cover only the material that is expressed. That is: the representations do not go beyond what is expressed.
- The optional-rules account (e.g., Bloom et al. 1975) holds that children have fuller linguistic representations, as in the deletion-rules account,

but that instead of deleting elements of the rule, these elements are simply optional, where the likelihood of producing them is given by a number of interacting extra-linguistic factors, such as discourse salience and complexity of the element.

- The processing-limitations account (Lebeaux & Pinker (1981) as cited by Pinker (1984)) explains early omissions with reference to general processing limitations. The child's representations are essentially identical to the adult's (as opposed to the other three accounts), but other, interacting, cognitive mechanisms mature such that the child will eventually produce adult-like utterances.

In principle, all four can be argued to be in accordance with the findings of covariation discussed before. Under the deletion-rules and optional-rule accounts, deletions or non-production (respectively) would be triggered by the factors described, whereas in the processing limitation account, the filtering of the mapping from conceptual structure to phrase structure would be driven by these factors. An incomplete-rules account would argue that the selection of certain incomplete rules over others is brought about by these factors. As such, identifying factors behind omission does not discriminate between the four conceptions, although some accounts would need additional machinery to link discourse-pragmatic as well as memory constraints to the representations. Two explanations have been central in the study of length-limitations over the last thirty years, viz. the processing-limitations account and, more recently, the incomplete-rules account. As the usage-based theory has not made specific claims about either, it is worth looking at them in some detail.

The processing-limitations account

Eliminating the deletion-rules, incomplete-rules and optional-rules accounts on a priori grounds, Pinker (1984) argues that the limited length of early grammatical production can best be accounted for with general processing limitations, although he admits his account is rather speculative. According to Pinker (1984, 160ff.), the memory buffer can, in the mapping of f(unctional) to c(onstituent) structure, only process a certain amount of functional elements early on. Suppose that the child wants to express that the doll sits on the chair, but the memory buffer only allows for two functional elements to be mapped onto the constituent structure (which Pinker, on the basis of the continuity assumption, assumes to be the full, adult-like tree). The child only selects the pragmatically most salient elements to be mapped, so that only something like *sit chair* is produced. The memory constraint is relaxed over development, so that longer productions are possible, although Pinker does not go into the nature of this development.

Similarly, Boster (1997) argues that children's early utterances are constrained by a linguistic processing constraint. In her model, every (lexical or syntactic) operation has a cost, and for production, the cost of the generated

representation underlying the utterance cannot surpass a processing-limit parameter. This parameter becomes less restrictive over time or the cost of the rules that are applied and the lexical elements that are retrieved decrease. Boster (p. 17) criticizes Pinker (1984) for not explaining why so many subjects, as opposed to objects and verbs, are dropped. Her account does explain this, referring to the order in which lexical elements are merged and moved in the derivation underlying the production.

The advantage of these approaches is that they create maximal continuity in the representations used. Furthermore, this approach is in line with the findings discussed in section 2.4.2: children allow those elements of meaning to pass through the filter that are pragmatically most salient (cf. Graf et al.'s (2015) findings on the pragmatics of omission) and given a fixed memory buffer, a modified noun phrase, for example, means that there is less buffer 'left' for other elements (cf. Bloom et al.'s (1975) findings of covariation).

A conceptual downside is that a new variable that changes over time is introduced, namely the memory buffer. This creates cognitive discontinuity, albeit on a non-linguistic level. A more important problem with having a memory buffer that changes over time, is that it provides us with little of an explanation. Rather, it seems to be a redescription of what we observe, viz. limited-length utterances, in terms of a changing memory buffer.

The incomplete-rules account

The accounts of Braine (1963) and Schlesinger (1971), discussed earlier, constitute prime examples of incomplete-rules accounts. The child operates early on with incomplete rules, which it has extracted from use (possibly by mapping them onto prototypical event structures, as in Schlesinger's explanation). Schlesinger argues that the transition to longer utterances happens when the child starts to combine shorter rules. An [[AGENT] [ACTION]] rule can be combined with an [[ACTION] [OBJECT]] schema, thus producing an utterance containing an AGENT, an ACTION, and an OBJECT. Schlesinger does not provide any reasons as to *why* and *when* the child starts doing so, which makes the analysis unsatisfactory from a developmental point of view.

Although the usage-based view does not say much about early argument omission, we do find suggestions that this view is most coherent with an incomplete-rules account. In studies such as Theakston et al. (2012) and Lieven et al. (2009), we find that early on one-argument constructions used with transitive verbs can be slightly different in their selection preferences from the two-argument constructions for transitive verbs acquired later. This provides evidence that one-argument productions used with transitive verbs are not simply generated by two-argument rules. This view is in line with the description of the process of acquiring further dependents (such as arguments) of a head (i.e., a verb) given by Tomasello (1992, 234;250-253). Tomasello argues that central to the acquisition of grammatical rules (or constructions) is the widening of syntagmatic relations, with paradigmatic abstraction only being

the by-product of this (cf. previous section). If syntagmatic development operates on certain representations only (lexically-specific rules, rules with a very limited amount of abstraction), then it may be the case that a two-argument rule is not simply a 'coalescence' of earlier one-argument rules, but an independent development.

The grammatical-representation account

A third account that can be added, is the grammatical-representation account, which depends among all accounts the most on the particular representational theory of grammar it is associated with, viz. generative grammar. Several lines of research start from the assumption that argument omissions can be explained with reference to an erroneous (default) setting of a grammatical parameter. The parameters regulate aspects of the grammar at a global level, for instance whether the language allows for the omission of subjects or not or whether tense is obligatorily expressed on verbs or not. The phenomenon of subject omission has been explained with recourse to various parameters (Hyams 1986, Yang 2002, Hyams 2011).

The omission of other obligatory elements of the argument structure has received far less attention. For object-omission, Pérez-Leroux, Pirvulescu & Roberge (2007) argue that the child is figuring out how the (innate) licensing and recovery constraints for null-arguments work. These are general grammatical principles, but for every lexical item it has to be specified how they work. It is in finding out the exact properties of lexical elements that children make errors: expressing transitive events with only a subject and a verb because they assume the verb allows for recoverable or generic null-objects (as the verb *eat* in English, for instance does: *I ate all afternoon*). Note that the errors are thus not a purely grammatical matter on Pérez-Leroux and colleagues' account, but rather an interaction between the developing lexical knowledge and grammatical principles.

In short, grammatical-representational accounts explain the omission of arguments with general grammatical principles and parameters. One interesting consequence of this is that different omissions (subject, objects, prepositions) may be caused by different underlying principles.

Discussion

Why do children leave obligatory lexical elements, such as nouns and prepositions out, despite their presence in the input? Different accounts point to different factors seemingly involved in this process. Most accounts come down to an interaction between the representations and the processing capacity of the child (the exception being the grammatical-representations account). On top of this, there can be pragmatic effects explaining under what conditions which argument is omitted. This identification of factors does not provide an account of the question why (at all) children fail to produce certain arguments.

Here, the interaction between representations and the processing capacity is the most salient answer.

Interpreting Berk & Lillo-Martin's (2012) study, I think that we can reverse the line of reasoning. Instead of saying that the constraints loosen, we can say that the constraints stay the same, but the cost of retrieving rules becomes lower as a product of experience. This is not a language-specific operation, but something applying to cognition at large. Under this reversal, the findings of Berk & Lillo-Martin are congruent with a usage-based account, which focuses exactly on this role of the frequency of experience on the nature and use of the representations.

Empirically, the central explananda associated with early argument omission are the general development from fewer to more arguments, the frequent omission of subjects, and the effects of covariation between argument complexity and salience found. These constitute the first three explananda of for a usage-based theory of language acquisition.

2.4.3 Argument-structure overgeneralization in early production

In Stage I, children not only omit obligatory elements of the argument structure, they also make errors of a different kind, albeit less frequently. These errors, known as errors of commission, are cases where the child seems to apply a rule in a way an adult would not. They start showing up in children's spontaneous productions around their second birthday (Bowerman 1974, Marcotte 2005). Arguments may have been added, as in example (5), repeated here as (17), and the order of the elements may deviate from the 'correct' adult production, as in example (6)-(7), repeated here as (18)-(19).

(17) Adam fall toy (Adam 2;3, dropped a toy)

(18) eat Benny now (Ben, between 1;7 and 2;6, wants to eat)

(19) the bridge knock down (Aran 2;4, knocked the bridge down)

The received opinion in the literature is that errors like these are vanishingly rare and that grammatical acquisition is overwhelmingly error-free. I believe there are two reasons not to take this as a discouragement for considering these errors crucial explananda of a developmental theory. Firstly, rarity has never been an argument for downplaying the importance of phenomena like, say, long-distance *Wh*-movement for theorizing about the necessary complexity of the cognitive representations underlying linguistic behavior. Second, the rarity of the errors depends on the method of counting. In a study of causative-inchoative alternations (cases like examples (17) and (19)), Marcotte (2005) gathered instantiations of these kinds of errors in a large portion of the CHILDES database. Marcotte found that in contexts where an error *could* occur, children across all age categories make errors like the ones in (17) and (19) in about one out of ten and one out of fifteen cases respectively (Marcotte 2005,

56-57). These numbers furthermore suggest, as Marcotte notes, that the often stated difference in frequency between the type in (17) and that in (19) may be due to the fact that the former are more salient to the researcher doing a diary study.

Going beyond the observed properties of these utterances, several accounts of the linguistic knowledge underlying them have been developed. Two main categories of interpretations can be discerned. First, 'errors' like these may emerge from the child struggling to put all the meaningful elements together, without applying any rule or rules for the linearization of such elements. Second, children may apply their inventory of lexical and grammatical representations to construct an utterance, but there are qualitative or quantitative differences in the selectional properties of the grammatical representations children and adults use.

The groping-patterns approach

A proponent of the former position for this stage of development, is Tomasello (1992). Tomasello argues that many productions in Stage I are not governed by any sort of grammatical system. Tomasello explains the fact that the vast majority of productions is in compliance with the adult order, by arguing that the children are imitating the input on a superficial level, i.e., without understanding the function of word order. In terms of the continuity assumption, this creates a strong discontinuity between the earliest productions and the later ones, and there is no account in Tomasello (1992) how children move from a pre-grammatical to a grammatical stage and whether this transition is gradual or immediate. A further problem of this view is that there is often more regularity in the kind of errors of commission children make, suggesting an account in which structure-building rules do seem to play a role.

Groping patterns are were first discussed by Braine (1976) as patterns in which the child attempts to compose a multi-word utterance without having command over the rules for doing so. Under Braine's (1976) analysis, it crucially involves linearly stringing together known units. This happens, according to Braine, when one of the elements occurs both as the first and the second element in different two-word utterances in Stage I. An example would be the element *all gone*, which occurs both in cases like *all gone daddy* and *daddy all gone*. Braine provides no account, however, of the possible mechanisms whereby these patterns are generated. The idea of young children using groping patterns is not unique to the usage-based view or its progenitors: although Pinker (1984, ch. 4) tried to explain the descriptive facts of groping patterns from the perspective that they are generated with rules and a single adult-like combination mechanism, he cites Chomsky (1975) as conjecturing that "Stage I speech reflects a prelinguistic system akin to a fledgling's first flutterings" (Pinker 1984, 97).

Clark (2003, 165-166) essentially agrees with the position that groping patterns are not generated by adult-like rules, but adds that they may be driven

by information structure (what is given and what is new in the discourse), thus assigning slightly more structure to them than Braine would. Under this analysis, the empirical task of finding out whether an utterance is generated by a hierarchical-structure-building rule or simply strung together is more difficult. Children's early patterns may structurally follow adult-like order and at the same time be just strung together linearly, because the information structure for children will be very similar to that of adults. Clark raises another point, namely that the pragmatic salience of the events and objects may play a central role in the early Stage-I productions. She cites the example of *get-down cart*, meaning 'I want to get down in order to get my cart' as a case in point. As the two words reflect an event and an entity from a complex proposition (in her semantic analysis), it seems hard to interpret this utterance as being generated by anything like a rule. A hypothesized (semantic) rule would have to be something like [[ACTION] + [OBJECT INVOLVED IN THE PURPOSE OF THE ACTION]] which is an unlikely, but logically not an impossible rule.

Pinker's broad and narrow rules

The other interpretation of cases such as (18)-(19) is that they are not instantiations of groping patterns, but rather the application of some grammatical rule that is somehow used differently from how adults would use it. One influential account is Pinker (1989). Pinker argues, for the causative-inchoative alternation, that children operate with two kinds of rules for deciding if a verb heard in one argument-structure pattern can also occur in the other argument-structure pattern. The broad-range rules, on the one hand, provide necessary conditions for verbs to alternate between argument-structure patterns. Because they provide necessary conditions, they are rather abstract. Children can apply them early on, because they are derived from the innate linking rules between the conceptual and the formal structure. Narrow-range rules, on the other hand, are acquired in a piecemeal fashion over development. The child will gradually find out that manner-of-motion verbs can occur in both causative-transitive and inchoative, that is: manner of motion is a sufficient condition for allowing the verb to alternate. Initially, Pinker argues, children without a well-developed set of narrow-range rules will, under discourse pressure, resort to using only the broad-range rule.

Syntactic accounts

Early errors of commission have received a rather extensive treatment within the generative nativist paradigm. Even though the present work does not start from this perspective, or share its basic properties, some of the explanations prove insightful for understanding what a child is doing.

An interesting account of the object-verb errors (such as example (19)) is that of Radford (1990, 231). Radford argues that in these cases, only the object noun and the verb, but not the subject noun, are presented to the syntax. Be-

cause the verb is required to merge with a nominal phrase to provide a subject role, the object-noun erroneously ends up in the subject position. In the case of subject-omission in transitive clauses, yielding a verb-object pattern, Radford argues that the subject is presented to the syntax but dropped (presumably because of parameter settings, such as the ones discussed in the section on argument omission). Aside from the theory-internal details, this account points to a general mechanism: if the learner for some reason cannot express an argument (in Radford's terms: present it to syntax), she may have to take recourse to whatever other syntactic pattern *is* available to express something close to what she means. From the constructivist perspective taken in this research, one could argue that the object is coerced into the subject slot of the intransitive construction, because the transitive construction for some reason cannot be used. A fuller account in these terms obviously would require a specification of the conditions under which the better-fitting transitive construction cannot be used.

Cases such as example (18), where the subject appears post-verbally, have been analyzed by Deprez & Pierce (1993, 43) as misinterpretations of the subject role as an object. They argue that in 90% of cases of this error, the subject is not an agent, but a theme of the verb's semantics, and hence realized as the direct object. The fact that this error occurs structurally with theme subjects is an argument against the groping-pattern account of early errors of commission: the high proportion of theme-subjects suggests that there is more structure to this error type than if the child was just 'groping' to construct an utterance.

Another interpretation of the pattern in example (18) is given by MacWhinney (1985, 1120), who argues that children initially go through a stage of placing all new, salient, or interesting information first. This is a pragmatic principle, rather than a structural one, and so MacWhinney's account is not a syntactic account, but rather a 'groping-pattern-plus-pragmatics' one. However, if it is dominantly theme-subjects for which this error occurs, as Deprez & Pierce (1993) note, MacWhinney's account seems to predict falsely that these errors would be made with agentive subjects as well.

Usage-based accounts

In a recent series of papers, Ambridge and colleagues (Ambridge & Lieven 2011, Ambridge et al. 2012, Ambridge 2013, Ambridge, Pine, Rowland, Freudenthal & Chang 2014, Bidgood, Ambridge, Pine & Rowland 2014) look at a number of factors involved in overgeneralization of argument-structure patterns and the retreat from overgeneralization. First, statistical pre-emption, as proposed by Goldberg (1995), plays a role. Statistical pre-emption is the process whereby overgeneralizations stop being made once a more concrete, competing form is part of the grammatical inventory. Second, the entrenchment of verbs in argument-structure constructions is shown to have an effect. If a learner observes a verb fifty times with one argument-structure construction, and never with another, she can be more certain that it is unlikely that

that verb will occur in the other construction than if she would have seen it five times in one construction and never in the other. That is to say: in the former case, the verb is more entrenched in the first construction than in the latter case. Third, and similar to Pinker, children appear to become increasingly sensitive to narrow verb-classes. Ambridge and colleagues showed this by looking at novel verbs that expressed certain classes of meanings (e.g., manner of motion, sound emission) and by studying whether children accepted the generalization of the novel verb into an argument-structure construction they did not observe it in previously. Children increasingly showed sensitivity to the narrow-range rules that govern the generalizability. Because the verbs were novel, pre-emption or entrenchment could not play a role, and it must be the verb semantics that the children used to accept or reject a generalization. Two further factors of interest are named, but not further worked out in these articles, viz. the frequency of the argument-constructions per se, and the pragmatics of the situation that may make the children's use of an overgeneralization more or less likely.

Although Ambridge and colleagues focus on a later age range, these results are insightful for the study of younger children's overgeneralization behavior. The errors discussed at the outset of this section wax and (partially) wane within this very developmental period. Under an assumption of continuity of processing mechanisms, this means that pre-emption, entrenchment, verb semantics, construction frequency and pragmatics can be expected to play a role as well. With the exception of the last factor (i.e., pragmatics), all of these can be operationalized in a computational model of the kind I propose later in this dissertation. In fact, as I will show there, the first four factors prove to be only separate mechanisms on an analytical level, but all follow from the process of selecting the optimal set of constructions to express the learner's conceptual intent with.

It has been argued that different regularities in the environment become salient to the developing child at different ages (for a general account along these lines, see, e.g., Hollich et al. 2000). In the case of overgeneralizations, it has been argued that statistical distributional information (such as pre-emption and entrenchment) have an effect on overgeneralization behavior before the semantic classes do (Tomasello 2003, 180). Ambridge et al. (2014, 221-222), however, argue that in the studies operationalizing this idea experimentally, the effect of earlier sensitivity to distributional statistics over verb semantics may well be due to task effects. Again, under a continuity assumption, it seems better to assume equal sensitivity to all properties, only to be given up when there is strong evidence to the contrary.

Discussion

Findings about the frequencies of the errors, such as Marcotte's (2005), should be taken as explananda for any theory of language development and thereby for a computational model. The fact that there are strong biases in the error

patterns, as noted by Deprez & Pierce (1993), suggests that it is safe to assume that the errors should follow from the grammatical representations the computational model has at a point in time, rather than being due to the child 'groping' to construct an utterance.

From a usage-based perspective, experimental findings such as Ambridge and colleagues' provide conditions on the kinds of overgeneralizations that should be expected: entrenchment, construction frequency, and pre-emption should play a role from very early on, but the effect of semantic classes should increase over time (regardless of whether we treat these as separate mechanisms or symptoms of a unifying process). This gives us two further explananda for a theory and model: to explain why overgeneralization takes place, as well as to account for the shifting role of the various factors involved in the retreat from overgeneralization.

2.4.4 Explananda for a usage-based model of language acquisition

Not all phenomena discussed in the previous sections are equally suitable to be studied with a computational model. Especially the notion of abstraction, however interesting, is to my mind undecided, with empirical claims in favor of both an early-abstraction and a conservatism point of view, and as such I will not consider them as empirical explananda in this dissertation.

On the side of production, however, five interesting phenomena can be found that seem uncontroversial:

E1 *An increasing number of arguments is produced over developmental time.*

E2 *Subjects are omitted more often than other arguments*

E3 *The amount of arguments co-varies with the complexity of the arguments.*

E4 *Argument-structure constructions are overgeneralized at some point in development, but the learner overcomes this overgeneralization.*

E5 *The role of various reasons for overgeneralization varies over developmental time.*

Besides these broad-level phenomena, there are of course also several more detailed phenomena. The more of these can be captured with the same computational model, the better. However, it seems good to have a baseline of global phenomena to account for, so that models are not developed with a single, narrow, purpose in mind.

2.5 Computational usage-based models of language acquisition

In the past two decades, computational modeling has been increasingly applied as a method of studying the nature of grammatical development in ontogeny, with several important modeling attempts within the usage-based framework being published in the past ten years. As this thesis deals with the usage-based perspective, I will focus only on computational modeling studies focussing on the mechanisms involved in language acquisition that start from this perspective, with one exception.

Even within the usage-based tradition, there are several interesting models that I do not discuss here nonetheless, because they do not have the explicit goal of being psychologically realistic (e.g., the tradition of grammar induction, see de la Higuera (2010) and references cited there), or because they are models that try to analyse the utterances children produce in a post-hoc fashion, thereby not being full input-output models (e.g. Bannard et al. 2009, Borensztajn, Zuidema & Bod 2009) Also worth mentioning is the approach of Fluid Construction Grammar, especially van Trijp's (2008) dissertation. Although the model contains a working set of mechanisms for acquiring constructions, the focus of the approach is not on language development.

2.5.1 Semantic-grammar models

The first kind of models we discuss are models directly operationalizing the constructivist tenet that the grammar consists of form-meaning pairings. The first three models discussed below constitute the direct starting points of the model I present in chapter 3.

Chang (2008)

In her dissertation, Chang (2008) presents a model of the acquisition of grammatical constructions that is aimed to fit in with the Berkeley Neural Theory of Language (Feldman 2006). The model assumes grammatical constructions, pairings of phonological form and meaning, as its representational format.

The model processes input items (pairings of an utterance and a situational context) one at a time. For every input item, the model tries to analyze it by using its inventory of constructions. The resulting analysis is a semantic specification of the composite meaning of all constructions used, which can then be pragmatically resolved against the situational context of the input item. As the model will initially have an incomplete grammar (i.e., one that might not be able to analyze every word), the analyzer allows for incomplete analyses and cases where there are multiple partial analyses (e.g., when the analyzer just recognizes two lexical constructions but has no construction to combine them).

Often multiple analyses are possible. In those cases, the model will select the analysis from among the analyses with the lowest number of roots (or partial analyses) that has the highest probability given the grammar (reflecting how often the used constructions have been observed before) and that covers as much of the context and utterance as possible.

On the basis of this best analysis, the model updates the counts of the used constructions and hypothesizes novel constructions by reorganizing the constructicon. Chang frames this as an incremental optimization process, in which the model looks for an 'optimal' grammar. To do so, she employs Bayesian Model Merging (Stolcke 1994). Bayesian Model Merging evaluates whether a reorganization step in a model of the data (i.e., a constructicon of the language) enhances the trade-off between compactness and good coverage of the data by using the Minimum Description Length principle (Rissanen 1978). Making an abstraction over two grammatical constructions, for instance, makes the grammar more compact (the commonalities are stored only once), but also typically decreases the coverage (the model now allows for the generation of structures that have not been observed before, so that the observed ones become less likely).

The reorganization steps Chang discusses fall into two classes. The mapping operations, first, specify how a novel representation can be formed on the basis of the unanalyzed parts of the utterance and meaning. 'Simple mapping' finds new pairs of unanalyzed phonological form and meaning to map to each other as a novel construction, whereas 'relational mapping' takes the relation between several partial analyses and, given the satisfaction of some constraints, hypothesizes that relation to be a novel construction. The merging operations, on the other hand, properly reorganize existing parts of the grammar. In the case of 'join', two constructions that share part of their structure are joined into a larger whole, creating, for instance, an [[ENTITY] [ACTION] [OBJECT]] construction out of an [[ENTITY] [ACTION]] and an [[ACTION] [OBJECT]] construction. With 'split', on the other hand, a construction is split into parts on the basis of its commonalities with another construction, creating, e.g., a lexical [BUTTERFLY / *butterfly*] construction out of a [[SEE / *see*] [BUTTERFLY / *butterfly*]] and a [[SEE / *see*] [ENTITY]] construction). 'Merge', the most powerful operator, takes the structural overlap in form and function between two constructions, and adds this overlap as a novel construction to the grammar. An example would be the case where the model has a [[GRAB / *grab*] [BALL / *ball*]] and a [[GRAB / *grab*] [DOLL / *doll*]] construction. The structural overlap in the phonological and conceptual structure constitutes the more schematic construction [[GRAB / *grab*] [TOY]], which is then hypothesized as a new construction. Note that with 'merge' the two daughter constructions are not discarded, as in Stolcke's (1994) version of Bayesian Model Merging, but rewritten as inheriting structure from their newly abstracted parent.

In the experiments, Chang lets the model start out with many lexical constructions in place and thus the simple mapping operation (used for learn-

ing new lexical constructions) is left out. Given this starting point, the model shows improvement over time in analyzing utterances (both in the amount of the utterance analyzed and the amount of the situation interpreted). The grammar furthermore stabilizes over time.

Alishahi and Stevenson (2008)

Alishahi & Stevenson (2008) develop a model that learns argument-structure constructions that generalize over particular verbs. The starting point for their model are input items are frames consisting of a predefined set of functional or conceptual and formal features which are specified with various values. The model starts out, in the training phase, with the knowledge of the meanings and distributional categories of words (sometimes left out if there is noise – see chapter 4 for a fuller treatment), so that the argument structure (e.g., 'argument-1 + verb' or 'argument-1 verb *on* argument-2') can be part of the set of features. Other features include the semantic representations of the event, the event roles, and the entities filling the event roles.

When processing a novel frame, the model tries to categorize it as belonging to one of an (initially empty) set of clusters over frames. These clusters represent abstractions over the frames that allow the model to go beyond the observed input. In the clusters, each feature specifies a probability distribution over the values as they were observed in the frames that were categorized as belonging to that cluster. As such, the clusters are centroid representations of the frames that were categorized with them.

The process of categorization (or clustering) is conceived of as a Bayesian inference problem, where the frame is added to the cluster with the maximum a-posteriori probability. This probability is given by the prior probability of the cluster (roughly, its frequency) and the likelihood of the frame given the cluster (roughly, how well the cluster fits the frame). A smoothed part of the probability mass is assigned to the possibility of letting a frame form a new cluster. This probability mass depends on the amount of observed frames, and, as such, decreases over time. After a frame has been categorized with a cluster, the probability distributions over the various features are updated with the values of the frame.

In the experiments, Alishahi shows that several observed phenomena in child language acquisition can be simulated, including Akhtar's (1999) Weird-Word Order experiments, and the overgeneralization and recovery thereof of argument-structure constructions (Bowerman 1982). Importantly, the process of categorization can also be used to make top-down predictions about missing features. This way, novel verbs can be 'bootstrapped' given the (known) arguments and the categorization of these arguments with a certain cluster.

Kwiatkowski (2011)

Although Kwiatkowski's (2011) model is not framed as a usage-based model, it is worth discussing here, because it contains one feature that no other model contains, namely the fact that the meaning of lexical patterns (words) and the parametrized meaning of grammatical patterns are acquired at the same time. In Kwiatkowski's model, a Combinatory Categorial Grammar (CCG) formalism (Steedman 2000) is adopted for the development of a semantic parser. In this formalism, grammatical rules are instructions to create larger, more encompassing syntactic and semantic representations. The model is therefore a strictly componential one: all lexical semantic representations are associated with words, and all rules for combining lexical semantic representations into larger wholes are associated with grammatical rules. Given this division of labor, the model has to learn two things: the set of correct rules given a hypothesis space of possible CCG rules, and a lexicon of pairing of phonological form and lexical meaning.

In an incremental fashion, the model updates the (pseudo-)counts of the word-meaning pairings as well as that of the various possible rules, and eventually learns to parse and analyze utterances. It does so by, at every input item, trying all possible parses and word-meaning pairings given the utterance and the situational context. Each parse with word-meaning pairings at the terminal nodes now has a a certain probability given the pseudocounts of the previous step, and pseudocounts for the next step are updated with this probability. By doing so, the model gradually assigns more pseudocounts to rules and lexical items it has seen before, and thereby learns a grammar and a lexicon at the same time. Interestingly, there is a bi-directional bootstrapping process in the model: if it is very certain about the word meanings, but not so much about the grammar rules, the grammar rules still receive a large pseudocount update. On the other hand, if the model is certain about the rules, the weaker representations of the word-meaning pairings will be strongly reinforced.

μ-DOP

As a first attempt at modeling the acquisition of a grammar with semantic representations, I developed a model starting from Data-Oriented Parsing (DOP: Scha 1990, Bod 1998, Bod, Scha & Sima'an 2003), more specifically from Unsupervised DOP, (U-DOP: Bod 2009). U-DOP is a distributional learner (i.e., only the form into account) that builds on a very simple principle: assume all possible binary trees over a corpus of strings, and store all possible partial trees (subtrees). When analyzing a novel utterance, the best analysis is taken to be the one that involves the combination of the fewest subtrees in order to derive that utterance. If multiple such combinations, or *shortest derivations*, are possible, take the one that is most likely given the relative frequency of the partial trees in the observed corpus (the Most-Probable Shortest Derivation,

or MPSD principle).

This perspective constitutes a promising starting point for developing a model that takes both form and meaning into account. The reliance of U-DOP on units of heterogeneous size and the fact that the model allows for redundant representations is congenial with usage-based constructivist tenets. The fact that the model tries to stay as close as possible to what it has seen most (i.e., the MPSD principle) is furthermore a clear operationalization of Langacker's idea that more concrete units have precedence over more abstract ones in language processing.

This idea was worked out in Meaningful Unsupervised Data-Oriented Parsing (μ-DOP: Beekhuizen & Bod 2014). To include the acquisition of meaning, we followed the same basic intuition as with U-DOP: try all possibilities, and let the frequencies of the various possibilities 'decide' what are the best rules. When we include meaning, the range of possibilities does not only contain all possible branching structures, but also all possible combinations of parts of the meaning with parts of the branching structure.

Starting with no knowledge of the grammar, μ-DOP processes one input item at a time. As the model is a semantic-grammar learner, the input consists of pairings of an utterance and a situational representation, in our case, a simple logical form. In processing the input item, the model tries all derivations given the grammar so far, as well as a set of unseen rules, consisting of combinations of meaning splits from the situational context with all possible binary branchings. The rules in the grammar have a probability relative to their pseudocounts in the grammar, whereas the unseen rules split up a small probability mass reserved for unseen events that decreases as more rules are learned.

The probability of each derivation then, is the product of the subtree probabilities. After creating all possible derivations over the utterance, the model updates the pseudocounts of the subtrees used in all derivations with the their (normalized) probability among all derivations. Note that μ-DOP does not instantiate the MPSD idea: the model simply takes all derivations and updates the grammar with them. However, derivations consisting of fewer subtrees are often more likely given the probability model (the combination of three subtrees involves the product of three probabilities, which is typically higher than the product of four probabilities, in case of a derivation consisting of four subtrees).

As with Kwiatkowski's model, μ-DOP incrementally figures out which grammatical as well as lexical representations occur more frequently and are therefore more useful in understanding novel utterances. A key difference from Kwiatkowski's representations is that the grammatical units (i.e., structures with more than one terminal node) may contain lexical semantic content, whereas in Kwiatkowski's model, they only contain instructions for combining lexical semantic content. This reflects a difference in starting assumptions: whereas Kwiatkowski's model is a componential one, with separate roles for the lexicon and the grammar, the μ-DOP learner is agnostic about the proper

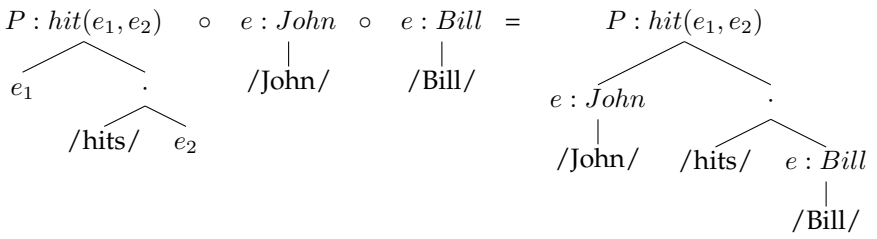


Figure 2.1: Some representations and the way they are combined in μ-DOP.

location of lexical semantic content: both words (being unary subtrees) and 'grammatical' rules (being all sorts of constellations of binary branching subtrees) may carry lexical semantic content.

Figure 2.1 illustrates several μ-DOP representations and how they are combined. The first construction can be seen as a verb-island construction (cf. Tomasello 1992), where a particular verb distributes its roles. Translating this format into a Cognitive Grammar representation, we have the three units in examples (20)-(22), being combined into the construct in example (23).

(20) [[E_1] [*hits*] [E_2]] | HIT(E_1,E_2)

(21) [E:JOHN / *John*]

(22) [E:BILL / *Bill*]

(23) [[E_1]→[E_1:JOHN / *John*] [*hits*] [E_2]→[E_2:BILL / *Bill*]] | HIT(JOHN,BILL)

In an experiment on toy data, μ-DOP was presented with input items consisting of a pair of seven situations and an utterance corresponding to one of the situations. As a global measure of evaluation, I evaluated how often the most likely derivation referred to the correct situation, that is: the situation that was paired with the utterance that the model processed. If the model picked out the correct situation, it achieved some sort of communicative success, and so the model should achieve a state of knowledge with which it can communicate succesfully.

The situations were simple semantic predicate-argument structures of one-place and two-place semantic predicates. Importantly, the model contained a number of non-compositional idioms. Besides the semantic predicate SEE(E_1, E_2), which was expressed by utterances such as *Bill sees Abe* or *Mary sees Jack*, there was also an intransitive semantic predicate TURN.50(E_1) that was expressed with the expression *X sees Abe*, with any entity filling the E_1 on the

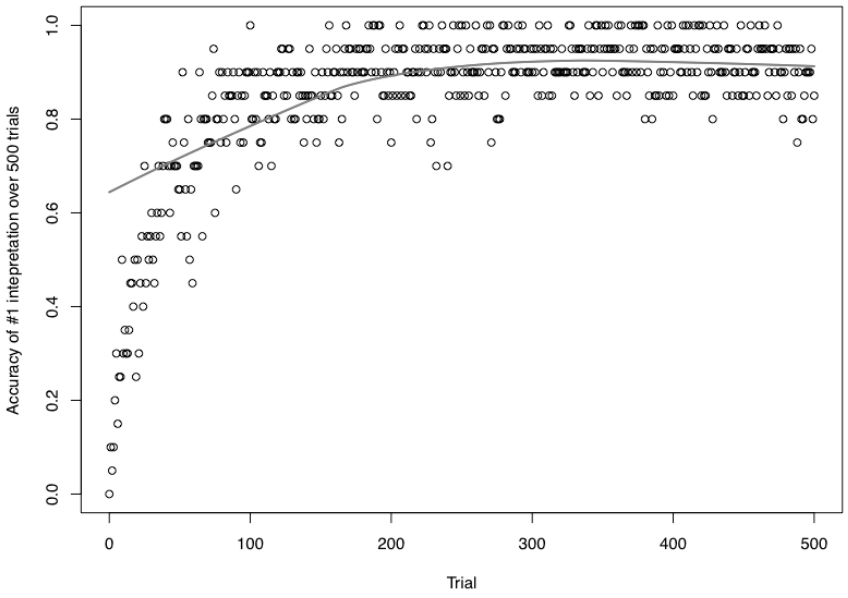


Figure 2.2: The average accuracy of μ-DOP in the first 500 trials.

X.[12] This means that the model faces two tasks: on the one hand, there is the acquisition of a non-compositional idiom, on the other: there is an ambiguity with the literal interpretation that hinders this acquisition.

Figure 2.2 gives the results of the accuracy of the most likely derivation over time (averaging over 10 simulations). We can see the model converging to a good performance. Given the limited nature of the toy example, it should not be very hard for the model to understand how the various meanings are expressed, but the amount of uncertainty (6 'distracting' situations) appears to be no problem for the learner.

On a more qualitative level, we can look at what derivations are being used. Of particular interest in Beekhuizen & Bod (2014) was the acquisition of idioms. Looking at the *sees Abe* idiom, we find the derivations in figure 2.3 (from different simulations) after about 500 input items. We can see here that in analysis 1, the model (correctly) combines an intransitive construction with the lexical construction referring to ED and the verbal idiom *sees Abe*, meaning TURN.50. In the second case, the model has associated TURN.50 solely with the word *Abe*, and has acquired a lexical construction in which *Ed sees* simply

[12]The idiom was modeled after the Dutch expression *Abraham zien*, 'lit. seeing Abraham, turning fifty'.

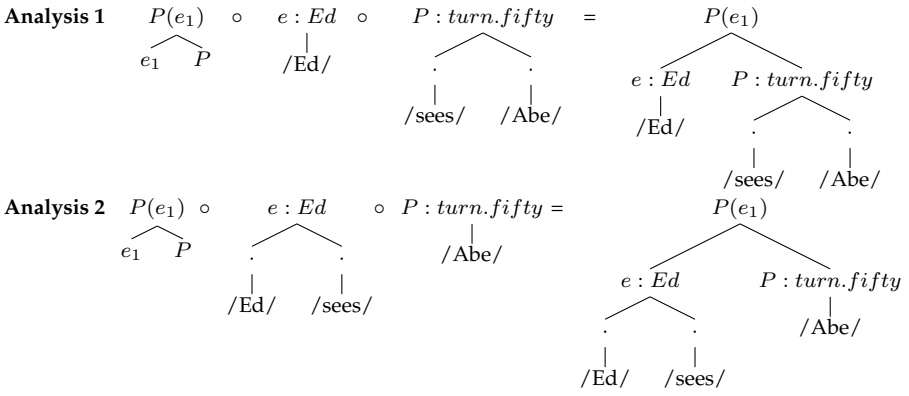


Figure 2.3: Two derivations of *Ed sees Abe* in a situation where TURN.50(ED) is present.

refers to ED, which are combined with an intransitive construction.

2.5.2 Usage-based distributional models

Besides models that directly operationalize the constructivist tenet of grammatical knowledge consisting of form-meaning pairings throughout, there are also models that take meaning out of the equation and focus on what can be learned from the formal distribution of elements in the input. Both models discussed here, MOSAIC and CBL, acknowledge the importance of meaning, but focus on the role of formal distributions in the input. In the case of the former, the focus is on the incremental build-up of a network of words leading to increasingly long productions, whereas the latter aims to model the role of multi-word units or 'chunks' in language development.

MOSAIC

The first distributional model to be discussed is MOSAIC (Freudenthal et al. 2010). MOSAIC processes utterances from a corpus one by one and incrementally builds up a network of phrases it has encountered. Crucially, the processing is limited by an utterance-final bias: an unknown word is only added to the network if everything that follows is already encoded in the network. In a later version, an utterance-initial bias is added as well, so that the model gradually builds up phrases at both edges of the utterance, as well as the concatenation of both (so that if it has seen *where* at the beginning of an utterance and *he go?* at the end, it can concatenate both to form *where he go?*). New nodes for unknown words are added to the network with a probability reflecting a

bias towards shorter phrases (adding a fourth word in a chain is less likely than adding a third) and an increasing ability to integrate nodes in the network.

Some variants of the model furthermore implement a second kind of link in the network, viz. a 'generative' link. With a generative link between two words, MOSAIC can substitute the current word for any word linked to it with a generative link. This allows the model some generalization beyond what it has directly observed.

After having seen a number of utterances, the model can be evaluated by having it generate a number of strings given the network. Each string is generated by following a path through the network. After few processed input items, the utterances are short, and over time they grow longer. Given this property, it can be shown how phenomena such as subject-omission follow from the nature of the input in combination with the simple edge-biased learner. More interestingly, the rate of root infinitives in different languages (utterances of the type *daddy grab*, where the inflected verb is left out) can be modeled as a product of the input: some languages display many such utterances in early child speech, whereas children learning other languages hardly ever leave the verb uninflected if it is obligatory to inflect verbs. MOSAIC accurately models the proportions of root infinitive errors found in various languages because the languages vary in how often they have infinitives in constructions with an inflected auxiliary. More infinitives in [auxiliary + infinitive] patterns means that, given a right-edge bias, the learner has more opportunity to pick up these infinitives without the inflected auxiliary, and therefore produce more utterances of the type *where he go?*.

Another phenomenon studied with MOSAIC is the nature of early generalizations, especially the Verb-Island hypothesis discussed in section 2.4.1 (Jones, Gobet & Pine 2000). Using the generative links, MOSAIC can be shown to simulate distributions of [noun + verb] and [verb + noun] combinations in production that are closer to the child's distributions than to the caregiver's distributions. Moreover, the amount of pronouns, proper nouns and common nouns in the model's outputted utterances matches more closely to the child's than to the caregiver's. Finally, the model can be shown to have both verb-specific constructions, of the [[noun] [*hit*] [noun]] type, as well as argument-frames, such as [[*you*] [verb]].

The CBL model

McCauley & Christiansen's (2014*a*) Chunk-Based Learner (CBL) is a similar distributional learner to MOSAIC, aimed to show the role of multi-word units in language development and language use. CBL processes utterances word by word, and keeps track of the backward transitional probabilities (BTPs) of every word given the next word. When it encounters a peak in the BTP between two words, that is: a probability of the current word given the next one that is *higher* than the average BTP over the entire corpus, the two words are

'chunked' together. When a dip in the BTP is encountered (i.e., a probability of the preceding word given the current one that is *lower* than the average BTP over the corpus), a boundary is placed and all chunked words preceding the boundary (at least the preceding word, but possibly more) are placed in the 'chunkatory', the inventory of chunks. These chunks are then used in subsequent processing: if, for any number of words, a chunk can be found, the words in the utterance that are subsumed by it are automatically chunked and no BTPs are calculated.

McCauley and Christiansen evaluate the CBL on various tasks. In a production experiment, they give the model a sentence that a child produced in an unordered form (i.e., as a multiset of words) under the assumption that this is an approximation of the meaning of the utterance. They then ask the model to find the most likely sequence of chunks given the start-of-utterance symbol. Importantly, this is a process that happens incrementally over the utterance: there is no whole-utterance optimization. When the most likely sequence is found, it is scored against the actual utterance. Using this production process, they are able to simulate children's utterances in a typologically varied sample of languages.

Furthermore, CBL simulates a number of interesting findings on the repetition of multi-word units in children. To give an example: Bannard & Matthews (2008) found that two and three-year olds were significantly more likely to correctly repeat a four word utterance if the first three words formed a chunk, and three-year-olds furthermore did so faster than for correct repetitions of four word phrases that contained no chunks. CBL closely mimics this behavior when trained on child-directed speech. An interesting conclusion McCauley and Christiansen draw is that "the importance of multi-word units may actually grow, rather than diminish, throughout development" (p. 428). That is: chunks are not only a stepping stone for the early language learner, but may continue to play an (even increasing) role in later processing.

2.5.3 A comparison

How well do the models discussed in this section instantiate the idea of a usage-based learner? And for which of the developmental phenomena do they account? Table 2.2 gives my, admittedly highly oversimplified, assessment. A '$+$' means the model satisfies that desideratum or has been succesfully used to simulate that empirical finding. A '$-$' means that either the model does not satisfy that desideratum, or has not succesfully been shown to simulate that empirical explanandum. A '$\Diamond$', finally, means that the model does not satisfy that desideratum or simulate that explanandum, but has, as I will discuss below, the potential to do so.

desideratum/explanandum	(Chang 2008)	(Alishahi & Stevenson 2008)	(Kwiatkowski 2011)	(Beekhuizen & Bod 2014)	(Freudenthal et al. 2010)	(McCauley & Christiansen 2014*a*)
D1 (explicitness)	+	+	+	+	+	+
D2 (comprehensiveness)	◊	−	◊	◊	−	−
D3 (simultaneity)	◊	−	+	+	−	−
D4 (representational realism)						
D4-1 (qualitative grounding)	+	+	−	+	+	+
D4-2 (quantitative grounding)	+	+	+	+	+	+
D4-3 (immanence)	+	+	−	+	+	−
D5 (processing realism)						
D5-1 (heterogeneous structure building)	−	−	−	−	+	−
D5-2 (linear processing)	−	−	−	−	+	+
D6 (ontogenetic realism)						
D6-1 (cumulative complexity)	◊	−	−	−	+	+
D6-2 (learning-by-processing)	−	+	+	+	+	+
D6-3 (parts-to-whole and whole-to-parts)	+	−	−	−	+	−
D6-4 (developmental continuity)	+	+	+	+	+	+
D7 (explanatory insight)	+	+/−	+	+/−	+/−	+/−
D3-1 (unification)	−	+	−	−	+	−
E1 (decreasing argument omission)	◊	−	−	−	+	−
E2 (prevalence of subject omission)	◊	−	−	−	+	−
E3 (co-varying complexity)	−	−	−	−	−	−
E4 (overgeneralization and retreat)	◊	+	−	−	−	−
E5 (mechanisms overgeneralization)	−	−	−	−	−	−

Table 2.2: A comparison of the various learners discussed in section 2.5.

D1: Explicitness

Explicitness is one of the hardest desiderata to evaluate. All models discussed here are explicit about what simplifying assumptions they make, and what they do and do not take into account. For instance, Chang, Alishahi & Stevenson, Kwiatkowski and Beekhuizen & Bod take conceptualizations of the situation as part of the input into account and do so with a varying degree of naturalism. However, all of them are very explicit about the artificial nature of the representation (in being derived from the utterance, for instance). Moreover, Chang and Alishahi & Stevenson discuss how their conceptual representations are grounded in ideas about how meaning and conceptualization work.

D2: Comprehensiveness

None of the models has been used for accounting for the full process of language use, that is: going from conceptualization to an utterance in production and from an utterance to a conceptualization in comprehension. Although Chang, Kwiatkowski and Beekhuizen & Bod certainly have the potential to do both, in none of these works the model is shown to be able to perform both tasks.

D3: Simultaneity

Which models learn words and their meanings and grammatical patterns and their meanings at the same time? Only the models of Kwiatkowski and Beekhuizen & Bod have been shown to do so. Chang (2008)'s model certainly has the mechanisms to do so. However, as she does not evaluate the model starting with no constructions at all, we cannot tell if the mechanisms actually let the learner build up an inventory of lexical and grammatical constructions at the same time.

D4: representational realism

D4-1 and D4-2 All of the models under scrutiny have their representations grounded, both qualitatively (in what they contain) and quantitatively (in keeping track of the frequency of their usage), in aspects of the usage events. The sole exception concerning the qualitative grounding is Kwiatkowski, whose combination rules come from a universal grammar and are thus not derived from properties of language use. This is obviously not a problem, as Kwiatkowski does not frame his model as a usage-based learner.

D4-2 All usage-based models that employ some notion of abstraction (i.e., Chang, Alishahi & Stevenson, Beekhuizen & Bod, Freudenthal et al.) satisfy desideratum D4-3, viz. that abstractions should be, at least conceptually, immanent in the constructions they are derived from. In Chang's model this is

achieved through the use of inheritance relations between constructions. Alishahi & Stevenson do so by taking a clustering approach, where the centroid representation of the cluster of usage events can be seen as the shared potential of a set of usage events. In Beekhuizen & Bod, the Data-Oriented Parsing dictum 'assume all substructures and let the statistics decide' (cf. Bod 1998) is taken to instantiate the property of immanence.

Nevertheless, all three models represent the abstract constructions as separate entities. I do not believe this to be a problem for the models. Although the discussion about immanence does not show up in any of them, none of them is incompatible with the view that abstractions are co-activation patterns over multiple exemplars. MOSAIC is interesting in this respect, because there the immanence is most faithfully implemented: abstraction over positions in the chain of a network is represented with the generative links, which are only made if the distribution of the words on two nodes is similar enough. Here, abstraction is truly not something distinct from the actual usage events.

D5: processing realism

D5-1 The models that employ methods for combining structure (i.e., all except Alishahi & Stevenson's) mostly employ a single means of doing so. This is a slot-filling operator in the case of Chang, Beekhuizen & Bod, and Kwiatkowski, and a concatenation operator in McCauley & Christiansen. Only MOSAIC allows for both concatenation, by following the regular links in the network, and a form of substitution, with the generative links.

D5-2 Interestingly, the focus on linear processing and non-global optimization is stronger in the two distributional learners than the models that involve meaning. In all four semantic learners, the best analysis or the relative goodness of the analyses is found through a probabilistic calculation that takes the full utterance into account, thus running counter to the idea that language users process utterances linearly and without doing utterance-wide optimization of the analysis. Both distributional learners, however, engage in a strongly constrained process of analyzing the utterance.

D6: ontogenetic realism

D6-1 Both distributional models satisfy the constraint that more complex representations are to be built up from simpler ones. In MOSAIC, this is done by letting the network incrementally grow as more input is processed. The network up until that point constitutes the simpler representation, which is used to bootstrap the more complex representation the netwerk contains after the processing. Similarly, CBL uses its chunks at a certain point to find bigger chunks, and thus uses the simpler chunks as bootstrapping devices. None of the semantic learners show how, for instance, longer argument-structure patterns can only be learned after having seen simpler ones. In Kwiatkowski and

Beekhuizen & Bod, the effect of considering all hypotheses is that the maximal level of syntagmatic complexity is in principle already within reach after having processed only the first exemplar. Alishahi & Stevenson's model assumes lexical mappings to be (largely) in place prior to the acquisition of the argument-structure constructions, and thus does not account for larger representations (with more arguments) being built up from simpler ones. Chang's model is interesting, because it has the potential in its learning mechanisms to build up more complex representations from simpler ones, but this potential is not evaluated, because the model starts with full knowledge of lexical constructions. Because of this, argument-structure constructions are first acquired with their full width rather than being bootstrapped using simpler constructions.

D6-2 The idea of learning-by-processing is instantiated in all models: all have an account on which an input item is processed and the results of that processing inform the learner about updates and extensions of the grammar. The reason I scored Chang with a '$-$' is that the acquisition of abstract representations crucially involves a decision making procedure on the basis of the Minimum Description Length principle. This constitutes, to my mind, a case of post-hoc decision making (especially in the case of the functions other than the two mapping functions), and one, moreover, in which the value of adding a rule in the light of the whole grammar is considered.

D6-3 Only two models allow for both parts-to-whole and whole-to-parts learning, viz. Chang and Freudenthal et al.. Chang's model is the clearest example: with the learning mechanisms defined in her model, constructions can both be joined and split, thus allowing the model to go from parts to wholes and from wholes to parts. MOSAIC allows for parts-to-whole learning by the build-up of the network, and whole-to-parts learning by the generative links. In Beekhuizen & Bod and Kwiatkowski, the parts are known (or: being learned). Although Beekhuizen & Bod allow for wholes with structure associated with them, the parts are always already present in them, and likely have some probability mass assigned to them as well. In Alishahi & Stevenson's model, the blame assignment is done in advance, under the assumption that the learner already knows the lexical constructions.

D6-4 All models assume the same set of mechanisms to be available throughout development. Interestingly, in Chang's model, the frequency of use of some mechanisms may increase over time, while the frequency of use of others may decrease, but all of them are available at all times.

D7: Explanation

Like explicitness, the amount of explanatory insight is hard to evaluate. The model of Chang stays very close to a usage-based theory of language acqui-

sition, but for all the others, the relation between theory and model takes a slightly larger interpretive step. This is in principle not a problem, and it may of course be the case that insights from modeling change some theoretical conceptions. Alishahi & Stevenson's and Freudenthal et al.'s model stand out for their unifying properties. In the former, it is shown how several processes, such as syntactic bootstrapping, overgeneralization, and decreasing verbal conservatism (in the weird-word-order studies) are all effects of the same probability model that employs the induced clusters. In the latter, the right-edge bias together with the incremental build-up of the network are factors that account for several empirical phenomena (argument omission, the presence of optional infinitives) that are often thought to be due to a distinct range of factors.

E1, E2, and E3

Issues concerning the limited length of early productions are generally not extensively studied using computational modeling techniques. Most models start out with the learner having the ability to process the full utterance and derive possible representations from it. The limitation on processing in MOSAIC allows this model to simulate the decreasing amount of argument omission, and with the right-edge bias, it can be shown how subjects are left out more often than other arguments. Of the other models, I believe only Chang's has the potential to simulate this, although we do not know if the model will actually do this, given that in the simulations the model knows the set of lexical constructions in advance, and can thus process the whole utterance, thereby building up representations of full width.

E4 and E5

The only model that has addressed the overgeneralization of argument-structure constructions and the retreat thereof, is Alishahi & Stevenson (2008).

Wrap-up

The models vary in the extent to which they meet the desiderata and explananda pose throughout this chapter. Of course, this comparison is slightly anachronistic: the modeling enterprise, like any field, proceeds in small steps, and the desiderata and explananda are formulated by someone who was able to look at more than a decade of progress in the field. The various points of criticism should therefore be read as an agenda: we would like a cognitive model of language acquisition to satisfy all of them. The model I will present in chapter 3 constitutes my attempt to do so.

CHAPTER 3

The Syntagmatic-Paradigmatic Learner

3.1 Introduction

In this chapter, I introduce the Syntagmatic-Paradigmatic Learner (SPL for short), a computational model of the acquisition of linguistic representations that constitutes the culmination of my inquiries presented in chapter 2. In that chapter, I discussed several desiderata and explananda for a usage-based learner. With those in mind, I developed a model that satisfies many desiderata and explains most of the explananda (as we will see in the later chapters) with a limited set of mechanisms and representations.

Globally speaking, SPL is an incremental learner that processes input items one by one. Each input item consists of an utterance, paired with a set of situations to which the utterance can refer. SPL tries to analyze the utterance on the basis of the situational context, its current state of linguistic knowledge, and several general processing operations. Using the resulting analysis, SPL updates and expands its linguistic knowledge. The learning gets off the ground by a procedure of analogical reasoning over recent exemplars. Using this procedure, the model is able to learn initial lexical mappings between form and meaning.

Unique features of SPL are that it performs the full comprehension and production task (desideratum D2), and acquires lexical and grammatical constructions at the same time (D3). The gradual build-up of the representations in the model through the syntagmatization and paradigmatization operations (defined below) furthermore makes SPL a faithful implementation of the usage-based conception. At the same time, it addresses those aspect of the

usage-based approach that I deemed unsatisfactorily worked out (cf. section 2.2).

As a note for readers not accustomed to reading set-theoretical and graph-theoretical definitions, and probability calculations: I will only employ the high degree of formalization in this chapter, and try to explain and motivate it in the text surrounding the formalization. The formalization is meant to show how one can operationalize certain usage-based notions.

3.2 General properties of input items to the model

3.2.1 Input items: utterances and conceptualizations of situations

SPL takes as its input pairings of an utterance and a number of conceptualizations of situations that the learner considers to be the possible conveyed meanings. The idea that the language-learning child has a conception of the possible meaning of an utterance (in a conceptualization of a situation) is a logical necessity for symbol acquisition to get started. At the very least, not all of the possible concepts a child can entertain should be considered to be signified by every utterance, as this would disallow any correct associations to be formed.

The assumption that the meaning of an utterance can be independently obtained has been commonly made, and has been labeled the Interpretatibility Requirement (O'Grady 1997, 260), put forward most eloquently by MacNamara:

> It is not too fanciful to think of the infant as treating the sentences he hears as glosses on the events that occur about him. The grammar he writes is not in Latin or in any other language, but in some neurological code of which as yet not a single letter has been deciphered. (Macnamara 1972, 12)

The most obvious source of this language-independent understanding is the perception of the situation in which the language is used (Gleitman et al. 2005, 28). In fact, the primary external source of obtaining a set of candidate meanings is experience. As we know from work such as Tomasello & Farrar (1986) and Baldwin (1993), this does not necessarily mean the *perceptual* experience of the *immediate* situation in which the utterance is produced (although that is the simplest imaginable source); it can also include concepts *inferred* on the basis of perceived situations (e.g., mental states such as intentions and attitudes) as well as non-immediate situations (concepts not present in the here and now of the speech situation, or in the child's visual field, but nevertheless deemed relevant by the child because of these inferential mechanisms). Nonetheless, the simplest source of potentially signified concepts is the perception of the situation that is spatially and temporally contiguous with the

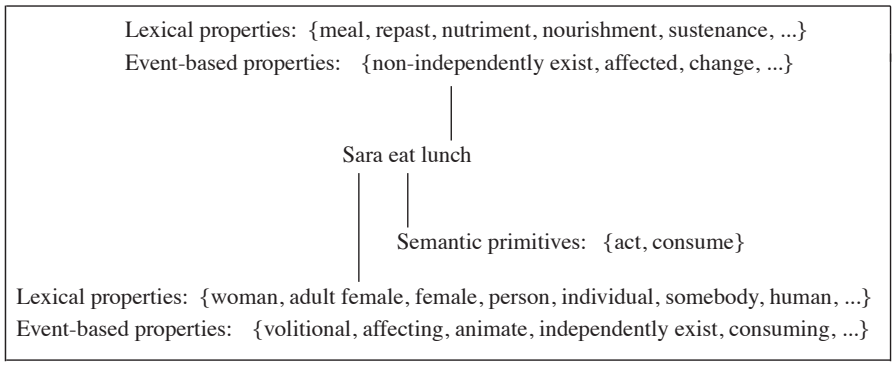


Figure 3.1: Semantic features extracted on the basis of the utterance in Alishahi & Stevenson (2010, 59).

utterance, as this is a source that requires little further cognitive sophistication to arrive at, and that is attested in other species as well (Goodall 1986, Savage-Rumbaugh, Murphy, Sevcik, Brakke, Williams & Rumbaugh 1993, Kaminski, Call & Fischer 2004).

The Interpretability Requirement may, however, be too strong compared to the situations the child finds herself in. It may be that the correct situation is not observed, for instance. Furthermore, there may be many situations besides the correct one that are initially equally likely to be the situation the utterance refers to. These issues constitute a topic that many computational models discuss, but the empirical grounding on the eventual decision they make concerning the frequency with which the correct situation is absent and the number of 'distracting' situations being co-present, is thin. For that reason, I decided to venture into this topic empirically by looking at videotaped caregiver-child interaction. The results of that exploration and an answer to the question how to provide the computational model with realistic input items are discussed in chapter 4.

3.2.2 The structure of the conceptual representations

In dealing with the acquisition of a constructicon, hierarchical representations of meaning are required. Rather than taking recourse to approaches to meaning based on formal logic, I do so by using a graphical structure with sets of features on the nodes that reflect the subtleties of conceptualization better. To do so, I make use of Alishahi & Stevenson's (2010) input-generation procedure. In their procedure, an utterance with a conceptual frame is generated. An example is given in figure 3.1. We can automatically extract hierarchical

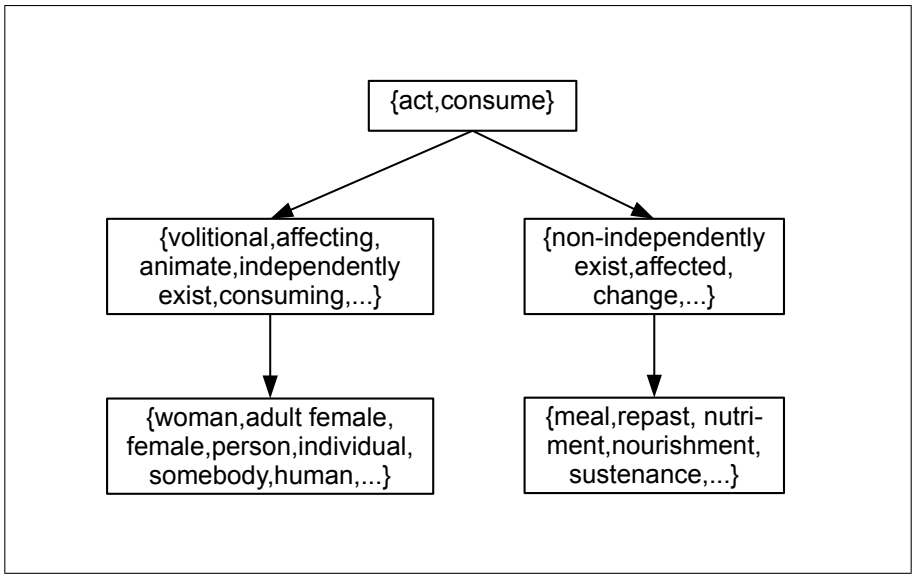


Figure 3.2: An example of a situation.

conceptual representations from Alishahi and Stevenson's procedure given the following three basic rules:

- The event node is the root node.
- The semantic role nodes, or event-based properties are daughters of the root node.
- The semantic argument nodes, or lexical properties, are daughters of a semantic role node.

For the example in figure 3.1, we obtain the structure found in figure 3.2. This structure constitutes (a conceptualization of) a situation s. As we will use conceptual graphs more in the model, it is useful to have some general definition. A situation is a graph G, which consists of a pair of a set of vertices V (or nodes), each of which contains a set of conceptual features, and a set of unlabeled directed edges (or links) E, connecting pairs of vertices in V. As we will see, the meanings of linguistic representations consist of meaning graphs as well.

In Alishahi & Stevenson's (2010) procedure, every generated situation is paired with a linguistic argument structure and a set of words filling the main

predicate and argument positions. Together, these constitute the utterance U. The argument structure for the situation in figure 3.2 would be ARGUMENT$_1$ + PREDICATE + ARGUMENT$_2$, but prepositions can also be part of the argument structure, for example in ARGUMENT$_1$ + PREDICATE + ARGUMENT$_2$ + *on* + ARGUMENT$_3$.

For now, this short exposition suffices to give an idea of the structure of the representations. Chapter 4 will deal with the exact properties of the input generation procedure.

3.3 Constructions

3.3.1 Constructions as representational primitives

The only representational unit of linguistic knowledge employed in SPL is the construction. While there are many perspectives on what a construction is within the theory of construction grammar, I start off from Verhagen's (2009) vantage point (cf. the discussion in section 2.1.1). Recall that Verhagen argues for the importance of the conceptual distinction between the contents of constructions and the roles these contents play. Crucially, a construction is a symbol, that is: a conventional pairing of a signifier and a signified. Signification entails that when the hearer observes a signifier, he infers that the speaker intends him to conceptualize the signified. The conventionality means that the signification process relies on a mutual understanding of the inferential process of signification between any two members of a language community.

The next question is what kinds of elements we assume to be present as the signifying and signified roles of a construction. Following desideratum D4-1, we assume only phonological and conceptual structure to be the elements out of which constructions are built. In the simplest case, that of words, the signifying element is a phonological string, and the signified element a conceptual representation. Grammatical constructions, however, often have non-phonological signifiers. As Verhagen argues, conceptual structure can be taken to fufill the role of a signifying element as well, and it is this content type that constitutes the signifying element in many grammatical constructions in SPL.

In grammatical constructions, we can also see a second property of the signifier, namely that it can be complex, that is: consisting of multiple elements. It is this property that allows language its expressivity: signification processes can be recursively applied to the outcomes of other signification processes, and multiple signifieds can function together as the signifier of a larger, more encompassing construction, effectively giving rise to a hierarchical interpretation of a phonological string.

We therefore assume that the signifier of a construction consists of a number of constituents, each specifying what kind of element (a phonological string, a conceptual representation, or both) should be satisfied for that con-

stituent to be recognized. The signified element of a construction is taken to be a conceptual graph that is a subgraph of a situation: as we will see, all signified conceptual structures are grounded in the situations, and as such, the meaning of the constructions is qualitatively grounded in linguistic usage events as well. Importantly, no matter how abstract, all constructions in SPL have a semantic representation as a signified. That is: I assume that there are no conventions in a language that are completely devoid of meaning.[1]

Finally, it has to be noted that this research just forms a proof of concept of the feasibility of operationalizing a usage-based constructivist approach to grammar learning, early production and comprehension. In order to model more complex phenomena than the ones we study here, richer conceptual representations and further extensions to the definition of a construction have to be assumed. To name a few: the current definition of conceptual structure as a graph without re-entrances is not suited for addressing issues of co-referentiality within sentences, as this would require multiple links in the conceptual graph to connect to the same node. In principle, there is no reason why this cannot be implemented in SPL. Furthermore, the signifiers of constructions are now strictly linearly ordered. This is unproblematic for a language like English, that relies heavily on word order and constrains the possible word orders rather strictly, but for languages with freer word order, we may want to loosen the strict linearity constraint and define constructions in terms of sets of signifying constituents which may or may not have some ordering constraints on them.

3.3.2 A formal definition of constructions and the constructicon

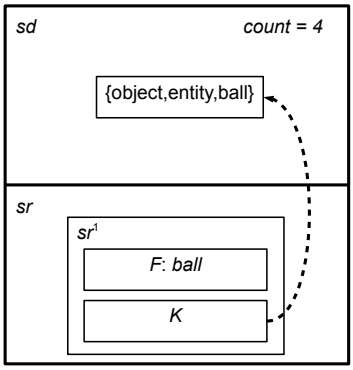
Formalizing these assumptions, we arrive at the following definition of a construction and a constructicon:

[1] This is an issue that has drawn some attention in the constructivist literature. Concerning the case of subject-auxiliary inversion in English, which is often considered to be a purely structural generalization, Goldberg (2006, ch. 8) argued that a common functional element to all cases can be found. A full treatment of the question whether purely structural generalizations exist falls outside the scope of this research, but one option that is rarely considered is that a generalization such as subject-auxiliary inversion in English may only be a *linguist's* generalization. This means that linguists may observe structural commonalities in several grammatical patterns, but that the language user does not have any sort of mental representation corresponding to these structural commonalities. That is to say: the abstraction over the various patterns is not made by the language user, but she rather maintains a number of semantically non-vacuous lower-level constructions.

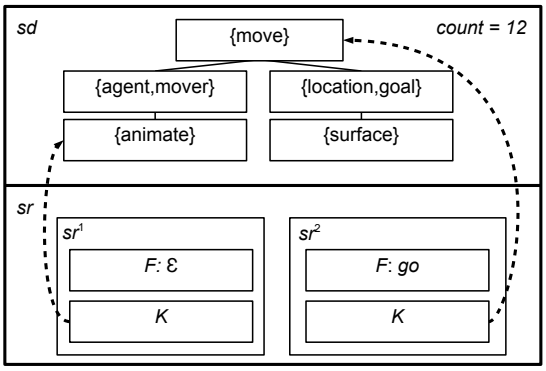
Definition of a construction and a constructicon

- Let α be a phonological element from the speaker's phonological inventory. In principle, there are no constraints on the size of α: it can consist of a single phoneme, or a string longer than a word. For the purposes of the experiments here, we set the lower bound on α to be a word in the input generation procedure.
- A construction c is a pair of a signifier sr_c and a signified sd_c, where:
  - sd_c is a conceptual graph G (as defined in section 3.2).
  - sr_c is an n-tuple (where $n \geq 1$) of constituents. Each constituent (denoted: sr_c^i for the i^{th} constituent) is a pair of a conceptual constraint K and a phonological constraint F, where
    - $K(sr_c^i)$ is a single vertex in G
    - $F(sr_c^i)$ is a string of phonological elements of any length greater than or equal to one ($F(sr_c^i) = \alpha^+$, where $^+$ denotes the Kleene plus) or unspecified ($F(sr_c^i) = \epsilon$)
  - A construction is furthermore associated with a $count_c = [0, \infty]$ reflecting how often that construction has been processed.
- We define the **head** constituent of a construction sr_c^{head} to be the constituent that has a conceptual constraint $K(sr_c^{\text{head}})$ such that $K(sr_c^{\text{head}}) = v_{\text{root}}(G)$.
- We define a **lexical** construction to be a construction c that has a single signifier, i.e., $|sr_c| = 1$.
- A constructicon Γ^t is a set of constructions $c_1, \ldots, c_n$, including their counts, at some time t

Figure 3.3 gives two examples of possible constructions. In Figure 3.3a we can see a lexical construction, containing a single signifying constituent. The construction's meaning is a conceptual graph consisting of a single vertex and no edges. The signifier consists of a conceptual constraint (K) pointing to the root vertex of G, and the phonological constraint (F) specifying that this construction can be recognized with the phonological string *ball*. Figure 3.3b, next, displays a grammatical construction, that is: a construction consisting of more than one signifying constituent. The signified meaning is a meaning graph G consisting of four vertices, each containing a set of conceptual features. The first signifying constituent sr^1 has a phonological constraint that is empty (represented as $F : \epsilon$) and a semantic constraint pointing to the vertex



(a) An example of a lexical construction.



(b) An example of a grammatical construction.

Figure 3.3: Two examples of constructions.

of G that contains the feature ENTITY. The second constituent sr^2 has a specified phonological constraint, viz. *go*, and a conceptual constraint stating that whatever is combined with this constituent must somehow combine with the feature set {EVENT,MOVE}. This second constituent, furthermore, is the head constituent of the construction, as its semantic constraint points to the root vertex of the constructional meaning.

As the box-diagrammatic format is often unwieldy, we will make use of a modified version of Langacker's (1987) bracket notation format, as introduced in chapter 1. The two constructions in figure 3.3 would be represented as follows in this format:

(24) [BALL / *ball*]

(25) [[ANIMATE] [MOVE / *go*]] |
MOVE(MOVER(ANIMATE),LOCATION(SURFACE))

3.4 Defining the space of possible analyses

When presented with an input item, the model employs its inventory of constructions and processing mechanisms to analyze it. Constructions can be applied if the string of signifying constituents is found to be present and if their meaning 'makes sense' in the context of one of the co-present situations.

I conceptualize the analysis of an utterance with constructions as a derivation process in which a fixed set of rules[2] is applied to an utterance. This is perceived for explanatory purposes as a top-down branching process (starting with a TOP-node, and terminating in the words of the utterance). However, as we will see in section 3.5.4, the model employs in the implementation a more realistic bottom-up process in which it does not keep track of all logical possibilities.

As the learner starts with no knowledge of the linguistic conventions, and as in the early stages of learning, the constructicon does not allow for full analyses of the utterance, the model will have to be robust enough to interpret parts of the utterance and situation on the basis of little knowledge. To this end, I define several rules that allow the model to create analyses of the utterance despite having little or no knowledge of linguistic constructions.

3.4.1 Mapping constructions to situations

I assume that SPL always interprets an utterance in the light of the observed situations in the input item. This means that the meaning of every used construction has to 'make sense' given at least one of the situations, or in other words: the model has to establish how the meaning of a construction is contextually resolved. In the simple case of a noun-like lexical construction, the

[2] I will call them 'rules' or 'mechanisms': these should be taken to be equivalent.

model can apply that construction if there is at least one element in the context to which the constructional meaning can refer. Potentially, there are more: a word may refer to a number of entities, possibly in multiple situations, in which case the hearer has to disambiguate to which entity the construction refers. What the model needs is a means of finding out what parts of the situational context S can be expressed by each of the constructions $c \in \Gamma$.

In order to link the constructions to the situations, subset mappings between signified meaning of the constructions and parts of situations are made. The (possibly empty) set of subset mappings $\mathbf{M}$ between the signified conceptual graph sd_c of a construction c and the situational context S consists of all legal subset mappings $\mathbf{map}_{\text{subset}}$ between sd_c and situations in the context. A mapping $\mathbf{map}_{\text{subset}}$ between sd_c and a subgraph of a situation $s \in S$ is established if and only if sd_c and the subgraph of s have the same edge structure and if the sets of conceptual features on the vertices of the subgraph of s are supersets of the sets of features on the vertices of sd_c.

Definition of subset mapping

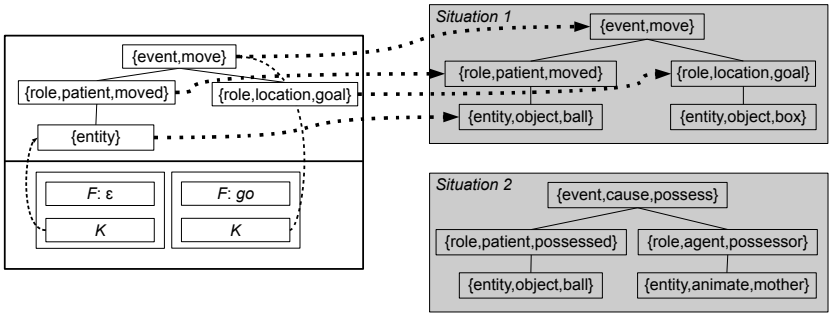
A subset mapping is an injective structure-preserving function $\mathbf{map}_{\text{subset}} = f : sd_c \rightarrow s$ between a signified constructional meaning sd_c of a construction c and a situation $s \in S$ such that

- $\mathbf{map}_{\text{subset}}(sd_c)$ is a connected subgraph of s
- The feature sets of all vertices in sd_c are subsets of the feature sets of the vertices $\mathbf{map}_{\text{subset}}(v) \in s$ they correspond with (i.e., $\forall v \in V(sd_c).v \subseteq \mathbf{map}_{\text{subset}}(v)$)

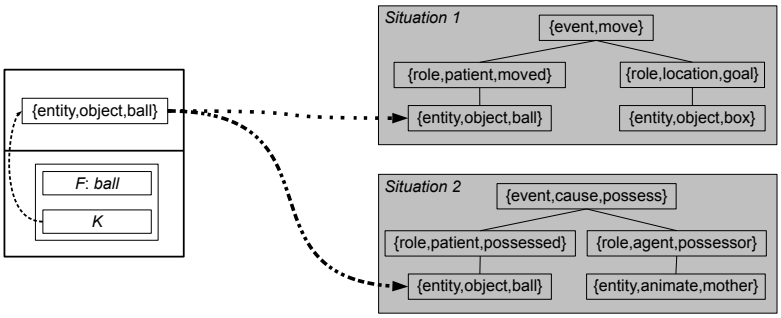
For any construction c, the set of possible subset mappings holding between sd_c and any $s \in S$ is denoted as $\mathbf{M}(sd_c)$. As per convention, we leave out the subscript when talking about subset maps, i.e., $\mathbf{map} = \mathbf{map}_{\text{subset}}$ (as opposed to other kinds of maps which we will encounter later).

Some examples of subgraph mappings are presented in figure 3.4. In the first example, we see a semi-open construction mapped to a subgraph of situation 1. Each of the four vertices of the constructional meaning maps to another vertex in situation 1, and all of the edges are preserved in the mapping. Furthermore, each vertex to which the vertices of the constructional meaning map is a subset of the conceptual features in the subgraph of the situation.

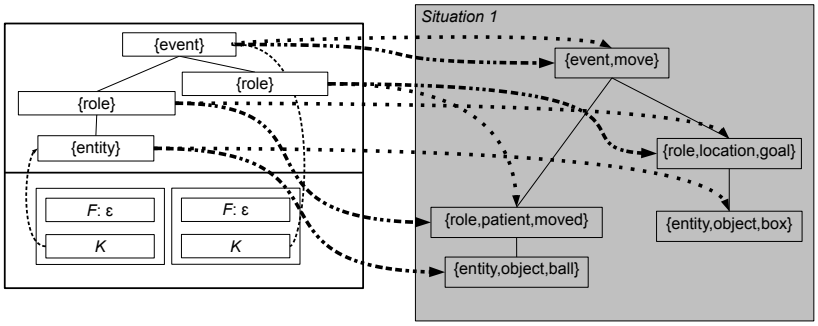
The second example shows a mapping of a lexical construction to two subgraphs, each in a different situation. The first mapping, represented as a dotted line, maps the vertex containing the feature set {ENTITY, OBJECT, BALL} to a vertex with an identical feature set in the first situation. The second mapping, represented as a triple-dash triple-dot line, maps that vertex to a vertex



(a) A subset mapping between a semi-open construction and the situational context.



(b) Two subset mappings between a lexical construction and the situational context.



(c) Two subset mappings between an open construction and the situational context.

Figure 3.4: Three examples of subset mappings. Different subgraph mappings are represented with differently patterned lines.

with an identical feature set in the second situation. In both cases, the edge structure is (trivially) preserved, and the content of the single vertex in the constructional meaning is a subgraph of each of the two vertices it maps to.

Finally, in the third example, we see what happens when we have a relatively abstract construction. Because the sets of conceptual features on the vertices of the constructional meaning are small, they have the potential of being the subset of many vertices in the situational context. In this case, the first vertex containing the conceptual feature set {ROLE} can be mapped onto either the vertex containing {ROLE, PATIENT, MOVED} of situation 1, or onto the vertex containing {ROLE, LOCATION, GOAL}, and similarly for the other vertices. Because of this, two subgraphs of situation 1 can stand in a subgraph mapping relationship with the construction.

The constraint that there needs to be at least one situational mapping in order to apply a construction is obviously an oversimplification: if a construction has a meaning that is not among the meanings considered to be relevant for communication, the model simply does not consider it. In adult linguistic communication, however, constructions can be referring to entities and events beyond what the hearer assumes the speaker to be considering, which means that these can nonetheless be retrieved and the communicative intent can be understood. Nonetheless, I believe that much of the infant's communication is based in the here-and-now of the situational context, and that, therefore, she will consider those primarily.

3.4.2 Three general constraints

Two general constraints on derivations furthermore hold. The first is that all constructions used in a derivation must be mapped, via a subset mapping, to the same situation. This is the principle of **coherence**, which ensures that the interpretation of the utterance is coherent. It relies on a communicative assumption that the speaker is trying to refer to a specific situation with her message.

The second constraint, **isomorphy**, states that the root vertices of the meanings of any two constructions used in a derivation may not be mapped to the same vertex in the situation. The principle of isomorphy constitutes a strong case of mutual exclusivity on the level of the sentence, similar to models of the acquisition of word meaning such as Fazly, Alishahi & Stevenson (2010) and Siskind (1996). It takes an intermediate position: whereas Fazly et al.'s notion of mutual exclusivity is a soft constraint, Siskind (1996, 43) goes further, stating that no two words may refer to the same part of a situation at all.

I believe Siskind's approach to be too strong: two words *can* refer to the same semantic elements. A verb like *leave* signifies a Source-Path-Goal image schema (Lakoff 1987), and a preposition like *from* does so as well. The fact that both refer to aspects of the same frame, does not preclude language users from using both in the same sentence (*I left from my house this morning*). The **isomorphy** constraint I define is non-probabilistic, but weaker than Siskind's

approach. It only states that the root vertices of the meanings of the used constructions may not map to the same vertex of a situation. In order to combine constructions in our model, we require the two constructions to share one vertex in a s, as we will see in section 3.4.5. Because of this, we need different constructions to be able to stand in a subset-mapping relationship to the same vertex in a situation (as in the case of *leave* and *from*), whereas Siskind's notion of isomorphy would preclude this.

One exception to **isomorphy** is the case in which the head constituent of a construction c is filled with another construction c'. In that case, the root vertex of $sd_{c'}$ points by necessity to the same vertex as the root vertex of sd_c. The other constituents of c and c' still have to obey **isomorphy**. The reason for this exception is that we want to allow for abstract argument structure constructions to be combined with verbs and more generally: for abstract valency patterns (i.e., without a phonological specification of the head constituent) to be combinable with lexical constructions giving a phonological specification of the head.

Unlike for **coherence**, the discrete nature of the **isomorphy** constraint is not self-evident. Exploring a more probabilistic version of isomorphy (in which multiple coverage of the same situational vertex is directly or indirectly penalized) may constitute an interesting future extension of the model.

A final constraint concerning heads is the the **single-dependent-distribution** constraint. This constraint states that the head constituent of a construction cannot be combined with another construction that has the same head, unless it is a lexical construction. This constraint prevents the recursive application of highly abstract constructions early on, which would otherwise lead to spurious bootstrapping behavior. The motivation for this constraint comes from the connection with dependency parsing the model has: given a head, certain patterns of dependents (other constituents) can be selected, but the selection can to be made only once. Cognitively, one could argue that, when selecting a verb, the speaker selects only a single, and not multiple, argument-structure constructions to express that verb with.

3.4.3 Starting a derivation: concatenation

Derivations are built using a set of processing mechanisms that are given to the model before it has any contentive knowledge of the grammatical constructions. As such, they should be considered 'innate' to the model, or at least existing prior to any linguistic input. However, they should be considered to be very general structure-building operations rather than domain-specific rules. The four operations defined by them (concatenation, rule application, ignoring, and bootstrapping) can be seen as general operations on information.

All processing mechanisms are applied to the left-most non-terminal symbol of a current derivation. A derivation starts with the TOP symbol. From a TOP symbol, we can start any number of concatenated derivations:

i TOP $\rightarrow$ START^{+}

The START symbol, then, forms the starting point for the application of pairings of a construction and a subset mapping (c, **map** pairings). We therefore add the following processing mechanism to the set:

ii START $\rightarrow (c, \mathbf{map})$

With mechanism **i**, any number of derivations can be concatenated as long as they obey the **coherence** and **isomorphy** constraints. Mechanism **i** gives the model the robustness to jointly interpret several partial analyses in early stages when it has little linguistic knowledge. As such, it can be seen as a general inferential strategy: the model understands several parts, assumes they are parts of the same message, and so interprets them jointly. Importantly, this processing mechanism remains available to the model throughout development (desideratum D6-4), although its relative importance may decrease.

3.4.4 Ignoring words

Furthermore, this concatenative top rule allows the model to integrate words that it cannot analyze into the derivation. This behavior, too, is needed in early stages, as the model simply does not have constructions to analyze all the words in the utterance. For ignoring words, we define the following rule, given that α is a minimal phonological string (in our case defined as a presegmented word).

iii START $\rightarrow \alpha$

Importantly, any $\alpha \in U$ can be ignored with rule **iii**. This is important in allowing the model the robustness to interpret complex constructions whose constituents are disjunct, i.e., by ignoring the intermediate words (applying rule **iii** for each ignored word).

3.4.5 Applying construction-mapping pairings

When applying a c, **map** pairing with rule **ii**, the constraints on its signifying constituents sr_c have to be satisfied in order to create a legal derivation. The processing mechanism **iv** specifies this, by instructing the model to replace c with its constituents sr_c.

iv $(c, \mathbf{map}) \rightarrow sr_c^1, \ldots sr_c^n$

Satisfying the constraints on each sr_c^i can be done in three ways, depending on the constraints on sr_c^i.

v $sr_c^i \rightarrow \alpha^{+}$ (if $F(sr_c^i) \neq \epsilon$)

vi $sr_c^i \rightarrow \alpha^+$ (if $F(sr_c^i) = \epsilon$)

vii $sr_c^i \rightarrow (c', \mathbf{map}')$ (only if c is not a **lexical** construction)

Rule **v**, firstly, terminates the derivation with any number of terminal nodes. In the generative process, these terminal nodes are specified by the non-empty phonological constraint $F(sr_c^i)$. In parsing, this phonological string α^+ has to be a substring of U. When the phonological constraint is not specified (i.e., $F(sr_c^i) = \epsilon$), we can bootstrap a substring of U into that constituent with rule **vi**. This is another operation, besides the concatenation process of rule **i** and ignoring words with rule **iii**, allowing the model to apply constructions despite not knowing certain lexical constructions.

Rule **vii**, finally, allows the model to fill any constituent of a construction c with another pairing of a construction c' and a subset mapping $\mathbf{map}'$. Apart from having to satisfy the general constraints of **isomorphy** and **coherence**, the new pairing $c', \mathbf{map}'$ has to satisfy the phonological and semantic constraints on sr_c^i.

Satisfying semantic constraints

Satisfying a semantic constraint means that whatever fills a constituent (from a top-down perspective) or whatever is used to recognize a constituent (from a bottom-up perspective) has a meaning that is compatible with the content of the semantic constraint. Recall that a semantic constraint on a signifier $K(sr_c^i)$ is a pointer to a single vertex v in the meaning of c. As such, it can be mapped, via the subset mapping **map** to a vertex in one of the situations.

Semantic constraint satisfaction is defined as the situation in which the root vertex of the meaning of the construction filling a constituent is mapped to the *same* vertex in one of the situations to which the semantic constraint on the constituent is mapped. More formally:

Definition of semantic constraint satisfaction

A semantic constraint $K(sr_c^i)$ of a construction c with a mapping **map** is satisfied by a pairing of a construction and a mapping $c', \mathbf{map}'$ iff

- $\mathbf{map}(K(sr_c^i)) = \mathbf{map}'(v_{\text{root}}(sd_{c'}))$

Satisfying phonological constraints

Satisfying phonological constraints in rule **vii** is a slightly more complex matter. After all, the construction c' itself is not a phonological element. However, if the head constituent of c' terminates into a phonological string α^+ that is identical $F(sr_c^i)$, we consider $F(sr_c^i)$ satisfied.

Formally, phonological constraint satisfaction works as follows:

Definition of phonological constraint satisfaction

A phonological constraint $F(sr_c^i)$ of a construction c with a mapping $\mathbf{map}_{\text{subset}}$ is satisfied by a pairing of a construction and a mapping $c', \mathbf{map}'_{\text{subset}}$ iff

- $F(sr_c^i) = \mathbf{yield}(sr_{c'}^{\text{head}})$, where the yield of a signifier $\mathbf{yield}(sr_c^i)$ is defined as the string of phonological elements α^+ governed by the derivation at sr_c^i.

The motivation for allowing derivations themselves to satisfy phonological constraints is that it allows us, for adult language, to parse modified idioms like *pull some family strings* or *pull political strings*. In those cases, the *strings* is a lexical constituent of a phonologically specified construction [[*pull*] [*strings*]]. I propose that the analysis of *pull political strings* is that the [[*pull*] [*strings*]] construction is combined with something like an [[PROPERTY] [ENTITY]] construction, where the [PROPERTY] constituent is replaced with the lexical element *political*. Similarly, if the child starts out with highly lexically specified constructions, as usage-based theory has it, allowing for the modification of a lexically specified constituent is a desirable feature of the model.

Finally, a special constraint on head constituents sr^{head} of constructions is that rule **vii** can only apply if c' is a lexical construction. I assume that a head can only distribute its roles once, meaning that if a construction is applied in which the dependent constituents of a head constituent are given, the head constituent of this construction cannot be filled with another construction which again gives the dependent constituents of the same head. We call this constraint the **single-dependent-distribution** constraint. One could argue that this constraint is overly strict: if a learner knows an [[AGENT] [*kicks*]] as well as a [[*kicks*] [PATIENT]] construction, why could we not apply both subsequently? It would give the learner more robustness for interpreting full(er) utterances early on, for instance in cases where the learner does not have an [[AGENT] [*kicks*] [PATIENT]] construction, but does have an [[AGENT] [*kicks*]] and a [[*kicks*] [PATIENT]] construction (for a proposal along those lines, see Langacker 2009).

One apparent problem is that this approach would allow for a lot of overgeneration: the head constituent of, say, a transitive construction can be filled with a transitive construction, whose head constituent can be filled with another transitive construction, and so forth. Of course, the **isomorphy** constraint limits this, and in practice it would not pose that much of a problem.

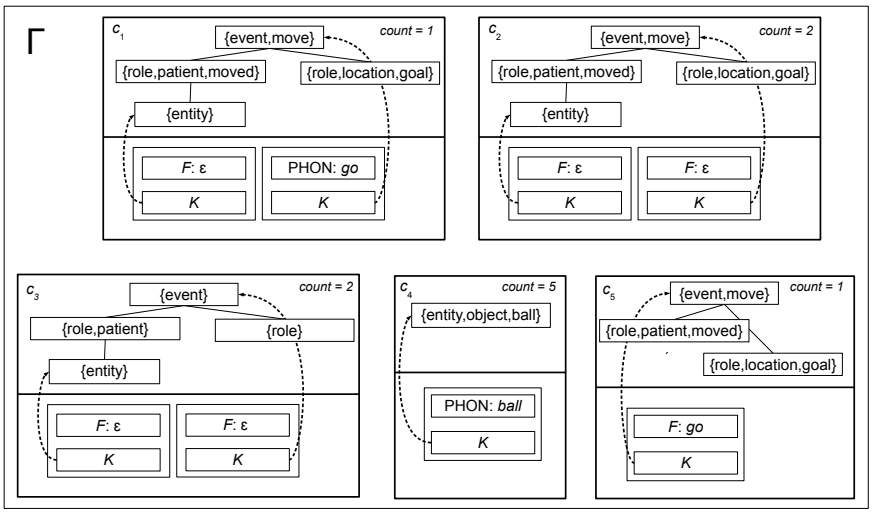


Figure 3.5: A constructicon Γ consisting of 5 constructions.

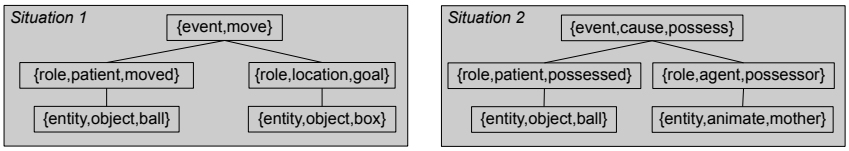
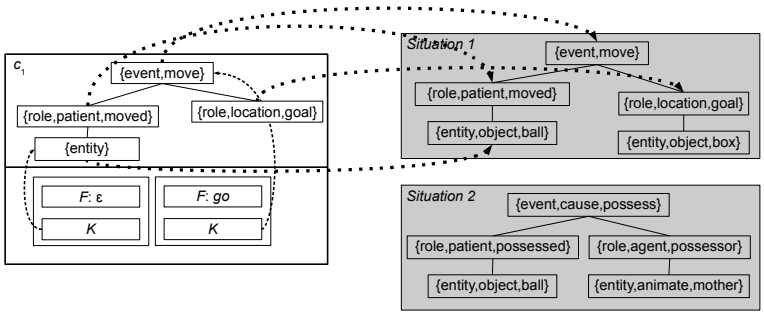


Figure 3.6: Two situations in the input item.

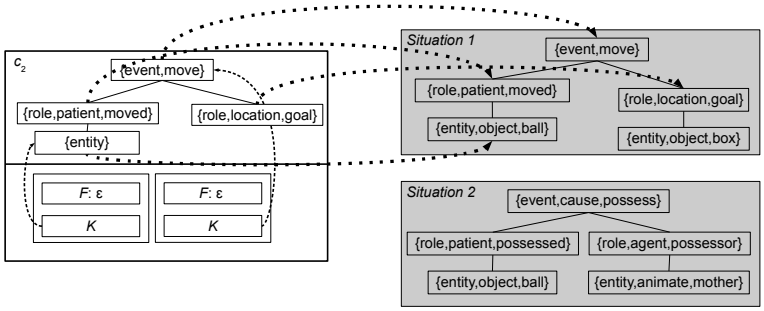
3.4.6 An example of the space of possible derivations

To illustrate the space of possible analyses of an utterance, let us take a look at an example. First, assume the constructicon in figure 3.5. This constructicon consists of five constructions. Let us further assume that the model is trying to create derivations over the utterance $U =$ *ball go there*. The situations S co-present are given in figure 3.6.

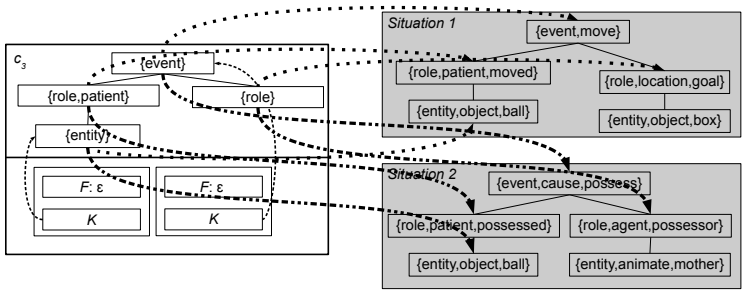
First, all subset mappings between the constructions and subgraphs of the situations are retrieved. Figure 3.7 gives all subset mappings for the five constructions and the two situations. Constructions c_1 and c_2 each have one mapping to situation s_1. Construction c_3, being more abstract, has two mappings: one to s_1 (let us call it $\mathbf{map}_1(c_3)$), and one to s_2 ($\mathbf{map}_2(c_3)$), as has construction c_4 ($\mathbf{map}_1(c_4)$ and $\mathbf{map}_2(c_4)$). The lexical construction c_5, finally, just has one



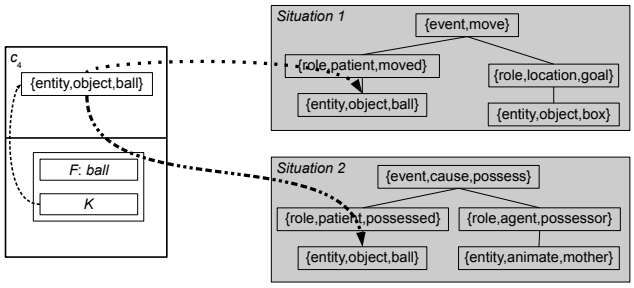
(a) The mapping between c_1 and the situations.



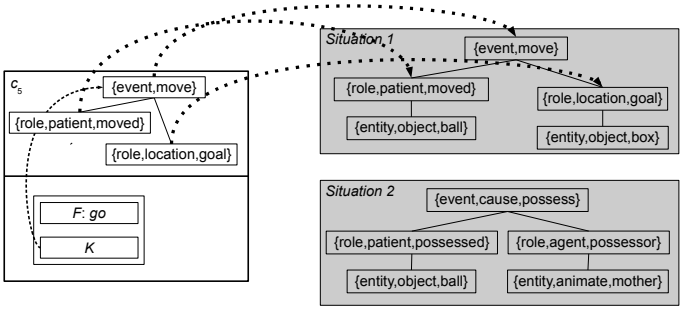
(b) The mapping between c_2 and the situations.



(c) The mapping between c_3 and the situations.



(d) The mapping between c_4 and the situations.



(e) The mapping between c_5 and the situations.

Figure 3.7: The mappings between the constructions in the constructicon and the situations.

subset mapping.

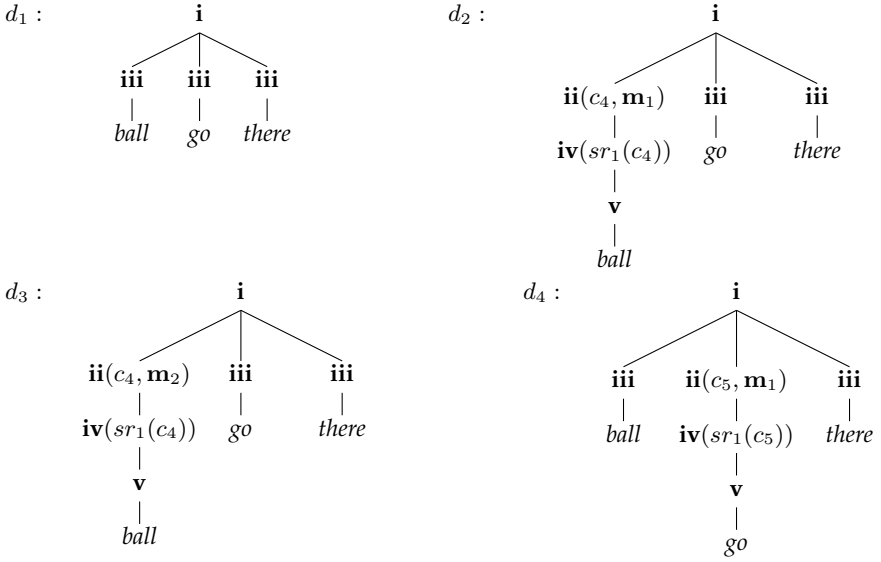


Figure 3.8: Derivations $d_1 - d_4$ for *ball go there*.

Which derivations are possible given this set of construction-situation mappings and the eleven processing mechanisms? Firstly, in the most trivial case, d_1, we ignore all words by applying rule **i** with an arity of three, followed by three times rule **iii**, with which we ignore a word. In the next three derivations, d_2-d_4, we apply one grammatical construction with rule **ii** and ignore all other words. Rule **ii** applies a c, **map** pair (represented as $(c, \mathbf{m}_i)$), which then splits into the constituents of c with rule **iv**. Because the single constituent of constructions c_4 and c_5 is phonologically specified and can be retrieved from U with rule **vi**, the derivation is valid. Note that in the case of d_2 and d_4, the construction is mapped to elements of situation s_1, and in the case of d_3 to s_2.

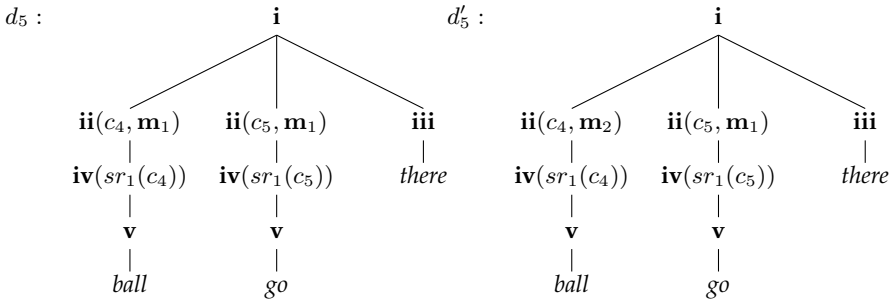


Figure 3.9: Derivations d_5 and d_5' for *ball go there*.

Concatenating the constructions c_4 and c_5 is also possible. Derivation d_5 exemplifies this: rule **i** is applied with an arity of three, after which constructions c_4 and c_5 are inserted with rule **ii**, and the final word is ignored with rule **iii**. Note that the derivation in d_5' is illegal: as c_4 is mapped to situation s_2 via $\mathbf{map}_2$ and c_5 to situation s_1 via $\mathbf{map}_1$, the **coherence** constraint is violated, rendering this derivation invalid.

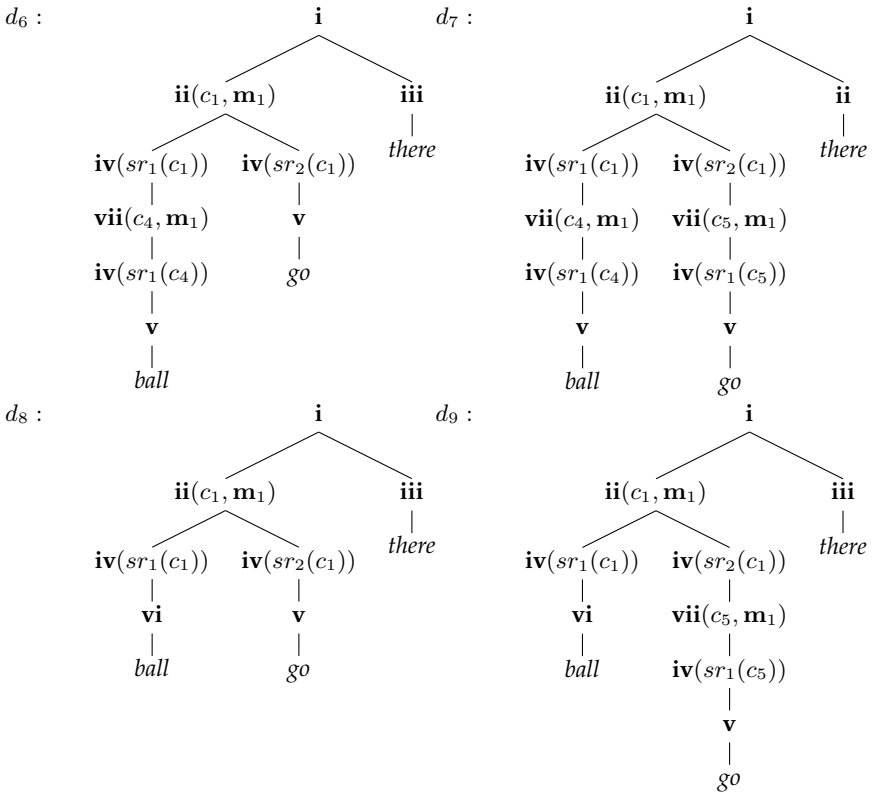


Figure 3.10: Derivations $d_6 - d_9$ for *ball go there*.

Then there are four derivations in which construction c_1 is applied and *there* is ignored. In the first two cases, d_6 and d_7, the open constituent of c_1 is filled with the pairing $c_4, \mathbf{m}_1$. In the latter two, d_8 and d_9, the word *ball* is bootstrapped by directly terminating the phonological open constituent $sr_1(c_1)$ with rule **vi**. Secondly, in d_6 and d_8 rule **vi** is applied to the recognition of the word *go*, whereas in d_7 and d_9 the second constituent of c_1 is filled with $c_5, \mathbf{m}_1$ via rule **vii**, which then terminates in the word *go*.

Construction c_2, with two open constituents, allows for more derivations. Derivation d_{10} gives the case in which c_2 is combined with c_4 and c_5 and the word *there* is ignored. However, we can also bootstrap either ($d_{11} - d_{14}$) or both

($d_{15} - d_{19}$) constituents.

Note that, although in principle construction c_1 could be combined with the second constituent of c_2, $sr_2(c_2)$, the **isomorphy** constraint precludes this. The combination of c_5 with $sr_2(c_2)$ *is* legal: although both $c_2, \mathbf{m}_1$ and $c_5, \mathbf{m}_1$ have root nodes mapped to the {EVENT,MOVE} vertex of situation s_1, $sr_2(c_2)$ is the head constituent of c_2, and hence this situation is exempt to **isomorphy**.

Finally, construction c_3 allows for even more derivations. As it has two mappings and is fully phonologically unspecified, many derivations involving bootstrapped constituents can be made. Below all 19 derivations can be found ($d_{20} - d_{37}$) in figures 3.13 and 3.14.

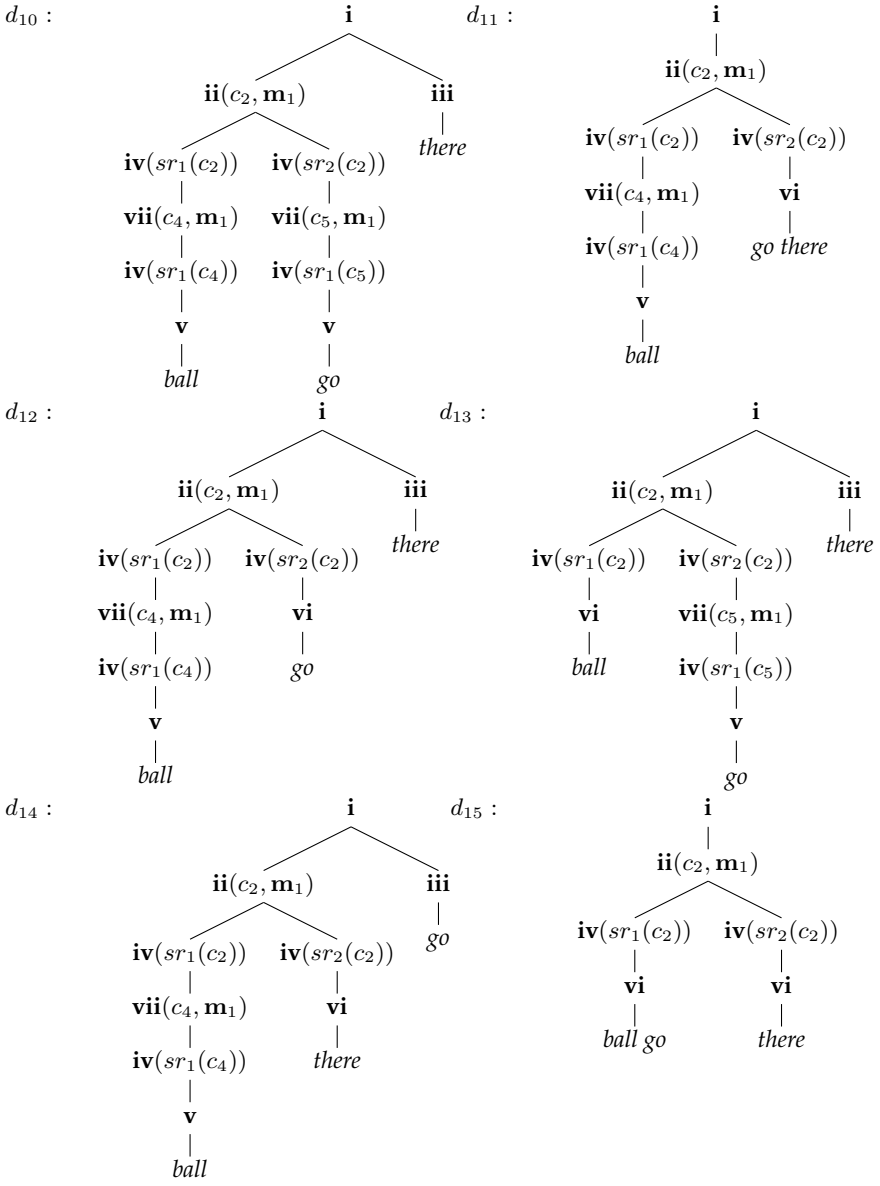


Figure 3.11: Derivations $d_{10} - d_{15}$ for *ball go there*.

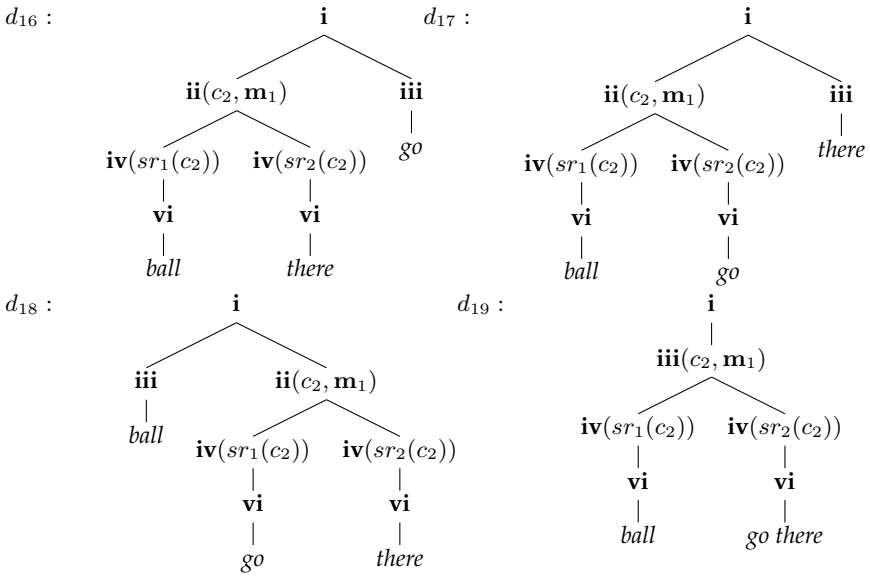


Figure 3.12: Derivations $d_{16} - d_{19}$ for *ball go there*.

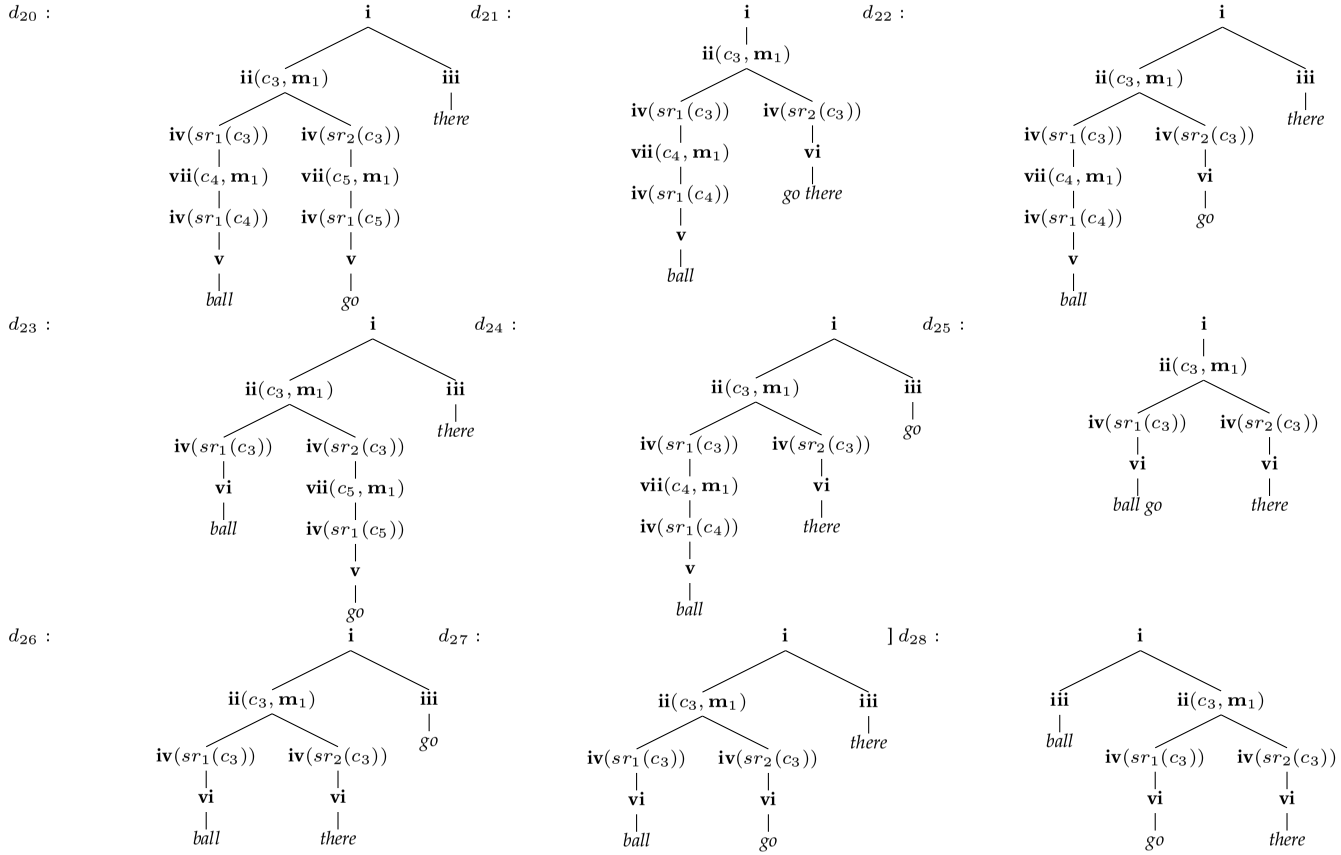


Figure 3.13: Derivations $d_{20} - d_{28}$ for *ball go there*.

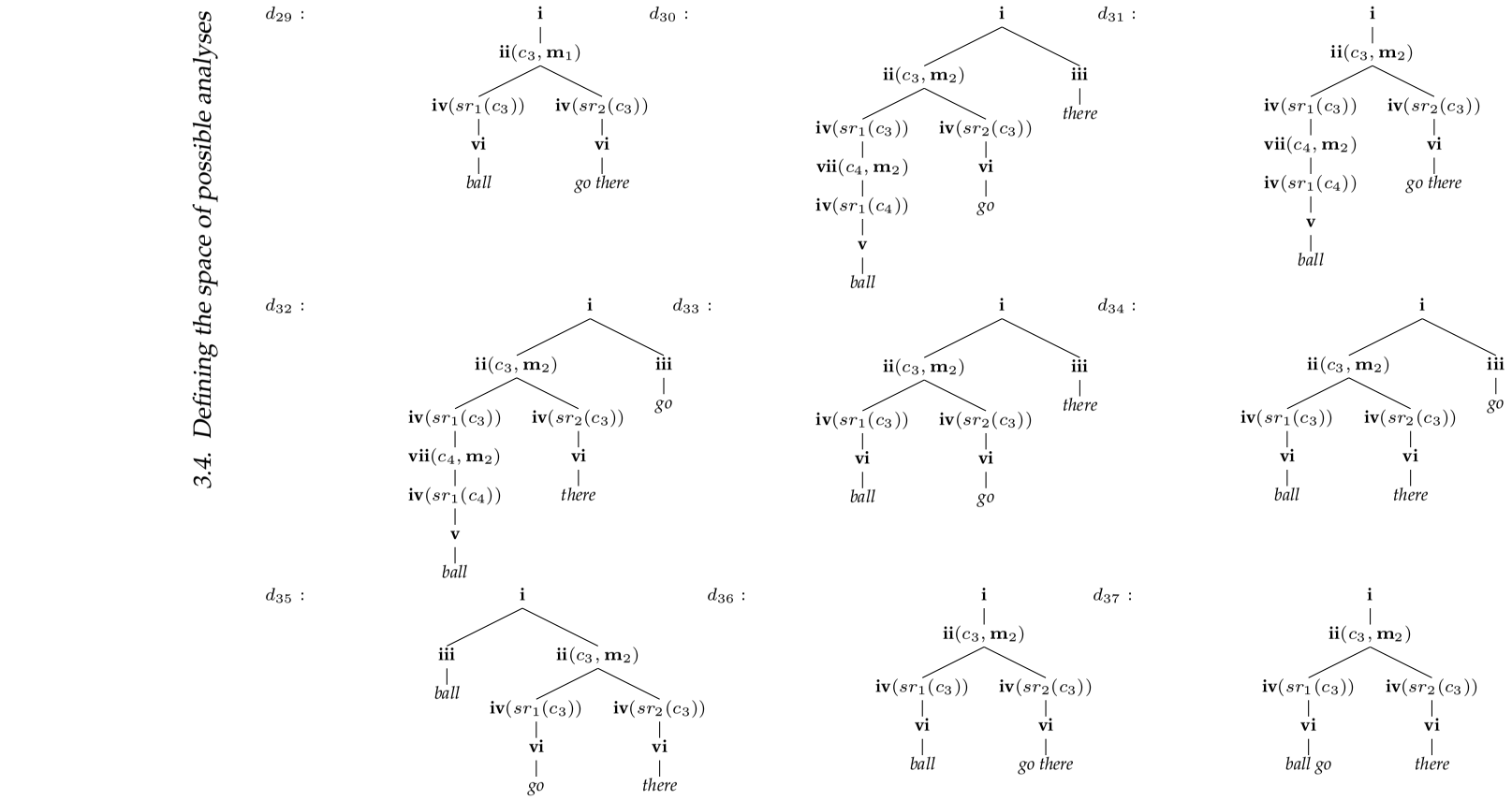


Figure 3.14: Derivations $d_{29} - d_{37}$ for *ball go there*.

	rule	probability
i	TOP $\rightarrow$ START$^+$	$\frac{1}{2}^{\lvert\text{START}^+\rvert}$
ii	START $\rightarrow (c, \mathbf{map})$	$P(c, \mathbf{map} \lvert CS_{\text{START}})$
iii	START $\rightarrow \alpha$	$P(\mathbf{u} \lvert CS_{\text{START}})$
iv	$(c, \mathbf{map}) \rightarrow sr_1(c), \ldots, sr_n(c)$	1
v	$sr_c^i \rightarrow \alpha^+$ (if $F(sr_c^i) \neq \epsilon$)	1
vi	$sr_c^i \rightarrow \alpha^+$ (if $F(sr_c^i) = \epsilon$)	$P(\mathbf{u} \lvert CS_{sr_c^i}) \cdot \frac{1}{2}^{\lvert\alpha^+\rvert} \cdot P(\mathbf{u} \lvert CS_{\text{START}})^{\lvert\alpha^+\rvert}$
vii	$sr_c^i \rightarrow (c', \mathbf{map}')$	$P(c', \mathbf{map}' \lvert CS_{sr_c^i})$

Table 3.1: Probabilities of the processing mechanisms in the analysis procedure.

3.5 Selecting the best analysis

The six processing mechanisms, along with the construction-mapping pairings used in them, will typically lead to a situation in which many derivations are possible, as we have seen in the example above. I assume that the learner selects a single best analysis among those different analyses. In this section, I describe the process for doing so, and the actual implementation, which makes the model more realistic in its processing of the utterance.

3.5.1 The probability model for derivations

We can consider the branching process defined by the seven processing mechanisms to be a probabilistic process, where the application of a rule at a point in the derivation has a certain probability of occurring given that point in the derivation. Some of these probabilities are fixed, whereas others change as the state of linguistic knowledge of the learner progresses. Table 3.1 gives the probabilities of the processing mechanisms, which will be explained below.

$P(c, \mathbf{map}|CS)$ in mechanisms **ii**, **vi**, and **vii** is defined as the smoothed relative frequency of the construction c out of all $c, \mathbf{map}$ pairings that can be applied at that point in the derivation, i.e., that compete with $c, \mathbf{map}$ for being applied. We call the set of all applicable $c, \mathbf{map}$ pairings at some point x the **competition set** given x, or CS_x. Formally, a competition set is defined as follows:

$$CS_{\text{START}} = \forall(c_{c\in\Gamma}, \mathbf{map}).\mathbf{map}(sd_c) = s \in S \tag{3.1}$$

$$CS_{sr^i_c} = \forall(c'_{c'\in\Gamma}, \mathbf{map}').\mathbf{map}'(v_{\text{root}}(sd_{c'})) = \mathbf{map}(K(sr^i_c)) \tag{3.2}$$

That is to say: given the START symbol, all pairings of a construction in the constructicon and a legal subset mapping compete with each other. Next, given a constituent of a construction sr^i_c, all pairings of a construction c' and a mapping $\mathbf{map}'$ for which the root vertex of the meaning of c' refers to the same vertex in a situation as the conceptual constraint of sr^i_c.

The smoothed relative frequency of the construction-mapping pairing is then defined as:

$$P(c, \mathbf{map}|CS) = \frac{count_c + 1}{\sum_{c', \mathbf{map}' \in CS} (count_{c'} + 1) + 1} \tag{3.3}$$

The probability of an unseen event **u**, applied in rules **iii** and **v**, is given by the remaining probability mass given a competition set CS, i.e.:

$$P(\mathbf{u}|CS) = \frac{1}{\sum_{c, \mathbf{map} \in CS} (count_c + 1) + 1} \tag{3.4}$$

Motivating the probability of rule i In the concatenation process of rule **i**, we set the probability of concatenating n derivations to $\frac{1}{2}^n$, that is: the more derivations are concatenated, the lower the probability of the overall derivation. This probability can be seen as a prior on the length of the concatenation, while, at the same time, it ensures that the probabilities of all generations given the constructicon and all possible situations sum to 1.

Motivating the probabilities of rules ii and iii Rule **ii** involves the application of a $c, \mathbf{map}$ pairing given the START symbol. This means that any construction, with any possible mapping to a situation in S can be applied. The competition set (as given in equation (3.1)) thus consists of all these construction-mapping pairings. The probability of selecting the pairing $c, \mathbf{map}$ out of all possible pairings is given by the smoothed relative frequency of c out of all applicable pairings. The fact that the probabilities are based on the counts of the constructions reflects desideratum D2-3, viz. the idea that the representational strength of the representations or their ease of retrieval should be grounded in their frequency of use.

Ignoring a word with rule **iii**, then, involves *not* selecting any construction-mapping pairing. That is: the model considers ignoring a word to be an unseen event **u**, and the remainder of the probability mass given CS_{START}, as defined in equation (3.4) is applied. Importantly, the probability of ignoring a word goes down as the size of the part of the constructicon that can be applied to

the current input item grows. This means that the more the model has learned, the smaller the probability of ignoring a word becomes.

Motivating the probability of rule iv Rule **iv** can be considered a dummy rule that expands the c of some $c, \mathbf{map}$ pairing into its signifying constituents sr_c. Because it can be trivially applied after rules **ii** and **vii**, I assign it a probability mass of 1

Motivating the probabilities of rules v, vi, and vii When substituting a signifying constituent sr_c^i of a construction for another element, several things can happen. Firstly, if sr_c^i is phonologically specified (i.e., if $F(sr_c^i) \neq \epsilon$), we can terminate the derivation directly into the phonological structure given by $F(sr_c^i)$ with rule **v**. In that case, the termination has a probability of 1.

Regardless of whether the phonological constraint on sr_c^i is specified, we can combine it with other $c, \mathbf{map}$ pairings with rule **vii**. Again, any $c, \mathbf{map}$ pairing applied at this point in the derivation stands in competition with all $c, \mathbf{map}$ pairings that can be applied at that point, that is: all $c, \mathbf{map}$ pairings that satisfy the phonological and semantic constraints on sr_c^i. The probability of applying a construction-mapping pairing $c', \mathbf{map}'$ thus is the smoothed relative frequency of c' out of all $c, \mathbf{map}$ pairings that can be used to fill sr_c^i (i.e., $CS_{sr_c^i}$), as given in equation (3.3). Again, this aspect of the probability model is grounded in desideratum D2-3, the idea that representational strength is grounded in the frequency of use.

Finally, if the phonological constraint on sr_c^i is empty, we may nonetheless terminate the derivation into a string of phonological elements α^+ with rule **vi**. This is the bootstrapping operation described earlier. The bootstrapping operation competes with all construction-mapping pairings that are applicable given sr_c^i, and, as with ignoring words, we assign it the remainder of the probability mass of $c', \mathbf{map}'$ pairings given sr_c^i. However, as the bootstrapped string α^+ can be of any length, it is undesirable if bootstrapped phonological strings of any length are equiprobable. This would lead the model to bootstrapping very long phonological strings too eagerly. Therefore, we apply the same principle as in the concatenation process of rule **i** to assign a quadratically decreasing probability over the length of the phonological string. Finally, we consider all elements α in α^+ to be ignored elements, and therefore multiply $P(\mathbf{u}|CS_{sr_c^i}) \cdot \frac{1}{2}^{|\alpha^+|}$ with the number of times rule **iii** would be applied if it was a regular 'ignore' operation, that is: with $P(\mathbf{u}|CS_{\text{START}})^{|\alpha^+|}$.

The probability of a derivation

The probability of a derivation can now be defined as the joint probability of all applications of the mechanisms in the derivation process. That is, $P(d|\Gamma, S)$ is the product of the probabilities of all rules r applied in it, as defined in table 3.1:

$$P(d|\Gamma, S) = \prod_{r \in d} P(r) \tag{3.5}$$

3.5.2 Equivalent derivations: parses

The 38 derivations we saw in section 3.4.6 give rise to different interpretations: d_3 and $d_{30} - d_{37}$ refer to situation s_2, d_1 to no situation, and the remaining derivations to s_1. Also within the groups of parses referring to the same situation, there is variation as to which parts of the utterance and the inferred linguistic structure point to which parts of the situation.

Under the usage-based assumption that linguistic knowledge can be redundantly stored at several levels of abstraction (Beekhuizen, Bod & Zuidema 2013), the model will apply constructions at varying levels of abstraction when analyzing an utterance. Several of these, however, have an identical derivational structure and refer in the same way to the same aspects of a situation. Therefore, for the purposes of analyzing an utterance they can be considered identical. We define DERIVATIONAL IDENTITY as follows:

Definition of DERIVATIONAL IDENTITY

Derivations d_1 and d_2 are derivationally identical iff

- $\text{rules}(d_1) = \text{rules}(d_2)$
- $\forall r_{i:r_i \in \text{rules}(d_1)}, r_{j:r_j \in \text{rules}(d_2)}.\mathbf{map}_i(c_i) = \mathbf{map}_j(c_j)$

Given this definition, d_1 and d_2 first have to satisfy the constraint that the strings of rules $\text{rules}(d_1)$ and $\text{rules}(d_2)$ be equal, that is: the same rules are applied in the same order. The $c, \mathbf{map}$ pairings applied in these strings of rules may, however, differ. At the same point in a derivation, the model can sometimes apply different construction-mapping pairings. Now, if for two strings of rules each mapping $\mathbf{map}$ applied at a certain point in d_1 has the same set of vertices in its codomain (the subgraph of a situation $s \in S$) as the mapping $\mathbf{map}'$ applied at the parallel point in derivation d_2, we consider the two derivations to be equal.

We define a parse or analysis a as the set of derivations that are derivationally identical to each other, and the set A as all parses given the utterance, the situations, and the constructicon. The probability of a parse can then be defined as the probability of either of the derivations subsumed by that parse being generated by the constructicon given the situation:

$$P(a|\Gamma, S) = \sum_{d \in a} P(d) \tag{3.6}$$

d_{30} :

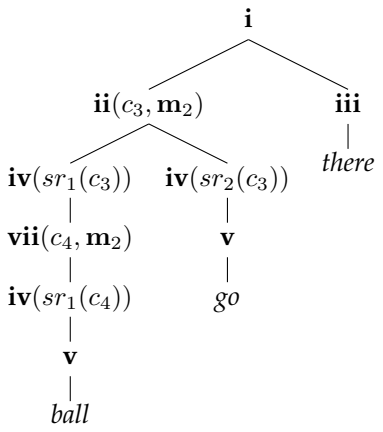


Figure 3.15: Derivation d_{30} for *Ball go there*.

Parallel c, **map** pairings in various derivations of a parse stand in a parent-child relationship to each other. As the derivational structure of the various derivations is identical, the constructions used should have the same number and types of constituents (otherwise the tree structure and choice of processing mechanisms would be different), and given the mapping equivalence, they should have meanings that are supersets or subsets of each other. We can relate this to Langacker's notion of immanence: the various derivations are not distinct events, but are all activation patterns over the same traces of linguistic usage events. The advantage of using both the parents and child constructions in the same parse is that we allow abstract constructions to back-up more concrete ones. This can be seen as a form of multiple licensing, albeit a very simple one (cf. Kay 2002)

The best parse a_{best} then, is taken to be the most-probable one.

$$a_{\text{best}} = \arg\max_{a \in A} P(a|\Gamma, S) \tag{3.7}$$

The situation mapped to by a_{best} is the identified situation $s_{\text{identified}}$, that is: the situation SPL thinks the speaker refers to. If the best parse has no mapping to any situation $s \in S$, for instance in the case when all words are ignored, one situation is selected at random to be the interpretation of the utterance. If multiple analyses are equally likely, one is selected at random to be a_{best}.

3.5.3 An example of the probability model

A single derivation

With the counts of the constructions, as given in figure 3.5, we can calculate the probabilities of all derivations. Let us look at derivation d_{30}, repeated here as figure 3.15. By substituting the left-most open symbol every time, we can order the rules as follows:

- **i**, **ii**$(c_3, \mathbf{m}_2)$, **iv**$(sr_1(c_3))$, **vii**$(c_4, \mathbf{m_2})$, **iv**$(sr^i_{c_4})$, **v**, **iv**$(sr_2(c_3))$, **vi**, **iii**

Rule **i**, applied with an arity of 2, has a probability of $\frac{1}{2}^2 = \frac{1}{4}$. Applying rule **ii** to the pairing $c_3, \mathbf{map}_2$ requires us to consider the competing construction-mapping pairings. Given the START symbol, this means we consider the competition set CS_{START}. As CS_{START} contains all possible construction-mapping pairings, it consists of $\{(c_1, \mathbf{map}_1), (c_2, \mathbf{map}_1), (c_3, \mathbf{map}_1), (c_3, \mathbf{map}_2), (c_4, \mathbf{map}_1), (c_4, \mathbf{map}_2), (c_5, \mathbf{map}_1)\}$. The probability of selecting $(c_3, \mathbf{map}_2)$ out of this competition set, or its smoothed relative frequency, is 3 (the count of c_3 plus one) over the sum of all smoothed frequencies of the elements in the competition set, plus one, or $(1+1)+(2+1)+(2+1)+(2+1)+(5+1)+(5+1)+(2+1)+1$:

$$P(\mathbf{ii}) = \frac{3}{27} \tag{3.8}$$

Next, the application of rule **iv** has a probability of 1. Applying rule **vii** afterwards again requires us to consider the competition set of the selected $c, \mathbf{map}$ pairing. In this case $c_4, \mathbf{map}_2$ is selected. The set of $c, \mathbf{map}$ pairings referring to the vertex {ENTITY, OBJECT, BALL} in situation 2 consists only of $c_4, \mathbf{map}_2$ itself, and the probability of selecting this pairing is

$$P(\mathbf{vii}) = \frac{5+1}{(5+1)+1} = \frac{6}{7} \tag{3.9}$$

The subsequent applications of rule **iv** and **v** each have a probability of 1, in the case of rule **v** because the phonological constituent of the first constituent of c_4 is specified.

After having terminated the first constituent of c_3, we look at the second constituent. Again, rule **iv** is applied with a probability of 1. After this application, *go* is bootstrapped into the constituent slot. The second constituent of c_3 has no phonological specification, and hence we take the second equation for rule **v**. This requires us to get the competition set for the second constituent, which consists of only the pairing $c_3, \mathbf{map}_2$ itself,[3] as well as for CS_{START}, which we saw in the application of rule **ii** before. The probability of an unseen event given $CS_{sr_2(c_3)}$ is 1 over 4 ($2+1$ for c_3, and 1 to smooth). The

[3]Note that the application of this pairing is ruled out by the **single-dependent-distribution** constraint, which in this case specifies that whatever fills the second constituent must be a lexical construction.

probability of an unseen event given CS_{START} is $\frac{1}{27}$, given that the denominator given this competition set is 27, as we have seen in the application of rule **ii** before. The $P(\mathbf{u}|CS_{\text{START}})$ is applied once, as the length of the phonological string is $|\alpha^+| = 1$. Similarly, the probability constraining the length of the concatenation $\frac{1}{2}$ is also raised to the power $|\alpha^+| = 1$, giving us the following probability

$$P(\mathbf{vi}) = \frac{1}{(2+1)+1} \cdot \frac{1}{2}^{1} \cdot \frac{1}{27}^{1} = \frac{1}{216} \tag{3.10}$$

Finally, we apply rule **iii** in order to ignore the word *there*. This amounts to an instance of an unseen event given the START symbol, which we have seen before, viz. $P(\mathbf{iii}) = \frac{1}{27}$. Table 3.2 below gives the probabilities of all derivations. I leave the calculation of the individual probabilities of the mechanisms as an exercise to the reader.

Several things can be learned from this example. First of all, because of the probability model, bootstrapping two adult words or bootstrapping one and ignoring one are equiprobable. Derivations d_{30} and d_{31} illustrate this.

Second, not every bootstrapping operation is equally likely. The higher the frequencies of the items in the competition set, the lower the probability of bootstrapping an element into it. Derivations d_{22} and d_{23} show this effect: because a highly frequent construction (c_4) can be fit into the first signifier, bootstrapping it becomes less likely. c_5 has a lower count, and hence bootstrapping the second constituent is relatively more likely, resulting in a probability of d_{22} that is twice as high as that of d_{23}. This effect can be seen as a pragmatic line of reasoning: I know some construction to be applicable given the constituent and the situation, so if that's a very likely construction, it is unlikely that the speaker would use a novel element to express it.

Third, we can see that the most likely derivations are those in which c_1 is used. This is a semi-open schema, with the phonological element *go* specified on the second constituent. As such, less rules have to be applied in order to arrive at a full derivations, and because of this, derivations with c_1 are globally more likely than those with c_2 and c_3, despite c_1 having a lower count than either c_2 or c_3. The most likely derivation is d_6 ($P(d_6) = \frac{1}{1701}$), in which c_1 is combined with c_4, and the last word is ignored. Here we see an effect akin to statistical pre-emption, which I will explore later in this thesis, namely that derivations with more concrete constructions use fewer rules and are thereby often more likely. It follows, however, from the general probability model and the rules, and is as such not a special built-in feature of the model.

Getting equivalent derivations

As discussed in section 3.5.2, we first look for all parses, that is: sets of derivations that are created by the same processing mechanisms, and for which every node in one derivation has the same subset mapping to a subgraph of a

derivation		P
d_1	$\frac{1}{2}^3 \cdot \frac{1}{27} \cdot \frac{1}{27} \cdot \frac{1}{27} =$	$\frac{1}{157{,}464}$
d_2	$\frac{1}{2}^3 \cdot \frac{6}{27} \cdot 1 \cdot \frac{1}{27} \cdot \frac{1}{27} = \frac{6}{157{,}464} =$	$\frac{1}{26{,}244}$
d_3	$\frac{1}{2}^3 \cdot \frac{6}{27} \cdot 1 \cdot \frac{1}{27} \cdot \frac{1}{27} = \frac{6}{157{,}464} =$	$\frac{1}{26{,}244}$
d_4	$\frac{1}{2}^3 \cdot \frac{1}{27} \cdot \frac{3}{27} \cdot 1 \cdot \frac{1}{27} = \frac{3}{157{,}464} =$	$\frac{1}{52{,}488}$
d_5	$\frac{1}{2}^3 \cdot \frac{6}{27} \cdot 1 \cdot \frac{3}{27} \cdot 1 \cdot \frac{1}{27} = \frac{18}{157{,}464} =$	$\frac{1}{8748}$
d_6	$\frac{1}{2}^2 \cdot \frac{2}{27} \cdot 1 \cdot \frac{6}{7} \cdot 1 \cdot 1 \cdot 1 \cdot 1 \cdot \frac{1}{27} = \frac{12}{20{,}412} =$	$\frac{1}{1701}$
d_7	$\frac{1}{2}^2 \cdot \frac{2}{27} \cdot 1 \cdot \frac{6}{7} \cdot 1 \cdot 1 \cdot 1 \cdot \frac{2}{12} \cdot 1 \cdot 1 \cdot \frac{1}{27} = \frac{24}{489{,}888} =$	$\frac{1}{20{,}412}$
d_8	$\frac{1}{2}^2 \cdot \frac{2}{27} \cdot 1 \cdot \frac{1}{7} \cdot 1 \cdot 1 \cdot \frac{1}{27} = \frac{2}{40{,}824} =$	$\frac{1}{20{,}412}$
d_9	$\frac{1}{2}^2 \cdot \frac{2}{27} \cdot 1 \cdot \frac{1}{7} \cdot 1 \cdot \frac{2}{12} \cdot 1 \cdot 1 \cdot \frac{1}{27} = \frac{4}{489{,}888} =$	$\frac{1}{122{,}472}$
d_{10}	$\frac{1}{2}^2 \cdot \frac{3}{27} \cdot 1 \cdot \frac{6}{7} \cdot 1 \cdot 1 \cdot 1 \cdot \frac{3}{12} \cdot 1 \cdot 1 \cdot \frac{1}{27} = \frac{54}{244{,}944} =$	$\frac{1}{4536}$
d_{11}	$\frac{1}{2}^1 \cdot \frac{3}{27} \cdot 1 \cdot \frac{6}{7} \cdot 1 \cdot 1 \cdot 1 \cdot \frac{1}{34{,}992} = \frac{18}{13{,}226{,}976} =$	$\frac{1}{734{,}832}$
d_{12}	$\frac{1}{2}^2 \cdot \frac{3}{27} \cdot 1 \cdot \frac{6}{7} \cdot 1 \cdot 1 \cdot 1 \cdot \frac{1}{648} \cdot \frac{1}{27} = \frac{18}{13{,}226{,}976} =$	$\frac{1}{734{,}832}$
d_{13}	$\frac{1}{2}^2 \cdot \frac{3}{27} \cdot 1 \cdot \frac{1}{378} \cdot 1 \cdot \frac{3}{12} \cdot 1 \cdot 1 \cdot \frac{1}{27} = \frac{9}{13{,}226{,}976} =$	$\frac{1}{1{,}469{,}664}$
d_{14}	$\frac{1}{2}^2 \cdot \frac{3}{27} \cdot 1 \cdot \frac{1}{378} \cdot 1 \cdot \frac{3}{12} \cdot 1 \cdot 1 \cdot \frac{1}{27} = \frac{9}{13{,}226{,}976} =$	$\frac{1}{1{,}469{,}664}$
d_{15}	$\frac{1}{2}^1 \cdot \frac{3}{27} \cdot 1 \cdot \frac{1}{20{,}412} \cdot 1 \cdot \frac{1}{648} = \frac{3}{714{,}256{,}704} =$	$\frac{1}{238{,}085{,}568}$
d_{16}	$\frac{1}{2}^2 \cdot \frac{3}{27} \cdot 1 \cdot \frac{1}{378} \cdot 1 \cdot \frac{1}{648} \cdot \frac{1}{27} = \frac{3}{714{,}256{,}704} =$	$\frac{1}{238{,}085{,}568}$
d_{17}	$\frac{1}{2}^2 \cdot \frac{3}{27} \cdot 1 \cdot \frac{1}{378} \cdot 1 \cdot \frac{1}{648} \cdot \frac{1}{27} = \frac{3}{714{,}256{,}704} =$	$\frac{1}{238{,}085{,}568}$
d_{18}	$\frac{1}{2}^2 \cdot \frac{1}{27} \cdot \frac{3}{27} \cdot 1 \cdot \frac{1}{378} \cdot 1 \cdot \frac{1}{648} = \frac{3}{714{,}256{,}704} =$	$\frac{1}{238{,}085{,}568}$
d_{19}	$\frac{1}{2}^1 \cdot \frac{3}{27} \cdot 1 \cdot \frac{1}{20{,}412} \cdot 1 \cdot \frac{1}{648} = \frac{3}{714{,}256{,}704} =$	$\frac{1}{238{,}085{,}568}$
d_{20}	$\frac{1}{2}^2 \cdot \frac{3}{27} \cdot 1 \cdot \frac{6}{7} \cdot 1 \cdot 1 \cdot 1 \cdot \frac{3}{12} \cdot 1 \cdot 1 \cdot \frac{1}{27} = \frac{54}{244{,}944} =$	$\frac{1}{4536}$
d_{21}	$\frac{1}{2}^1 \cdot \frac{3}{27} \cdot 1 \cdot \frac{6}{7} \cdot 1 \cdot 1 \cdot 1 \cdot \frac{1}{34{,}992} = \frac{18}{13{,}226{,}976} =$	$\frac{1}{734{,}832}$
d_{22}	$\frac{1}{2}^2 \cdot \frac{3}{27} \cdot 1 \cdot \frac{6}{7} \cdot 1 \cdot 1 \cdot 1 \cdot \frac{1}{648} \cdot \frac{1}{27} = \frac{18}{13{,}226{,}976} =$	$\frac{1}{734{,}832}$
d_{23}	$\frac{1}{2}^2 \cdot \frac{3}{27} \cdot 1 \cdot \frac{1}{378} \cdot 1 \cdot \frac{3}{12} \cdot 1 \cdot 1 \cdot \frac{1}{27} = \frac{9}{13{,}226{,}976} =$	$\frac{1}{1{,}469{,}664}$
d_{24}	$\frac{1}{2}^2 \cdot \frac{3}{27} \cdot 1 \cdot \frac{1}{378} \cdot 1 \cdot \frac{3}{12} \cdot 1 \cdot 1 \cdot \frac{1}{27} = \frac{9}{13{,}226{,}976} =$	$\frac{1}{1{,}469{,}664}$
d_{25}	$\frac{1}{2}^1 \cdot \frac{3}{27} \cdot 1 \cdot \frac{1}{20{,}412} \cdot 1 \cdot \frac{1}{648} = \frac{3}{714{,}256{,}704} =$	$\frac{1}{238{,}085{,}568}$
d_{26}	$\frac{1}{2}^2 \cdot \frac{3}{27} \cdot 1 \cdot \frac{1}{378} \cdot 1 \cdot \frac{1}{648} \cdot \frac{1}{27} = \frac{3}{714{,}256{,}704} =$	$\frac{1}{238{,}085{,}568}$
d_{27}	$\frac{1}{2}^2 \cdot \frac{3}{27} \cdot 1 \cdot \frac{1}{378} \cdot 1 \cdot \frac{1}{648} \cdot \frac{1}{27} = \frac{3}{714{,}256{,}704} =$	$\frac{1}{238{,}085{,}568}$
d_{28}	$\frac{1}{2}^2 \cdot \frac{1}{27} \cdot \frac{3}{27} \cdot 1 \cdot \frac{1}{378} \cdot 1 \cdot \frac{1}{648} = \frac{3}{714{,}256{,}704} =$	$\frac{1}{238{,}085{,}568}$
d_{29}	$\frac{1}{2}^1 \cdot \frac{3}{27} \cdot 1 \cdot \frac{1}{20{,}412} \cdot 1 \cdot \frac{1}{648} = \frac{3}{714{,}256{,}704} =$	$\frac{1}{238{,}085{,}568}$
d_{30}	$\frac{1}{2}^2 \cdot \frac{3}{27} \cdot 1 \cdot \frac{6}{7} \cdot 1 \cdot 1 \cdot \frac{1}{162} \cdot \frac{1}{27} = \frac{18}{3{,}306{,}744} =$	$\frac{1}{183{,}708}$
d_{31}	$\frac{1}{2} \cdot \frac{3}{27} \cdot 1 \cdot \frac{6}{7} \cdot 1 \cdot 1 \cdot \frac{1}{8748} = \frac{18}{3{,}306{,}744} =$	$\frac{1}{183{,}708}$
d_{32}	$\frac{1}{2}^2 \cdot \frac{3}{27} \cdot 1 \cdot \frac{6}{7} \cdot 1 \cdot 1 \cdot \frac{1}{162} \cdot \frac{1}{27} = \frac{18}{3{,}306{,}744} =$	$\frac{1}{183{,}708}$
d_{33}	$\frac{1}{2}^2 \cdot \frac{3}{27} \cdot 1 \cdot \frac{1}{378} \cdot 1 \cdot \frac{1}{162} \cdot \frac{1}{27} = \frac{18}{178{,}564{,}176} =$	$\frac{1}{9{,}920{,}232}$
d_{34}	$\frac{1}{2}^2 \cdot \frac{3}{27} \cdot 1 \cdot \frac{1}{378} \cdot 1 \cdot \frac{1}{162} \cdot \frac{1}{27} = \frac{18}{178{,}564{,}176} =$	$\frac{1}{9{,}920{,}232}$
d_{35}	$\frac{1}{2}^2 \cdot \frac{1}{27} \cdot \frac{3}{27} \cdot 1 \cdot \frac{1}{378} \cdot 1 \cdot \frac{1}{162} = \frac{18}{178{,}564{,}176} =$	$\frac{1}{9{,}920{,}232}$
d_{36}	$\frac{1}{2} \cdot \frac{3}{27} \cdot 1 \cdot \frac{1}{378} \cdot 1 \cdot \frac{1}{8748} = \frac{3}{178{,}564{,}176} =$	$\frac{1}{59{,}521{,}392}$
d_{37}	$\frac{1}{2} \cdot \frac{3}{27} \cdot 1 \cdot \frac{1}{20{,}412} \cdot 1 \cdot \frac{1}{162} = \frac{3}{178{,}564{,}176} =$	$\frac{1}{59{,}521{,}392}$

Table 3.2: Probabilities of derivations $d_1 - d_{37}$.

situation as the parallel node in the other derivations. In our set of 37 derivations, we find several that are derivationally equivalent. There are ten cases in which in one derivation the pairing $c_2, \mathbf{map}_1$ is applied, and in the other $c_3, \mathbf{map}_1$. As the mappings of these pairings point to exactly the same vertices in situation 1, it is possible that two derivations, when otherwise using the same rules in the same order, are equivalent. Furthermore, there are two cases (a_7 and a_9) in which three derivations are equivalent, namely when c_1, c_2 and c_3 are applied at the same point in the derivation. In these cases, the first constituent of the three constructions is combined with c_4 and the second constituents with c_5.

What this table tells us is that parse a_6, consisting of derivations d_6 is the most likely analysis or a_{best}. In parse a_6, c_1 is combined with c_4 as its first constituent, and the second constituent is phonologically specified and can hence be terminated. This analysis is only minimally different from the second-best parse, the slightly more compositional a_7. In this parse, the second constituent is combined with another construction (c_5). The third-best analysis is a_5, consisting of just the derivation d_5. In this derivation, the two lexical constructions c_4 and c_5 are concatenated and no overarching construction is used. The three best analyses, in the bracket notation, are given below:

(26) a_6: [[ENTITY]→[BALL / *ball*] [MOVE / *go*]]

(27) a_7: [[ENTITY]→[BALL / *ball*] [MOVE]→[MOVE(MOVED,GOAL) / *go*]]

(28) a_5: ([BALL / *ball*] [MOVE(MOVED,GOAL) / *go*])

3.5.4 Implementation: lineair processing and pruning

The use of a probability model to find the best analysis is inspired by the statistical parsing tradition (Jurafsky & Martin 2009), where the disambiguation between multiple possible analyses is a massive practical problem. We may, however, doubt its cognitive reality. The most elementary of these concerns, that human beings do not actually perform such calculations, can be considered well-addressed by Jurafsky's (2003) discussion of the use of probability models in language comprehension and production. Jurafsky acknowledges that it is unlikely that people actually perform these calculations, but argues that probability models constitute a well-understood tool to model aspects of frequency and the competition between units (words, constructions).

Nonetheless, it remains unlikely that, even if the probability model is but an analytical tool, language users 'consider' all of the possible derivations that a model like SPL allows for. Starting from the insight that processing takes place linearly, and that language users do not keep track of all possible analyses (as evidenced by studies on garden-path sentences, see for instance Ferreira, Bailey & Ferraro (2002)), SPL performs the actual analysis in a bottom-up way, pruning away all but the most likely analyses (similar to the model developed by Jurafsky (1996)). As this aspect of the model was not at the heart

parse	derivations	probabilities derivations	probability parse
a_1	d_1	$\frac{1}{157,464}$	$\frac{1512}{238,085,568}$
a_2	d_2	$\frac{1}{26,244}$	$\frac{9072}{238,085,568}$
a_3	d_3	$\frac{1}{26,244}$	$\frac{9072}{238,085,568}$
a_4	d_4	$\frac{1}{52,488}$	$\frac{4536}{238,085,568}$
a_5	d_5	$\frac{1}{8748}$	$\frac{27,216}{238,085,568}$
a_6	d_6	$\frac{1}{1701}$	$\frac{139,967}{238,085,568}$
a_7	d_7, d_{10}, d_{20}	$\frac{1}{20,412} + \frac{1}{4536} + \frac{1}{4536}$	$\frac{116,640}{238,085,568}$
a_8	d_8	$\frac{1}{20,412}$	$\frac{11,664}{238,085,568}$
a_9	d_9, d_{13}, d_{23}	$\frac{1}{122,472} + \frac{162}{238,085,568} + \frac{1}{1,469,664}$	$\frac{2268}{238,085,568}$
a_{10}	d_{11}, d_{21}	$\frac{1}{734,832} + \frac{1}{734,832}$	$\frac{648}{238,085,568}$
a_{11}	d_{12}, d_{22}	$\frac{1}{734,832} + \frac{1}{734,832}$	$\frac{648}{238,085,568}$
a_{12}	d_{14}, d_{24}	$\frac{1}{1,469,664} + \frac{162}{238,085,568}$	$\frac{324}{238,085,568}$
a_{13}	d_{15}, d_{25}	$\frac{1}{238,085,568} + \frac{1}{238,085,568}$	$\frac{2}{238,085,568}$
a_{14}	d_{16}, d_{26}	$\frac{1}{238,085,568} + \frac{1}{238,085,568}$	$\frac{2}{238,085,568}$
a_{15}	d_{17}, d_{27}	$\frac{1}{238,085,568} + \frac{1}{238,085,568}$	$\frac{2}{238,085,568}$
a_{16}	d_{18}, d_{28}	$\frac{1}{238,085,568} + \frac{1}{238,085,568}$	$\frac{2}{238,085,568}$
a_{17}	d_{19}, d_{29}	$\frac{1}{238,085,568} + \frac{1}{238,085,568}$	$\frac{2}{238,085,568}$
a_{18}	d_{30}	$\frac{1}{183,708}$	$\frac{1296}{238,085,568}$
a_{19}	d_{31}	$\frac{1}{183,708}$	$\frac{1296}{238,085,568}$
a_{20}	d_{32}	$\frac{1}{9,920,232}$	$\frac{24}{238,085,568}$
a_{21}	d_{33}	$\frac{1}{9,920,232}$	$\frac{24}{238,085,568}$
a_{22}	d_{34}	$\frac{1}{9,920,232}$	$\frac{24}{238,085,568}$
a_{22}	d_{35}	$\frac{1}{9,920,232}$	$\frac{24}{238,085,568}$
a_{23}	d_{36}	$\frac{1}{59,521,392}$	$\frac{4}{238,085,568}$
a_{24}	d_{37}	$\frac{1}{59,521,392}$	$\frac{4}{238,085,568}$

Table 3.3: All parses A for *Ball go there*.

of my research, the implementation of the parser simply satisfies these constraints, but more realistic processing models can be thought of. In line with desideratum D5-2, I implemented the parser of SPL as follows.

SPL processes the words one by one. Over a span of words up to a certain word, a (possibly empty) set of derivations can be formed, which can be derivationally equivalent, and hence form a set of parses. From among this set of parses over a span, SPL only keeps the most likely one (or ones if there are multiple equiprobable parses) and discards the rest. When processing the next word, only the parses that are still active can be used to be combined into larger parses. Technically, the model employs an adaptation of the Cocke-Younger-Kasami algorithm that allows for words to be ignored, and prunes every cell in the matrix to the most likely analysis.

The motivation for this way of implementing the model not only comes from processing studies, but also from Langacker's (1988) discussion of processing, where he argues that when multiple units are in competition, a language user only selects a single one as the active unit. The implementation therefore not only constitutes an attempt to adhere to processing studies, but is also faithful to the description of processing within the theoretical framework.

3.5.5 SPL as a usage-based processing model

The derivation process described in this section allows the model to do comprehension on the basis of an utterance and a set of situations. As we will see later in this chapter, the production of an utterance on the basis of a situation is also among the model's possibilities, and therefore the model satisfies desideratum D2 (comprehensiveness). The model furthermore satisfies desiderata D5-1 (heterogeneous structure building) and D6-4 (developmental continuity) by having a set of diverse processing mechanisms that remain available over time. In the actual implementation of the way SPL performs its analyses, the model can be said to satisfy desideratum D5-2, although this aspect is not at the center stage of this research, and likely more realistic models of processing can be developed.

3.6 Learning

The resulting best parse a_{best} from every input item constitutes the input for the learning procedure. The constructions used in the best parse are reinforced (**reinforcement**), the result of any concatenative process is stored (**syntagmatization**) and any new possible abstractions that can be made are added to the constructicon (**paradigmatization**). Finally, the learner stores a limited number of recent best parses and the situations they were assumed to refer to, and employs a simple form of **cross-situational learning** to extract initial representations.

The learning model presented here aims to make three contributions to usage-based theory. First, SPL uses the syntagmatization operation to gradually build up longer constructions. The build-up of increasingly long constructions is a feature needed by the model to satisfy the law of cumulative complexity (D6-1). Second, the paradigmatization operation extracts any overlaps between the more concrete constructions and involves no grammar-wide evaluation of how useful the abstraction is. As I will argue, these features make the model conceptually congruent with the learning-by-processing and immanence view (cf. desiderata D4-3 and D6-2). Third, the model is the first usage-based model of language acquisition that is shown to acquire both lexical and grammatical constructions at the same time (cf. desideratum D2-8).

3.6.1 Reinforcement

The simplest form of learning is the reinforcement of the constructions employed in the derivations of the best parse.

Maximally concrete constructions

Langacker (2009) argues that the maximally concrete representations of the situation and the linguistic units used leave a trace in memory when they are processed. I operationalize this idea as follows. Recall that in an analysis a, there are several parallel derivations (i.e., every step in the derivation is the same, although different constructions may be used). For every such parallel step $\mathbf{s}$ in the derivation where a construction-mapping pairing $c, \mathbf{map}$ is applied, a maximally concrete construction $\mathbf{mcc}$ is extracted. We assume $\mathbf{mcc}$ to have as its meaning $sd_{\mathbf{mcc}}$ the subgraph of the situation to which the meaning of c maps via $\mathbf{map}$. The meaning of this novel construction thus directly reflects the conceptualization of the usage event in full detail. The signifiers of $\mathbf{mcc}$ consist of the signifiers of c, where the phonological constraints will be specified with whatever substring of the utterance is filling them, thus reflecting the utterance of the usage event in full detail.

In the case of a string that has been bootstrapped into a signifier of a construction c, we assume a novel construction $\mathbf{mcc}$ with as its signified meaning $sd_{\mathbf{mcc}}$ the vertex in the situation to which sr_c^i maps via the mapping $\mathbf{map}$ paired with c. The signifier of this bootstrapped construction $\mathbf{mcc}$ is then a pair of the string of words bootstrapped and a semantic constraint pointing to the vertex that constitutes $sd_{\mathbf{mcc}}$.

All maximally concrete constructions ($\mathbf{mcc}$s) are then added to the constructicon Γ (if they are not already present in it) with a count of 0. Example 3.16 illustrates the extraction of the maximally concrete constructions out of parse a_7. Importantly, the $\mathbf{mcc}$ for the second step of the derivation (the application of rule **ii**) is a phonologically specified grammatical construction. This construction, as opposed to the second and third $\mathbf{mcc}$s, is not present yet in the constructicon and added with a count of zero if a_7 were the best parse.

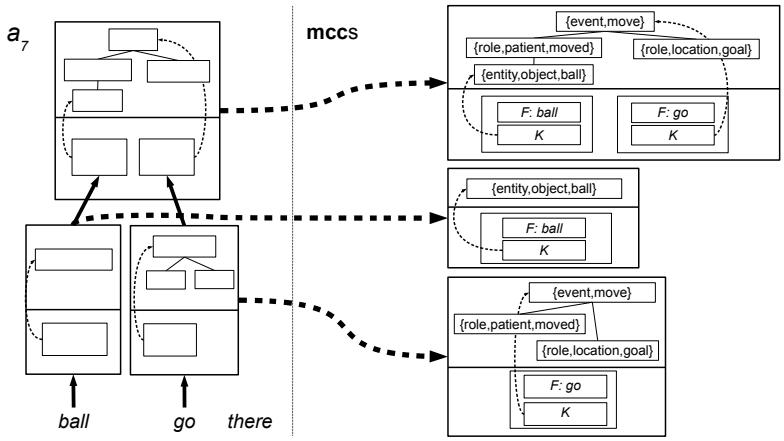


Figure 3.16: Extracted maximally concrete constructions from parse a_7.

The storage of maximally concrete representations is needed to account for prototype effects. A group of constructions in the constructicon whose meaning stands in superset-subset relations to each other may regularly map to the same subgraph of a situation in the analysis. However, if some more concrete constructions (i.e., constructions having more conceptual features specified in the constructional meaning sd) are used more frequently, we expect them to be more readily applicable than equally concrete constructions that are not as frequent. Now, if we only reinforce the more abstract constructions used, we cannot keep track of the frequency of the more concrete ones. Therefore, adding the maximally concrete ones, and generalizing over them in the abstraction step of the learning procedure (cf. section 3.6.3) allows us to keep track of this information.

This approach is similar to Alishahi & Stevenson's (2010) clustering approach, where frequently occurring conceptual features have more weight in the recognition of a construction. However, because in SPL the cluster can be said to be stored in a distributed fashion (a cluster in Alishahi & Stevenson's (2010) approach would correspond to a number of constructions in the constructicon in my approach), a 'cloud' of constructions may have multiple prototypes. That is to say: there may be two distinct sets of features being prototypical for a construction and both would be seperately stored in my approach, whereas in Alishahi & Stevenson's (2010) approach the association strength of the features is averaged over when they are clustered together.[4]

[4]However, if they are too distinct, they will form different cluster. The point is that with a hard

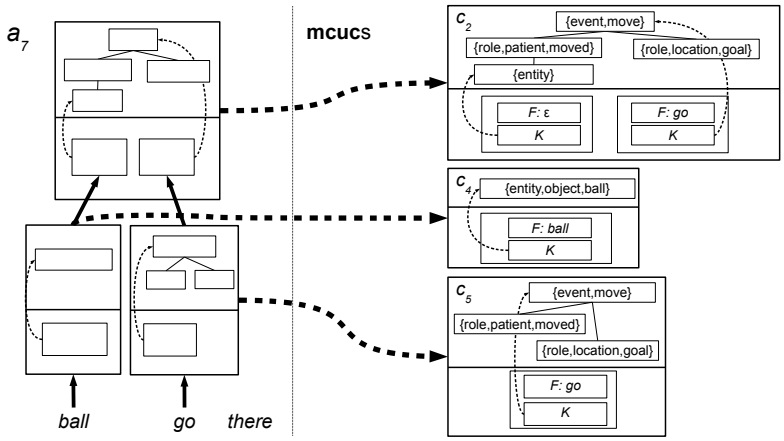


Figure 3.17: Extracted maximally concrete used constructions from parse a_7.

Reinforcement for maximally concrete used constructions

Secondly, not only the most concrete constructions are added to the constructicon, the constructions that are actually used are reinforced as well. The model does not reinforce all constructions used in all derivations of the best analyses, but only the most concrete *used* constructions per parallel step $\mathbf{s}$ in the derivation or: $\mathbf{mcuc}(\mathbf{s})$. Note that these are not necessarily the $\mathbf{mcc}(\mathbf{s})$ as defined in the previous paragraph. A construction c in any derivation in a_{best} is a maximally-concrete used construction if there is no other construction c' at the same step $\mathbf{s}$ in another derivation in a_{best} whose meaning is a superset of the meaning of c.

For simplicity's sake, I assume that after every input item, one 'count' can be distributed over the various most concrete constructions per step. That is: if there is a single derivation in the best parse, all constructions in that derivation are updated with 1. However, we will sometimes have multiple maximally-concrete used constructions for a certain step of the derivation. In that case, we distribute the count of 1 uniformly over the maximally-concrete used constructions at that step of the derivation. The update function thus is defined as follows:

$$count_{c^t} = \begin{array}{ll} count_{c^{t-1}} + \frac{1}{|\mathbf{mcuc}(\mathbf{s})|} & \text{if } c \in \mathbf{mcuc}(\mathbf{s}) \\ count_{c^{t-1}} & \text{otherwise} \end{array} \tag{3.11}$$

Figure 3.17 gives an example of the extraction and update of the maximally-concrete used constructions for parse a_7. The main difference with

the **mccs** in figure 3.16 is that the reinforced construction is not fully phonologically specified and has a more abstract signified conceptual representation.

The reason for using maximally-concrete used constructions, is that only those constructions are reinforced that are used productively. Their 'parents' in the constructional network that may be used in parallel steps in other derivations of a_{best} do not get reinforced, as there is a more concrete construction 'blocking' their update. This could be regarded as a form of pre-emption in the reinforcement procedure. Alternatively, we could say that the more abstract constructions are only motivating the use of the use of the maximally concrete constructions, backing them up with their probability mass.

A desirable effect of this procedure is that more abstract constructions are only reinforced when they are used productively, that is: in novel situations where no more concrete daughter constructions of those constructions can be used. This reflects Bybee's (2006) ideas about type and token frequency: the more *novel* instances of an abstract pattern are found, the more distinct types it can be said to have, and the more it will be reinforced.

3.6.2 Syntagmatization

Syntagmatization allows for the gradual build-up of the valency of constructions (i.e., the number of slots they have). Postponing the formal definition of the process for now, syntagmatization as a learning process is derived from the same general gradualist starting points many usage-based developmental theorists start off from (Tomasello 2003, Goldberg 2006). Despite being a gradualist take on the growth of grammar, this notion has not been worked out in detail by either of these theorists. If we want to adhere to Brown's law of cumulative complexity (desideratum D6-1), we have to assume that at least something akin to this learning process has to take place in the language-learning child

The fact that early productions often have fewer arguments expressed can, to my mind, be explained most readily if we assume that the constructions underlying these productions have more restricted valency patterns than later constructions. Most developmental approaches assume a combination of richer linguistic structure plus the deletion of some elements (Bloom et al. 1975). I believe this to be (1) a less parsimonious explanation, and (2) not in line with findings such as those presented by Theakston et al. (2012), who show that productions with transitive verbs and a single argument (SV and VO-utterances) have a different profile than productions at the same age with transitive verbs and two arguments (SVO-utterances). I interpret this fact as suggesting that the child uses different representations to generate SV, VO, and SVO-utterances respectively.

Similarly to the hypothesis that the various paradigms of a construction are gradually learned (which is typically called 'abstraction'), I assume that the

clustering operation, the model needs to decide and the cluster takes on a centroid representation.

syntagms constituting adult constructions are also acquired in an item-based, piecemeal way. Most developmental theorists, and many usage-based computational models discussed in the previous chapter assume that the learner is able to process the complete utterance and understand the valency relations between several elements of the utterance. Over these maximally concrete valency relations, then, more abstract constructions are learned. To my mind, this approach overlooks two other steps which we should expect to take place simultaneously, viz. the acquisition of lexical constructions and the acquisition of the linguistic realization of the semantic valency relations of these words. Syntagmatization takes care of this latter process. The gradual build-up of grammatical syntagms is reminiscent of Freudenthal et al.'s (2010) approach. Their MOSAIC model gradually builds up an inventory of strings of words to process utterances. The SPL model takes a similar approach, but combines it with a semantic parsing approach.

Implementation

Recall that a derivation can contain a number of concatenated constructions by the application of rule **i**. These constructions are understood by the model as being part of the same communicative intent. What syntagmatization does, then, is to take these concatenated constructions, look for constructions whose meanings stand in a semantic head-dependent relation to each other, and extend the 'head' constructions expressing that semantic head with the 'dependent' constructions.

Formally, the set of concatenated derivations consists of all applications of construction-mapping pairings that are directly governed by rule **i**. We use the maximally-concrete construction **mcc** for every construction-mapping pairing. For every construction c in this set with the meaning sd_c, we take all other constructions c' in this set whose meaning $sd_{c'}$ refers to a child or grandchild of the root vertex of sd_c. If the root vertex expresses an event, this involves the semantic roles it projects and the referents filling these roles. The reason we include grandchildren is that event roles are specified on a separate vertex in the meaning representation, and we want to capture events and their participants. This particular design choice thus depends on the semantic formalism used, and has to be modified to accomodate different representational formats.

Next, we take the constituents of c and the head constituents of any other construction c' that refers to semantic dependents of sd_c, and linearly consider those to be the signifiers of a novel construction c_{syn}. The meaning of c_{syn}, viz. $sd_{c_{\text{syn}}}$, consists of the meaning of c, the root vertices of the meanings of all dependent c' and any vertices from the situation needed to make $sd_{c_{\text{syn}}}$ connected. c_{syn} is then added to the constructicon with a count of 0.

To give an example, let us assume a_5 was the best parse. In this parse, three partial derivations are concatenated with rule **i**. The third, however, is the ignoring of *there*, and hence is not considered. The set of concatenated constructions thus consists of the **mcc**s for c_4 and c_5. For the former, there

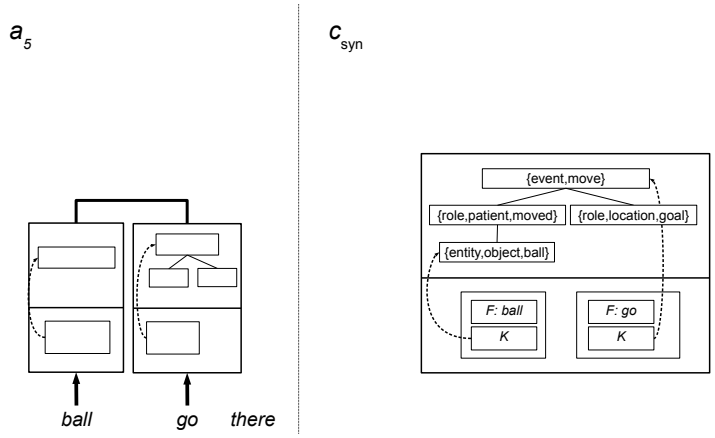


Figure 3.18: An example of the syntagmatization process applied to parse a_5.

are no other constructions in the set that express semantic dependents of it and hence no novel syntagmatizations can be made. c_5, on the other hand, expresses the {EVENT,MOVE} vertex of the situation, and c_4 expresses {ENTITY,OBJECT,BALL}, which is a grandchild of the {EVENT,MOVE} vertex. We therefore take all constituents of the **mcc** of c_5 and the head constituent of the **mcc** of c_4 and consider those to be a novel construction. The meaning of this novel construction consists of the meaning of the **mcc** of c_4 and the root vertex of the **mcc** of c_5 and any vertices needed to connect them (i.c., none). This novel construction is then added to the constructicon as c_6 with a count of 0. Figure 3.18 illustrates this process.

3.6.3 Paradigmatization

As we saw before, the model is able to parse utterances using a mixture of concrete and abstract constructions. How does it obtain these more abstract constructions? We can consider abstraction as the formation of paradigms of linguistic elements that can be substituted for each other, and hence call the process **paradigmatization**. In my implementation of the notion of abstraction, I again follow Langacker (2009), who argues that abstraction is not so much the creation of a novel hypothesis about the constructicon, but rather a by-product of processing several more concrete instantiations of a pattern. The overlap between these more concrete instantiations then becomes a potential to generalize. Whether one describes this in terms of abstract schemas or as a set of exemplars plus a rule to analogize over these, does not matter ac-

cording to Langacker (2009), as long as one is aware that these abstractions are not new cognitive 'entities' created from other 'entities' but rather a potential that is 'immanent' in these exemplars. In my implementation, however, the abstractions are separate entities. This should be seen as reflecting an implementational rather than an ontological issue, and as such it does not conflict with desideratum D4-3.

It is the idea of acquiring a grammar as a hypothesis testing procedure that underlies Bayesian Model Merging. What I propose, for abstraction, is to take the view seriously that there is no such thing as selection between levels of abstraction, i.e., that the organization of the abstraction in the constructicon is not governed by a selection mechanism deciding which level or clustering is the most appropriate one given the data and some prior conception on what the constructicon, or grammars in general, should look like (e.g., compact, or uniform). Rather, all possible abstractions over reinforced constructions (i.e., constructions with non-zero counts) are made, and the reinforcement of some of these abstractions, but not others (as discussed in section 3.6.1) leads to a constructicon that is highly general, but probabilistically constrained (i.e., utterances analyzed with both abstract and concrete constructions will have higher probabilities than utterances analyzed with only abstract constructions). This way, the abstraction in the SPL model differs from that of Chang (2008) and Beekhuizen, Zuidema & Bod (2013), who apply a Minimum Description Length criterion (Rissanen 1978) to the selection of abstractions, as well as Alishahi & Stevenson (2010), who cluster maximally concrete frames, thereby forcing the model to categorize an input item discretely with one or the other centroid cluster. Incorporating an element of 'selection' in one's model fits better with a deductionist view on language acquisition than an inductionist. For that reason, I think having a model that does allow for a gradual, bottom-up search through the hypothesis space, but without selection, is the preferable computational approach for a usage-based account of grammar acquisition (cf. desideratum D6-2).

Implementation

Whenever a construction c obtains its first reinforcement, it is compared to all constructions $c' \in \Gamma$ at that point in time that have also been reinforced (i.e., have a $count_{c'} > 0$). If from the overlap between c' and c a new construction c_{para} can be formed, and if c_{para} is not in Γ yet, c_{para} is added to the grammar.

The formation of an abstraction requires a comparison between c and c'. Not all comparisons lead to novel abstractions. Crucially, the model has to be able to find parallels between c and c' in both their signifiers and signifieds. This does not constitute a selection process, but rather reflects what comparisons the learner can and cannot make. A novel construction c_{para} can be formed from c and $c' \in \Gamma$ under the following conditions:

Conditions for creating an abstraction over two constructions

---

- Let $\mathbf{map}_{\text{overlap}}$ be a bijective structure-preserving mapping $f : sd_c \rightarrow sd_{c'}$ between the signified meanings sd_c and $sd_{c'}$ of two constructions c and c' such that
  - $\forall v_{v \in sd_c}.(\mathbf{map}_{\text{overlap}}(v) \cup v) \neq \emptyset$
- Let $\mathbf{M}_{\text{overlap}}(c, c')$ be the set of all possible $\mathbf{map}_{\text{overlap}}$ between c and c'.
- For each $\mathbf{map}_{\text{overlap}} \in \mathbf{M}_{\text{overlap}}(c, c')$, a novel construction c_{para} is created iff
  - $|sr_c| = |sr_{c'}|$
  - $\forall i_{i \in [1,\ldots,|sr_c|]}.\mathbf{map}_{\text{overlap}}(K(sr_c^i)) = K(sr_{c'}^i)$
  - $\neg((|sr_c| = 1) \wedge (F(sr_c^1) \neq F(sr_{c'}^1)))$
- where c_{para} consists of
  - sd_c o contains the intersection between all elements in $\mathbf{map}_{\text{overlap}}$ as well as their edge structure.
  - $sr_{c_{\text{para}}}$, where for each $i \in [1, \ldots, |sr_c|]$, $sr_{c_{\text{new}}}^i$ consists of
    - $F(sr_{c_{\text{new}}}^i = F(sr_c^i)$ if $F(sr_c^i) = F(sr_{c'}^i)$ else ϵ
    - $K(sr^i(c_{\text{new}})) = \mathbf{map}_{\text{overlap}}(K(sr_c^i)) \cup K(sr_c^i)$

The starting point for the abstraction over two constructions c and c' is an intersection mapping $\mathbf{map}_{\text{overlap}}$ between their meanings sd_c and $sd_{c'}$. For every possible intersection mapping between two constructions, we create a novel construction if all signifiers in both c and c' have conceptual constraints mapped to each other per $\mathbf{map}_{\text{overlap}}$.

Two further constraints are that the number of signifying constituents must be equal for both constructions, and that if one of the constructions is a lexical construction, the two constructions must have the same phonological constraint on that signifier. This last constraint is intended to obviate the possibility of having phonologically-unspecified single-constituent constructions. These constructions add little to the potential for analyzing an utterance, and even though they could be extracted, using them would always result in derivations of a lower probability than derivations mapping to the same part of the same situation without them. Chang (2008), however, does allow for them.

The constraints proposed above are motivated, but not cast in stone. One

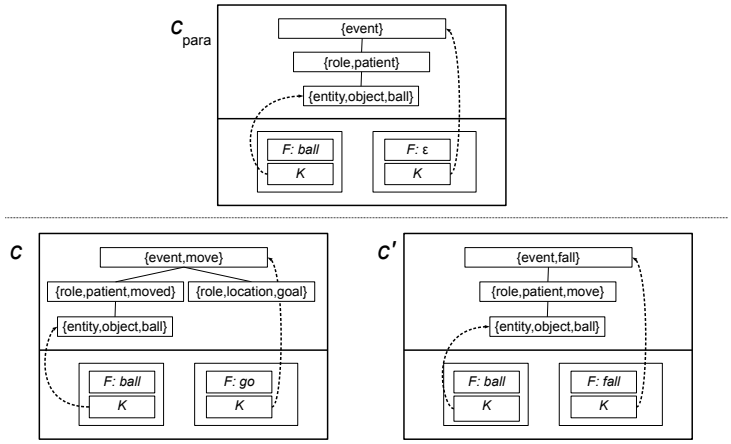


Figure 3.19: An example of succesful paradigmatization.

may drop the latter two constraints. The final constraint (identical phonological signifiers in the case of lexical constructions) only keeps the model from making spurious abstractions that, due to the probability model, will not be used.[5] The equal-number-of-signifiers constraint is more interesting. Loosening or dropping this constraint may result in different kinds of constructions being abstracted (in a similar way to Chang (2008)), but for the current purposes, this would needlessly complicate the model.

To give an example of the paradigmatization procedure, assume that a_5 was the best parse, and that the syntagmatized construction in figure 3.18 was added to the grammar. Through some subsequent input item, that construction becomes reinforced. Assume furthermore that another construction, represented as c' in figure 3.19, was present in the grammar as well. The meanings of the two constructions can be mapped with an overlap mapping, and the signifying constituents of both constructions point to each other via this mapping, so an abstraction can be made. This abstraction, c_{para} in figure 3.19, contains the intersection between c and c' as its meaning. The first constituent is phonologically specified and points to the {OBJECT,ENTITY,BALL} vertex in the meaning of c_{para}. The second, on the other hand, is not phonologically specified (as the phonological constraints on the second constituents of c and c' differ) and points to the root vertex of the meaning of c_{para}. With the paradigmatization operation, the model has now extracted a semi-open [[BALL / *ball*]

[5] An 'abstract' lexical construction is simply useless in creating derivations as for every use of an abstract lexical construction plus a concrete one, the competing analysis involving only the concrete lexical construction is more probable.

[EVENT]] construction, which is then added to the grammar with a count of 0.

Paradigmatization, I claim, is congenial with the view that abstractions are immanent in the more concrete patterns that instantiate them. Recall that Langacker argues that abstraction is essentially the co-activation pattern of several more concrete patterns. I believe the overlaps correspond to these co-activation patterns. Because SPL 'extracts' any and all of these patterns, the set of paradigmatized constructions can be seen as the potential for abstraction immanent in the more concrete ones. If a selection between paradigmatizations were made, on the basis of some criterion, the immanence would be, at least, harder to defend, as it would require a bridging hypothesis between the selection of certain paradigmatizations but not others. The view that abstractions are immanent, but discretely represented in a model is not new: we can find a similar idea in Skousen's (1989) Analogical Modeling, where all abstractions over a feature set are abstracted, and the model performs analogical reasoning over these.[6]

Note that paradigmatized constructions can be reinforced without the more concrete constructions instantiating them receiving further reinforcement. If an abstract construction is frequently used as the maximally-concrete used construction in many cases, it will receive much reinforcement, and hence be established as a unit, without the more concrete constructions achieving unit status. If an abstraction, however, is hardly used, for instance, because it generalizes over only two more concrete patterns that are themselves often used as **mcuc**s, the abstraction will stay rather weakly reinforce. As we will see in the following chapters, this dynamic leads to interesting insights in the development of constructional networks.

3.6.4 Cross-situational learning

When we assume a usage-based perspective on language acquisition, the model starts with an empty inventory of signs. Therefore, it should have learning operations at its disposal to get an initial inventory of constructions off the ground. A prime candidate for such learning operations is cross-situational learning.

Broadly speaking, cross-situational learning is the process whereby a learner observes multiple situations in which utterances are produced and extracts or reinforces recurring matching pairs of parts of the utterances and parts of the situations. An intuitive example would be the case in which the learner first hears the utterance *you grab the ball!* and sees a ball on the table and understands the intention that the caregiver want her to grab it, and next the utterance *oh, now the ball is on the floor!*, paired with a situation where the child just threw the ball off of the table. The phonological substring *the ball* is

[6] The difference being that the abstractions themselves can receive reinforcement and thus obtain a degree of representational autonomy, whereas this in not possible in lazy learners such as Analogical Modeling or Memory-Based Learning.

shared between the two utterance-situation pairs, as is the semantic element of an entity 'ball' being present in both understood communicative intentions.

The acquisition of form-meaning pairings via cross-situational learning can be interpreted in several ways. Many word learning models (Xu & Tenenbaum 2000, Frank, Goodman & Tenenbaum 2009, Fazly et al. 2010) assume a probabilistic model, where the connections between phonological strings and their referents get reinforced every time a pair of a word and a semantic element co-occur in an utterance-situation pairing. Recently, this view has been contested by researchers who argue that this places too much of a burden on the learner, as she has to maintain and update an $m \times n$ matrix for all m seen words and all n seen semantic elements (Stevens 2011). Instead, Stevens proposes, the learner forms hypotheses at random, and validates these in next rounds, either reinforcing them if the new utterance-situation pairing corroborates it, or discarding them if not.

There are things to be said for both views. The probabilistic view, as opponents of this view argue, creates behavior that is too gradient in nature. Learner's behavior seems more categorical than would be expected on the basis of a probabilistic view. On the other hand, one could argue that a probabilistic system interacts with more discrete decision-making systems which are addressed (possibly in different ways) in experiments and natural processing and production behavior. This could resolve the issue of apparent discreteness in behavior, but until it is worked out, it remains hand-waving. On the other hand, the creation of hypotheses at random seems like a strange starting point. It is not clear to me why a learner would form a hypothesis about a form-meaning pairing on the basis of no evidence.

The instantiation of cross-situational learning I assume here can be seen as taking a halfway position between the two. Because I do not think the metaphor 'language acquisition as hypothesis testing' is the right one (see section 3.6.3 as well), I consider these initial form-meaning pairings to be reflections of the processing of utterance-situation pairs, despite the learner not having any contentive linguistic knowledge yet. The cross-situationally extracted patterns are not random guesses, but reflect a simple form of analogical reasoning. On the other hand, I do not want to assume too much keeping track of every contingency between possible forms and possible meanings.

An exemplar is a structured representation of an experience. Importantly, it is *structured* by linguistic processing. That is: whatever linguistic structure is found in the input item (the U, S pairing) is stored alongside the U, S pair. For the current purposes, I assume that a linguistic exemplar is a pair of the selected situation $s \in S$, and the best parse a_{best}. Let us furthermore assume that the learner keeps track of the most recent n exemplars, where $n = [1, \infty]$.

Implementation

The form of cross-situational learning I assume extracts overlaps in form and meaning between a new exemplar and the most recent n exemplars. It only

extracts those overlaps about which it is sure, that is: only if a single maximal overlap in the ignored parts of the utterance strings and a single maximal shared subgraph between the analyses can be found, a novel construction containing exactly this overlap is extracted and added to the grammar with a count of 0. More formally:

Cross-situational learning

For every new exemplar s^t, a^t_{best} and every exemplar $s^{t-i}, a^{t-i}_{\text{best}}$ in the range $i = [1, \ldots, n]$:

- Let U_I be the yield of a parse a that is governed by rule **iii** (i.e., ignored words).
- Let G_I be the subgraph of s to which no root node of any construction used in a has a mapping, or:
  $G_I = \forall v_{v \in V_s}.\ \nexists c, \mathbf{map}_{c,\mathbf{map} \in a}.\mathbf{map}(sd_c) = v' \rightarrow v' \neq v$
- Extract a novel construction c_{xsl} iff:
  - $U^t_I \cup U^{t-i}_I \neq \emptyset$
  - $U^t_I \cup U^{t-i}_I$ is a contiguous substring of the yield of a^t_{best} (i.e., the original utterance U^t).
  - $U^t_I \cup U^{t-i}_I$ is a contiguous substring of the yield of a^{t-i}_{best} (i.e., the original utterance U^{t-i}).
  - Assuming the set $\mathbf{M}(G^t_I)$,
    - which consists of all possible structure-preserving bijective functions $\mathbf{map}_{\text{identical}}$ between a connected subgraph of G^t_I and a connected subgraph of G^{t-i}_I containing identical feature sets on the mapped vertices,

    there is exactly one most-encompassing mapping $\mathbf{map}_{\text{mem}} \in M(G^t_I)$ such that all other functions $\mathbf{map}'_{\text{identical}} \in M(G^t_I)$ specify a domain and a codomain that are subsets of the domain and codomain of $\mathbf{map}_{\text{mem}}$.
- where the new construction c_{xsl} consists of
  - $sd_{c_{\text{xsl}}}$, which is the subgraph of G^t_I being the domain of $\mathbf{map}_{\text{mem}}$
  - $sr_{c_{\text{xsl}}}$, being a single constituent sr^1,
    - where $K(sr^1_c) = v_{\text{root}}(sd_{c_{\text{new}}})$, and
    - $F(sr^1_c) = U^t_I \cup U^{t-i}_I$

Whenever a new exemplar s^t, a^t is added, it is compared to all exemplars in this 'memory buffer'. In this comparison, for s^t, a^t and some s^{t-i}, a^{t-i}, the part of the utterance that is ignored in a^t is compared to the part of the utterance that is ignored in a^{t-i}. If the maximal overlapping substring is not a contiguous substring of the full yield of a^t and of a^{t-i}, or if the maximal overlapping substring is empty, no new construction is extracted.

On the side of the meaning, a similar process takes place. The model compares the part of the situations s^t and of s^{t-i} that are not analyzed with a^t and a^{t-i} respectively (i.e., G_I^t and G_I^{t-i}). If, out of all functions mapping a subgraph of G_I^t to an identical subgraph of G_I^{t-i}, there is exactly one function $\mathbf{map}_{\text{mem}}$ such that all other mapping functions specify subgraphs of the two graphs in $\mathbf{map}_{\text{mem}}$, a construction can be extracted. Otherwise, no construction can be extracted.

This precludes the situation in which more than one most-emcompassing mapping can be made, i.e., the situation in which one mapping points to one part of the situation graph, and another mapping to a (at most partially) overlapping other part of the situation. In those cases the learner cannot be fully sure which analogy to make, and hence extracts no novel construction. Admittedly, this strict rule for extracting initial constructions is relatively brittle and simplistic, but given the complexity of the rest of the model I believe an overly constrained learning mechanism is preferable over an underconstrained one. Furthermore, the constraints all derive from more general principles of making analogies and reasoning with uncertainty (beit in a very simple form: if the learner encounters any uncertainty, it will do nothing).

Importantly, substrings of more than a single word in the adult representation can be extracted with the cross-situational learning procedure. These holistic chunks correspond to undersegmentation in Peters's (1983) sense.

To given an example of the cross-situational learning procedure, assume that a_6 was the best parse of the utterance, and that there is a previous exemplar consisting of a^{t-1}, s^{t-1}, depicted in figure 3.20 below. As the maximal overlap in unanalyzed parts of the situations consists of the {ROLE,LOCATION, GOAL} vertex, and the overlap in the utterance of *there*, a novel construction, linking these two can be extracted.

3.6.5 SPL as a usage-based learner

The mechanisms described in this section jointly embody many of the insights discussed in the previous chapter. Despite being operationalized as learning operations applying to the best analyses, they can be conceptualized as neural effects of the processing of the best analysis itself. That is to say: there is no decision-making process outside the processing of the usage events (D6-2). Novel constructions are learned without evaluating their value: if they are of use to the model in analyzing utterances, they will receive subsequent reinforcement, if not, they will remain one-off patterns with a zero count. The model embodies developmental continuity (D6-4), with all learning mecha-

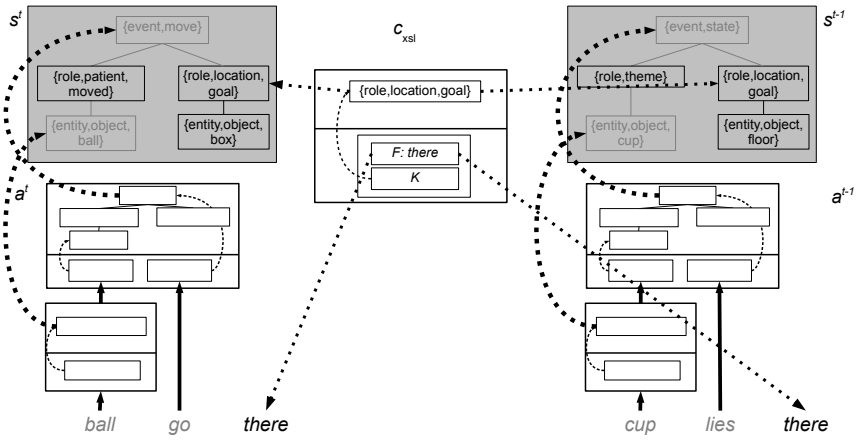


Figure 3.20: An example of succesful cross-situational learning. Analyzed parts of the utterances and the situations are marked in grey.

nisms being available throughout developmental time and simultaneity (D3) because both lexical and grammatical constructions can be learned with the same set of mechanisms. Although in practice, **cross-situational learning** operations will precede **syntagmatization** operations, which in turn precede the **paradigmatization** operations on them, all operations are used all the time.

Interestingly, the model has several ways of acquiring novel lexical constructions, viz. cross-situational learning, bootstrapping, and the use of maximally-concrete constructions. As bootstrapping will only take place after slightly more abstract constructions have been acquired, it typically takes place later in development than the cross-situational learning. These two ways of acquiring lexical constructions can be seen as reflecting Gleitman et al.'s (2005) ideas of the various ways in which lexical mappings can be learned, but from a usage-based perspective.

The two core mechanisms for the acquisition of grammatical constructions, syntagmatization and paradigmatization, further instantiate a few other desiderata. With syntagmatization, the acquisition of longer grammatical rules is qualitatively grounded in the usage-events: there is no preconception that longer rules will be part of the language, but the joint processing of several shorter rules leaves a trace in the mind of the learner (D5-1 and D6-2). At the same time, this means that longer constructions can only emerge if their parts are known, which satisfies D6-1, but having this as the only mechanism of the acquisition of rules conflicts with D6-3, the idea that we want a learner that does parts-to-whole and whole-to-parts learning. The latter is not

instantiated in this model, and, as I expressed before, I am skeptical about the necessity of such an operation and whether it fits in with ideas about learning-as-processing (D6-2).

The notion of abstraction engendered in the model is similar to Chang's, but differs in that it contains no post-hoc decision mechanism, thus being closer to the idea of learning-as-processing, to my mind. By extracting any and all abstractions, we can easily dereify the (reified) discrete representations as the potential for abstraction immanent in the most concrete constructions. Furthermore, the reinforcement mechanism only 'rewards' the most concrete *used* constructions, thereby boosting the potential of abstract representations (giving them more of a 'unit' status in Langacker's (2000) terminology) only if they are productively used.

3.7 Generation

An important property of SPL, as a generative model, is that it is bidirectional: we can analyze given utterances with it as well as generate new ones given a situation. Doing so, the model can simulate both processes of language comprehension and production (desideratum D2). Generation works largely by the same processes as analyzing an utterance, in that we generate a derivation that corresponds to the situation and take the phonological symbols at the leaf nodes of the derivation to be the utterance the model produces. Again, many analyses are possible and the model has to select the best one.

3.7.1 Differences with the analysis procedure

One aspect of the model differs from the comprehension procedure in the generation procedure. Several processing mechanisms defined in section 3.4 are geared towards processing input of which a part is not understood. In particular, the concatenation, ignoring and bootstrapping operations, as defined by rules **i**, **iii**, and **vi**, are operations allowing the model to interpret utterances despite having a limited inventory of linguistic signs. Generation works on the basis of known signs, and hence these three rules are not used. We assume that any derivation starts at rule **ii**, that is: with the application of a construction-mapping pairing. That is, the set of rules applicable in generation consists of rules **ii**, **iv**, **v**, and **vi**.

The probability of the derivation is again the product of the rules that are applied in it, as in equation (3.5). The probabilities of the $c, \mathbf{map}$ pairings are the same as for the comprehension procedure, and are repeated here as equations (3.12) and (3.13).

$$P(c, \mathbf{map}|CS) = \frac{count_c + 1}{\sum\limits_{c', \mathbf{map'} \in CS} (count_{c'} + 1) + 1} \tag{3.12}$$

	rule	probability
ii	START $\rightarrow (c, \mathbf{map})$	$P(c, \mathbf{map}\|CS_{\text{START}})$
iv	$(c, \mathbf{map}) \rightarrow sr_c^1, \ldots, sr_c^n$	1
v	$sr_c^i \rightarrow \alpha^+$ (if $F(sr_c^i) \neq \epsilon$)	1
vii	$sr_c^i \rightarrow (c', \mathbf{map}')$	$P(c', \mathbf{map}'\|CS_{sr_c^i})$

Table 3.4: Probabilities of the processing mechanisms in the generation procedure.

$$P(\mathbf{u}|CS) = \frac{1}{\sum\limits_{c,\mathbf{map}\in CS}(count_c + 1) + 1} \tag{3.13}$$

The probability of a derivation is, as in comprehension, given by:

$$P(d|\Gamma, S) = \prod_{r\in d} P(r) \tag{3.14}$$

3.7.2 Expressivity

In producing an utterance, a language user wants to be as expressive as possible (with as little effort as possible). I operationalize this idea as follows. Assume that the model creates analyses consisting of equivalent derivations, as in the comprehension procedure. The model penalizes an analysis for every feature of the situation that it does not express. It does so by taking the summed proportion of features in the situation s not expressed by the analysis a ($\mathbf{unexpressed}(\mathbf{a}, \mathbf{s})$) and raises the probability of an unseen event to the power of $\mathbf{unexpressed}(\mathbf{a}, \mathbf{s})$.

The calculation of $\mathbf{unexpressed}(\mathbf{a}, \mathbf{s})$ is defined as follows. For every vertex, the model checks which features are expressed by any construction in the derivation and takes the proportion of unexpressed features per vertex, after which the proportions are summed. Vertices that are completely unexpressed will thus contribute a proportion of 1 to $\mathbf{unexpressed}(\mathbf{a}, \mathbf{s})$, whereas partially expressed vertices add a value between 0 and 1 (exclusive) to the score. The referential penalty over a, or referential penalty(a) can thus be defined as:

$$\text{referential penalty}(a) = P(\mathbf{u}|CS_{\text{start}})^{\mathbf{unexpressed}(\mathbf{a},\mathbf{s})} \tag{3.15}$$

Obviously, the notion of referential expressivity employed here is a rather naïve one. The speaker, while interacting with the hearer, has a shared common ground with the hearer in the speech situation often allowing for the correct identification of the situation referred to with minimal means (see, e.g., Clark 1996). Furthermore, framing communicative success as the correct identification of a situation is rather simplistic as well. I leave it to future research to operationalize and implement more complex notions, involving for instance construal (Croft & Cruse 2004, ch. 3), argumentativity (Verhagen 2005), and the many other dimensions of communication.

3.7.3 Selecting the best analysis and utterance

The full probability of an analysis can now be given as follows:

$$P(a|s,\Gamma) = \sum_{d \in a} P(d|s,\Gamma) \cdot \text{referential penalty}(a) \tag{3.16}$$

Again, we can select the best analysis a_{best} given the situation, using the same definitions as in section 3.5.

Equally interesting is the selection of the best utterance. Multiple analyses may have the same yield (i.e., the same string of phonological structures) as the leaf nodes of the derivations. I take the probability of an utterance U given a situation s and a grammar Γ to be the disjoint probability of all analyses that have U as their yield.

$$P(U|s,\Gamma) = \sum_{a:\mathbf{yield}(a)=U} P(a|s,\Gamma) \tag{3.17}$$

3.7.4 An example of the generation procedure

Assume the grammar with the five constructions in figure 3.5, and the first situation (s_1) from figure 3.6. All of these constructions have subset mappings to the situation and hence all can be applied as the first rule. Constructions $c_1 - c_3$ are grammatical constructions and can therefore be combined with other constructions. Figure 3.21 gives all possible derivations given the situation and the grammar in the box-diagrammatic notation.

Calculating the derivational probabilities and the referential penalties, we get the probabilities of the six derivations in table 3.5. As we can glean from the table, derivations d_2, d_3 and d_4 are derivationally equivalent, while the other derivations are not equivalent with any other derivation. We thus arrive at the four analyses in table 3.6.

The penalty is calculated by looking at the summed proportion of the features on the vertices of the situation. For analysis a_1, the entire vertex {ENTITY, OBJECT,BOX} is left out. This means that the penalty is the probability of an

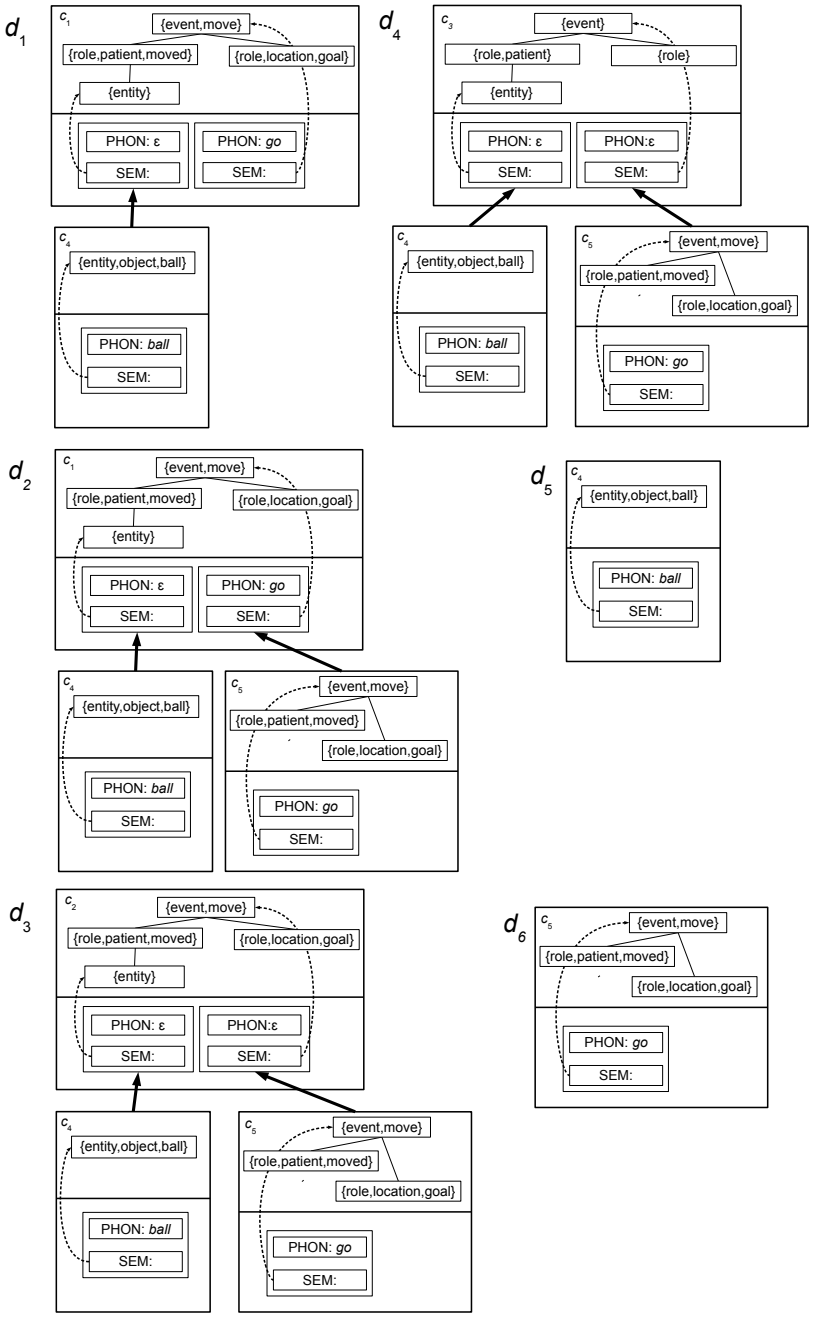


Figure 3.21: All derivations for the generation of utterances given s_1.

d	rules	yield	$\prod_{r \in d} P(r)$	$P(d\|s,\Gamma)$
d_1	**ii**, **iv**, **vi**, **iv**, **v**, **v**	*ball go*	$\frac{1}{13} \cdot 1 \cdot \frac{5}{6} \cdot 1 \cdot 1 \cdot 1$	$\frac{5}{78}$
d_2	**ii**, **iv**, **vi**, **iv**, **v**, **vi**, **iv**, **v**	*ball go*	$\frac{1}{13} \cdot 1 \cdot \frac{5}{6} \cdot 1 \cdot 1 \cdot \frac{2}{8}$	$\frac{10}{624}$
d_3	**ii**, **iv**, **vi**, **iv**, **v**, **vi**, **iv**, **v**	*ball go*	$\frac{2}{13} \cdot 1 \cdot \frac{5}{6} \cdot 1 \cdot 1 \cdot \frac{2}{8}$	$\frac{10}{624}$
d_4	**ii**, **iv**, **vi**, **iv**, **v**, **vi**, **iv**, **v**	*ball go*	$\frac{2}{13} \cdot 1 \cdot \frac{5}{6} \cdot 1 \cdot 1 \cdot \frac{2}{8}$	$\frac{10}{624}$
d_5	**ii**, **iv**, **v**	*ball*	$\frac{5}{13} \cdot 1 \cdot 1$	$\frac{5}{13}$
d_6	**ii**, **iv**, **v**	*go*	$\frac{2}{13} \cdot 1 \cdot 1$	$\frac{5}{13}$

Table 3.5: The probabilities of the six derivations in figure 3.21.

analysis	derivations	P derivations	$\sum_{d \in a}$	penalty	$P(a)$
a_1	d_1	$\frac{5}{78}$	$\frac{5}{78}$	$\frac{1}{13}^1$	$\frac{5}{1014} \approx 4.93e-3$
a_2	d_2, d_3, d_4	$\frac{10}{624} + \frac{10}{624} + \frac{10}{624}$	$\frac{30}{624}$	$\frac{1}{13}^1$	$\frac{30}{8112} \approx 4.93e-3$
a_3	d_5	$\frac{5}{13}$	$\frac{5}{13}$	$\frac{1}{13}^4$	$\frac{5}{371{,}293} \approx 1.35e-5$
a_4	d_6	$\frac{5}{13}$	$\frac{5}{13}$	$\frac{1}{13}^2$	$\frac{5}{2197} \approx 2.28e-3$

Table 3.6: The probabilities of the parses given the six derivation in figure 3.21.

unseen event to the power 1, or $\frac{1}{13}^{(1)}$. Analysis a_3, on the other hand, leaves out four full vertices, and thus has a penalty of $\frac{1}{13}^4$.

Two analyses are equally likely, viz. a_1 and a_2. If we look at the yields, we find that the utterance *ball go* is the most likely utterance given the situation and the grammar, with a probability of $P(U = \textit{ball go}|s,\Gamma) = 9.86e-3$. At about a quarter of that probability is the utterance *go*, supported only by analysis a_4 ($P(U = \textit{go}|s,\Gamma) = 2.28e-3$). The utterance *ball* is least likely to be generated by the model, with a probability of $1.35e-5$.

3.8 Meeting desiderata with SPL

SPL was developed with the theoretical discussion about the mechanisms necessary to account for language acquisition in mind. The close adherence to linguistic theorizing is therefore an aspect of this research that warrants its own section. In this section I evaluate whether SPL meets the various desiderata we set out in chapter 2. Throughout this chapter, I have discussed why the several aspects of SPL do so. Here, I briefly summarize them.

desideratum	evaluation
D1 (explicitness)	+
D2 (comprehensiveness)	+
D3 (simultaneity)	+
D4 (representational realism)	
D4-1 (qualitative grounding)	+
D4-2 (quantitative grounding)	+
D4-3 (immanence)	+
D5 (processing realism)	
D5-1 (heterogeneous structure building)	+
D5-2 (linear processing)	+
D6 (ontogenetic realism)	
D6-1 (cumulative complexity)	+
D6-2 (learning-by-processing)	+
D6-3 (parts-to-whole and whole-to-parts)	+/−
D6-3 (developmental continuity)	+
D7 (explanatory insight)	
D7-1 (unification)	+

Table 3.7: Evaluating SPL against the desiderata.

Concerning the explicitness of the model's simplifying assumptions (D1), I believe SPL to be relatively clear. Several aspects of the model were more at center stage than others, and, for instance, the implementation of cross-situational learning and the linear parser constitute highly simplified versions of obviously much richer cognitive processes.

Second, the model is (in principle) able to comprehend and produce utterances using the process of forming derivations and selecting the best one from among those (D2). It can, however, only be gradually expected to acquire this skill, as the model starts with an empty set of constructions. Over time, the model learns lexical and grammatical constructions, where the processes for the acquisition of both apply at the same time and are available to the model throughout developmental time (D3, D6-3).

The representations used by SPL, constructions, consist solely of conceptual and phonological structure, as well as a symbolic link, and can thus be said to be qualitatively grounded in the linguistic usage events (D4-1). By reinforcing the used constructions (more specifically, the maximally-concrete used construction), the model is sensitive to the frequencies of aspects of usage events (D4-2).

SPL processes utterances by using a derivation process, and selecting the most likely set of equivalent derivations from among all possibilities. Because such a global optimization process can be considered cognitively unrealistic, the actual analysis is done with a parser that goes over the utterance linearly and prunes all but the most likely analyses as it goes (D5-2). In building up analyses, the model has several processing mechanisms at its disposal: simple slot-filling, but also the creation of non-hierarchical analyses by means of concatenation, the top-down interpretation of a word by means of bootstrapping, and the possibility to ignore words, and as such the model has a robust toolkit of processing mechanism (D5-1).

Learning in SPL can be considered to be a by-product of processing (D6-2): the model processes an utterance, and the resulting best analysis as it is mapped to a situation leaves a trace by adding syntagmatized constructions and maximally-concrete constructions. Syntagmatization can be said to instantiate a cognitive take on Brown's law of cumulative complexity (D6-1) and part-to-whole learning: more complex constructions can only be learned on the basis of concatenations of simpler structures that leave traces in the mind of the speaker in the form of syntagmatized constructions.

Multiple constructions may share structure, in which case the model extracts a more abstract construction by means of paradigmatization. The paradigmatization operation involves no selection process whereby it is decided whether an abstraction is useful or not. As such any and all abstraction are extracted, and, as I argued, this makes this implementation of a notion of abstraction congruent with the idea of abstractions being immanent (D4-3).

Whole-to-parts learning was not the focus of this model, and, for instance, the model does not break down acquired unanalyzed chunks any further, and hence I evaluated D6-3 as $+/-$. There are some aspects of whole-to-parts

learning available to the model. One is the bootstrapping operator, whereby the meaning of an unknown part is assigned to it on the basis of a whole. This, however, does not constitute a case of the decomposition of a thitherto unanalyzed whole, and as such not all whole-to-parts learning operations conceivable are done by the model. Discussing this, I argued that the decomposition of chunks is perhaps an unlikely kind of cognitive operation from the perspective of D6-2: it requires the learner to engage in a post-hoc adjustment of the acquired chunks, which seems to be a kind of off-line reasoning for which more evidence would be needed. I leave it to proponents of the starting-big perspective to reconcile the decomposition of chunks with the learning-as-processing perspective, or to find evidence for off-line operations that break down chunks.

The issue of explanatory insight (D7) can of course only be discussed sensibly after we have seen some results. Before we turn to these, there is one aspect of the realism of the model that I would like to consider, viz. the nature of the input items, especially the situational contexts, given to the model. The next chapter deals with this issue.

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).

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

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

$noise$ 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-

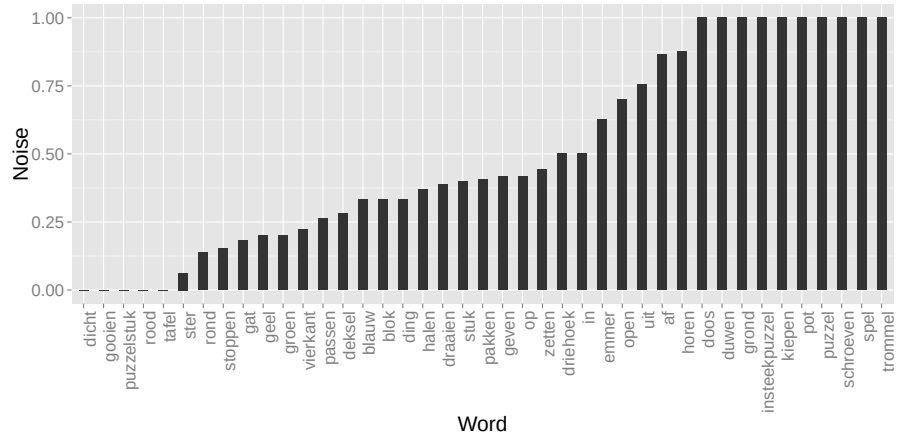


Figure 4.1: Noise per word.

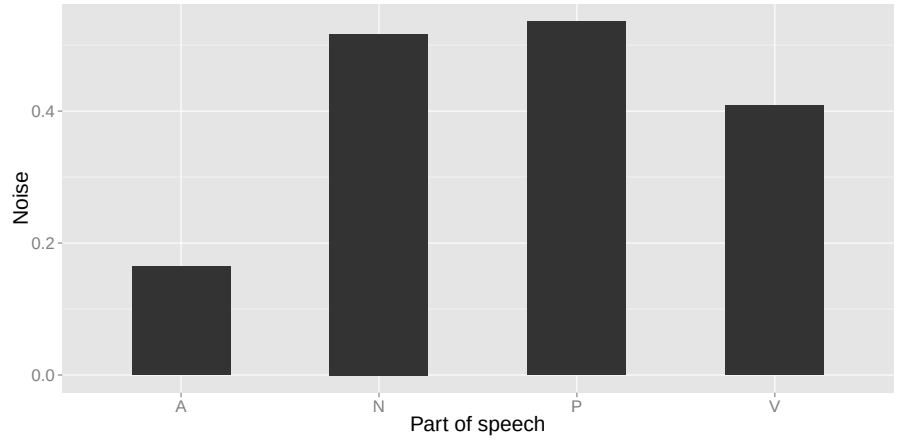


Figure 4.2: Noise averaged over the four part-of-speech categories.

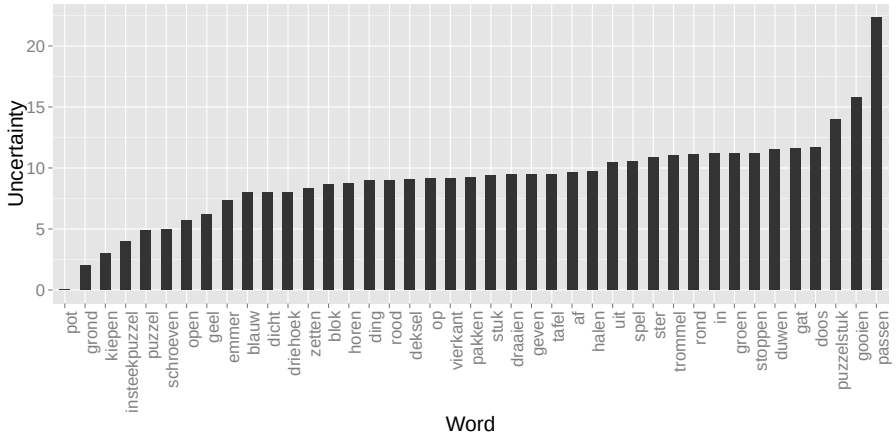


Figure 4.3: Referential uncertainty per word.

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

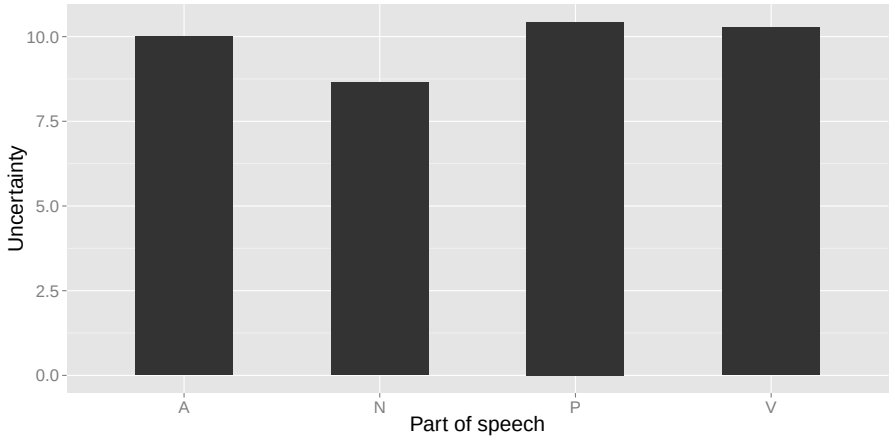


Figure 4.4: Referential uncertainty averaged over the four part-of-speech categories.

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

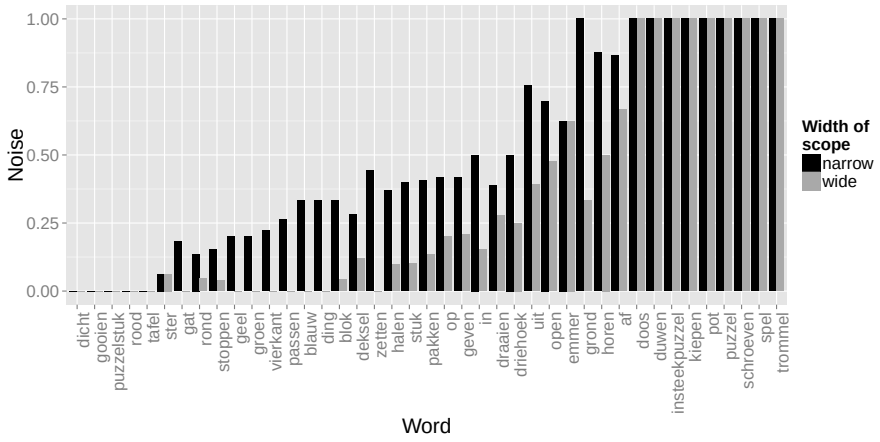


Figure 4.5: Noise per word, for both the narrow and wide set of candidate meanings.

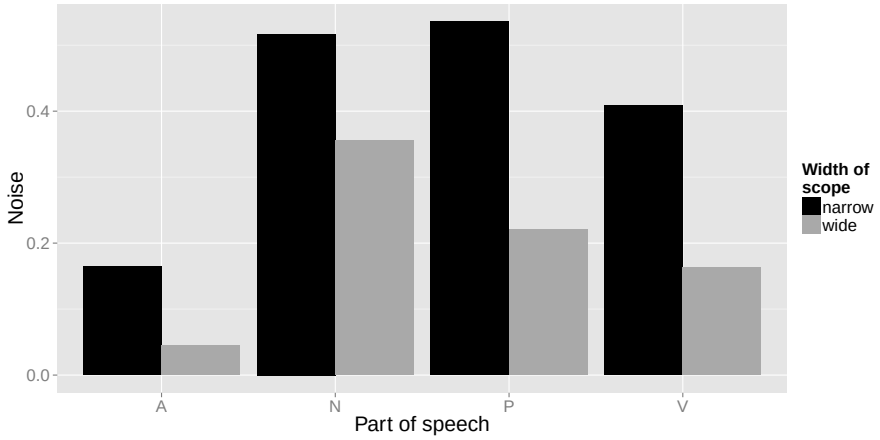


Figure 4.6: Noise averaged over the four part-of-speech categories, for both the narrow and wide set of candidate meanings.

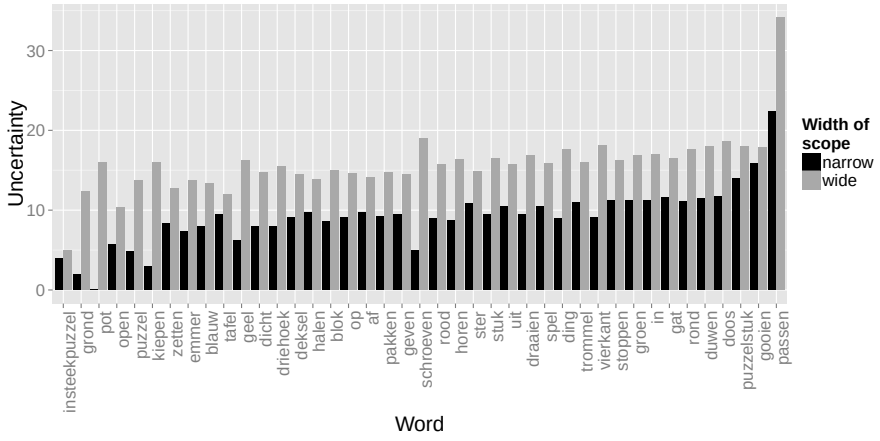


Figure 4.7: Uncertainty per word, for both the narrow and wide set of candidate meanings.

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

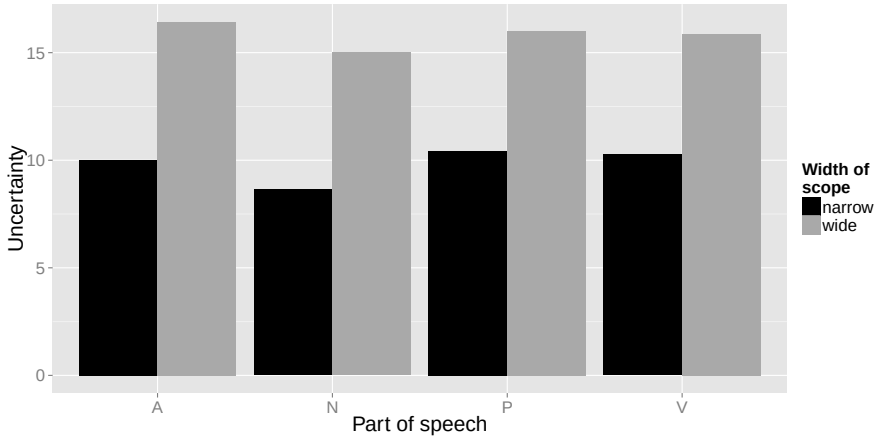


Figure 4.8: Uncertainty averaged over the four part-of-speech categories, for both the narrow and wide set of candidate meanings.

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

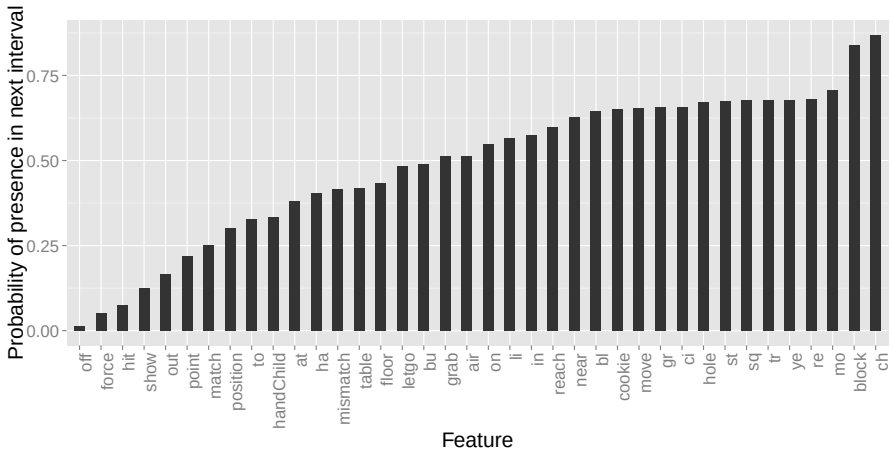


Figure 4.9: Probability of a feature being present in the next interval given presence in the current interval.

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

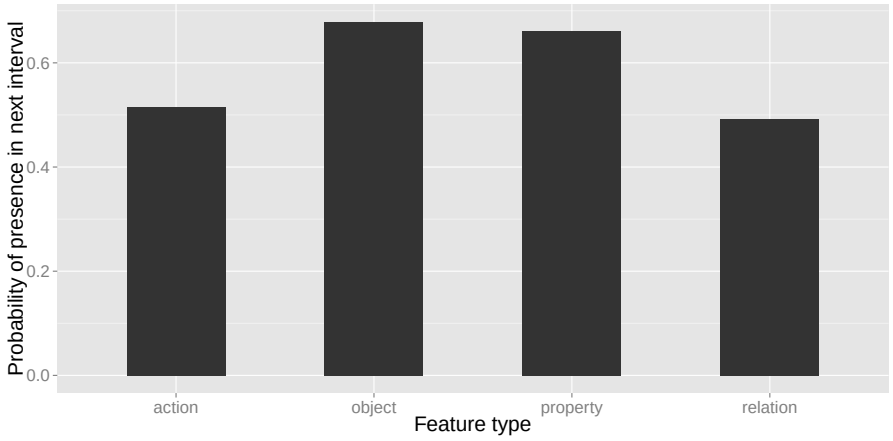


Figure 4.10: Probability of features being present in the next interval given presence in the current interval, averaged over feature types.

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-

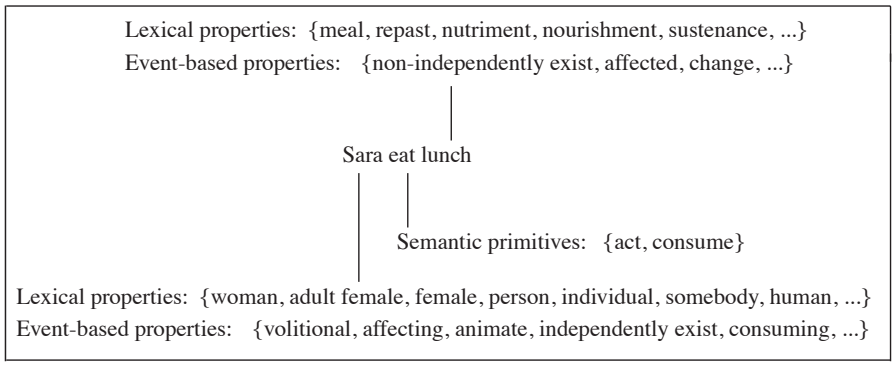


Figure 4.11: Semantic features extracted on the basis of the utterance in Alishahi & Stevenson (2010, 59).

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

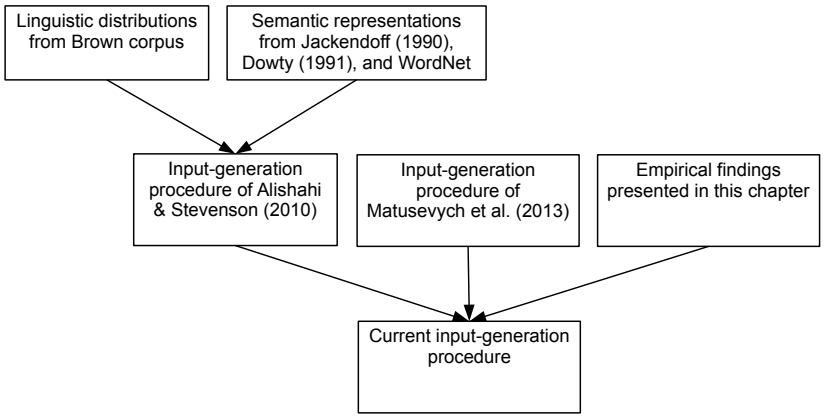


Figure 4.12: An overview of the components of the current input generation procedure.

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-

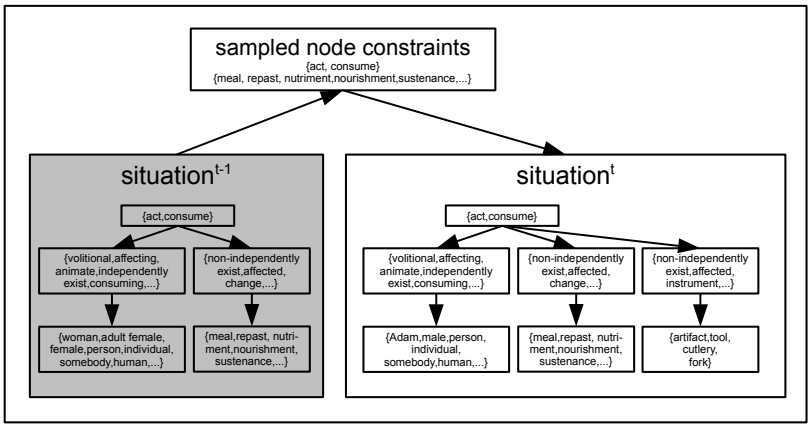


Figure 4.13: An example of sampled node constraints.

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

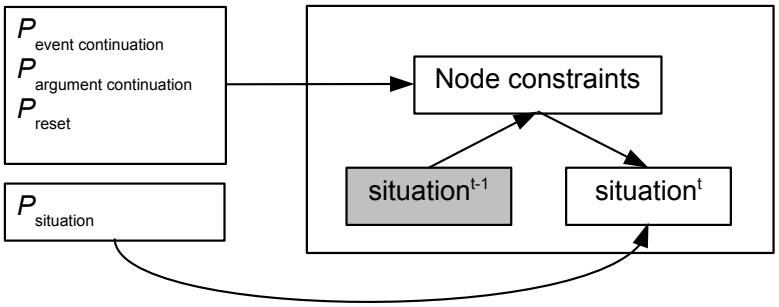


Figure 4.14: The procedure for sampling a situation.

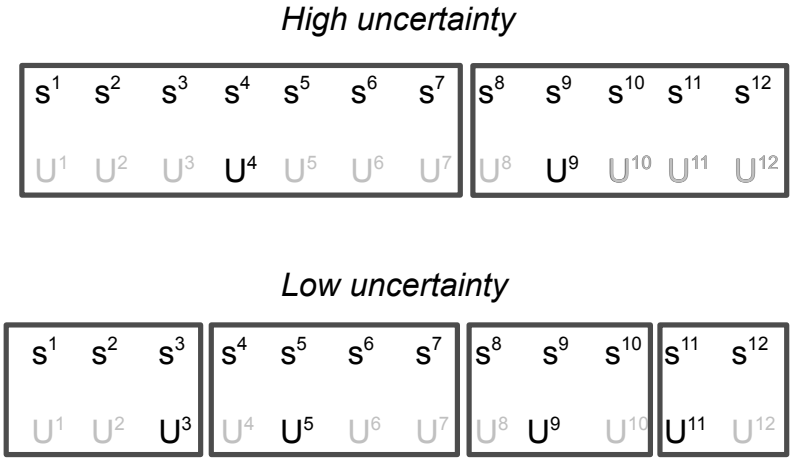


Figure 4.15: Two chains of situations, one subdivided with high uncertainty, the other with low uncertainty. 'U' denotes an utterance and 's' a situation. The grey utterances are non-selected. An input item consists of all black marked objects within one rectangle.

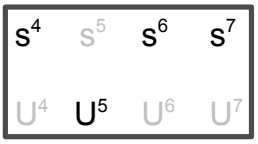


Figure 4.16: A noisy input item. 'U' denotes an utterance and 's' a situation. The situation corresponding to the selected utterance has been removed and is not part of the input item.

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.

CHAPTER 5

Comprehension experiments

5.1 Measuring comprehension

The previous two chapters set out a computational model of early grammar acquisition and a procedure for generating realistic input items. The time has come to look at the behavior of the model given these two. In this chapter, we look at the ability of the model to understand the utterances it processes. Recall that, at every turn, the model is presented with an utterance in the context of a number of situations, one of which may be the situation the speaker refers to. Can SPL, given noise and uncertainty in the situation, build up an inventory of symbolic units allowing it to comprehend the utterances? This question first requires us to define what understanding means in formal terms. That is: how do we define and operationalize 'comprehension'?

Because the input items are generated randomly, we run 10 simulations of $10,000$ input items. The latter number was established on the basis of prior testing to be the amount of input items when most scores had become stable. Recall that the referential $uncertainty$ was found to be 15 entities (events, entities) in section 4.3.2. Translating this to a number of situations, we set the number of situations co-present with the utterance to be 6 (I will henceforth call the propositional uncertainty parameter $uncertainty$). It is hard to establish a motivated number of situations, but given the overlap between situations (given the $continuation$ parameters), having six situations co-present is roughly equivalent to having 15 unique entities (not counting the roles). One of these six situations is the target situation, while the other five are distractors. Furthermore, we set the value for propositional noisr, P_{noise}, to 0.1, meaning

that in one out of ten situations, the target situation is absent.

5.1.1 General evaluation

A first measure of successful comprehension is the ability of the model to identify the target situation s_{target} out of all candidate situation S. Recall that SPL always identifies a situation $s_{\text{identified}}$ as the situation the speaker was thought to refer to. The **identification** score of an input item, then, is 1 if $s_{\text{identified}} = s_{\text{target}}$ and 0 otherwise. Because the $noise$ is set to 0.1 and the $uncertainty$ to 5 situations, there are 6 situations in the situational context S in 90% of the cases, and 5 in 10%. In that latter 10%, the model can, moreover, not retrieve the target situation, because it is simply absent. A chance baseline for **identification** is therefore $0.9 \times \frac{1}{6} = 0.15$, or one out of six for all situations in which the model can be expected to identify the target situation. Similarly, the maximum proportion of situations the model can correctly identify, or ceiling level for **identification** is 0.9, as in 10% of the cases, the target situation is not present.

The input items do not have a single correct mapping of the parts of the utterance to the target situation, and without such a gold standard, we cannot evaluate how well the linguistic analysis maps to parts of the situation. What we can evaluate, however, is what proportion of the utterance the model has processed, and what proportion of the identified situation (whether it is correct or incorrectly identified) is being mapped to by the best analysis. The first of these, **utterance coverage** is given by the proportion of the utterance U that is governed by rules other than rule **iii**, i.e., the rule for ignoring words. In other words: the proportion of U that is assigned a proper function in the analysis. Let U_{analyzed} be the substring of U that is governed by rules other than rule **iii** in the derivations underlying a_{best}. The utterance coverage can then be given by:

$$\textbf{utterance coverage} = \frac{|U_{\text{analyzed}}|}{|U|} \tag{5.1}$$

The second of these measures, **situation coverage**, works similary, but applies to the situation. The combined mappings of all constructions used in the best analysis specify a subgraph of the (correctly or incorrectly) identified situation $s_{\text{identified}}$ that is analyzed by a_{best}. Let us call this subgraph s_{analyzed}. The **situation coverage** is then given as the proportion of vertices of s that s_{analyzed} constitutes, or:

$$\textbf{situation coverage} = \frac{|V(s_{\text{analyzed}})|}{|V(s_{\text{identified}})|} \tag{5.2}$$

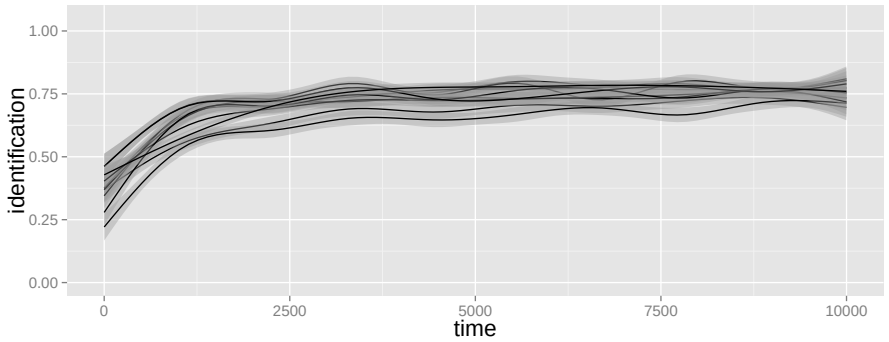


Figure 5.1: Identification scores for 10 simulations over time.

5.1.2 Evaluating the used representations

Foreshadowing the study of the representations acquired by SPL in section 6, we can also inquire what the representations are that the model actually uses. For the grammatical constructions, two interesting parameters are their length (in number of constituents) and their abstraction. From Brown's law of cumulative complexity, it follows that the inventory of linguistic representations grows more complex over time, which I take to mean that the representations become longer and the number of abstract slots increases. How this affects the choice of representations that the model actually uses in comprehension, is not evident from Brown's law itself.

Furthermore, we cannot speak of true 'evaluation' of the used representations: after all, we simply do not know what representations an actual language user employs when trying to comprehend an utterance. In section 5.3 we will look at the representations and mechanisms the model employs in analyzing input items, and compare them to hypotheses within the usage-based framework.

5.2 Global evaluation

5.2.1 Identification

As can be seen in figure 5.1, the model is increasingly able to identify the correct situation, reaching an **identification** score between 0.7 and 0.8 after $10,000$ input items, with a stabilization around 2500 items. Given a chance baseline of 0.15, the model performs well above chance, suggesting that it has learned to function relatively successfully as a communicative agent. Given a

ceiling level of 0.9, I consider the scores to be relatively close. Nonetheless, a score of 0.7 means that the model still makes a fair amount of errors (3 out of 10 cases, one of which is due to the noise).

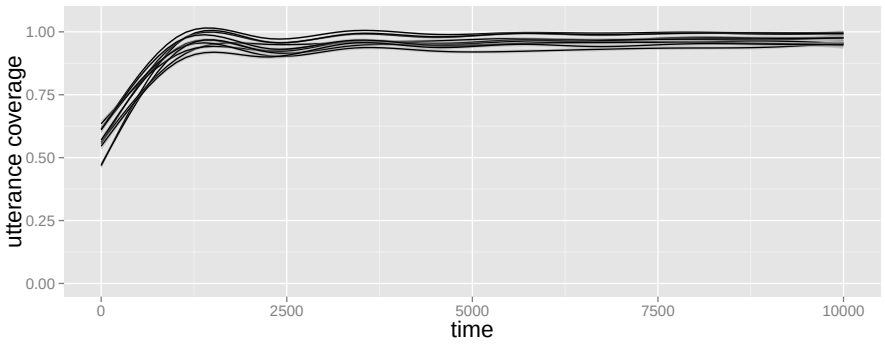
What are those cases in which the model erroneously identifies a situation as the target situation? Looking at the errors after $10,000$ input items, it seems that the only cases where the model makes errors are input items in which there are multiple, highly similar situations, and the model does not have the representational potential to tell them apart. This happens for instance when the model has misidentified attribute words like *pretty* or *happy* as markers of the role of that attribute, i.e., in an erroneous construction such as [[PERSON] [GET / *get*] [CHANGE-ROLE / *pretty*]]. When SPL has learned this construction, and next encounters two situations, one of which involves someone getting happy, and the other one involving that person getting pretty, the model is unable to choose between them, and guesses one, with a 50% chance of being correct. Other cases involve correctly learned constructions, but situations to which such constructions can equally well apply.

5.2.2 Utterance coverage

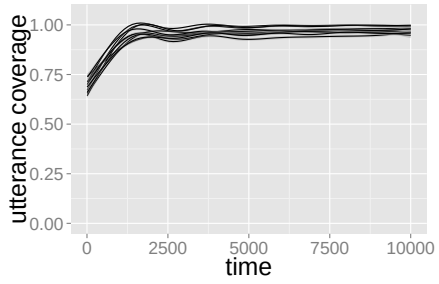
Secondly, how much of each utterance is covered by the parses at the various times? Figure 5.2a gives the results over time. The model reaches a state after approximately 1500 input items in which it is able to process almost the full utterance. It has to be kept in mind that this rapid peak may also be due to the fact that the model applies bootstrapping relatively eagerly.

When we split the values for **utterance coverage** over the correctly and incorrectly identified situations (figures 5.2b and 5.2c), we can observe that throughout the simulation, the analyses with incorrectly identified situations have lower **utterance coverage** scores. This is due to two things. First of all, there are (especially initially) several cases in which the model simply only ignores all words. Secondly, the model, in several cases, misidentifies the situation based on a partial understanding of the utterance. Given the continuity between subsequent situations, it is likely that the event and/or some participants of one situation are present in the next situation as well. When such a string of situations constitutes S, it is easy to see how the model, having understood one or two words, maps the analysis to the wrong situation.

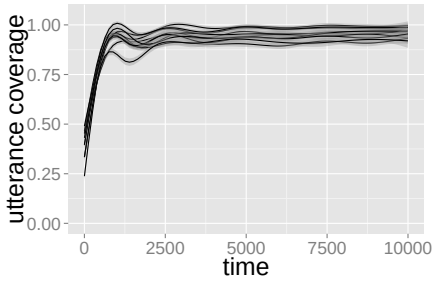
Interestingly, in all simulations, the model reaches a peak in the coverage of the utterance before suffering from a slight dip in the **utterance coverage**, from with it recovers afterwards. When we look at the scores split over correctly and incorrectly identified target situations, we can see that the peak is found slightly earlier for the incorrectly identified ones (around $1100 - 1200$) than for the correctly identified ones (around $1300 - 1400$). The dip in the utterance coverage mostly occurs at the time when the model is reaching convergence in the correct identification of the situation. This means that just before the convergence, the model is applying representations that cover more of the utterance, but do so with less success. In the next stage, the model uses slightly



(a) Utterance coverage for all input items.



(b) Utterance coverage for correctly identified target situations.



(c) Utterance coverage for incorrectly identified target situations.

Figure 5.2: Utterance coverage for 10 simulations over time.

shorter representations to analyze the utterances, covering slightly less of the utterance, but making more accurate analyses. Finally, the model starts using the longer representations again, but now in an accurate way.

What the model does here is reminiscent of a phase of syntactic creativity that is only later constrained by more 'fitting' representations. As we will see in section 5.3 below, and in the closer inspection of the learning mechanisms in the next chapter, the period around 2500 input items is also the moment when the model has just acquired abstract representations and has ceased to apply the syntagmatization operation frequently. This means that by then the potential for generalization, in the form of abstract constructions (constructions with few semantic constraints, obtained through paradigmatization), is present, and that afterwards the model 'recovers' from applying these abstractions too frequently by building up an inventory of more concrete constructions that 'pre-empt' the use of the abstract constructions in the analysis. The continuing accrual of relatively concrete constructions allows the model to overcome overgeneralization. As such, this robustness provides an argument for the apparent redundancy of storage, as many within the usage-based approach have argued (Langacker 1988, Beekhuizen, Bod & Verhagen 2014).

Let us have a look at an example that illustrates this. In one of the simulations, the model encounters, after some 200 input items, the utterance in example (29). The utterance illustrates a construction which is relatively rare (compared to other kinds of three-word utterances that are formed on the basis of a transitive construction). The optimal analysis the model assigns to this utterance is given in example (30). It involves an abstract intransitive construction and the bootstrapping of *go*. Some 300 input items later, the model encounters the same utterance, but now uses the analysis in example (31). This is a regular transitive construction, in which the action of a person on an object is expressed. With this construction, SPL erroneously takes the utterance to refer to a caused-motion event. Nonetheless, it covers the full utterance, as opposed to the analysis with the intransitive construction.

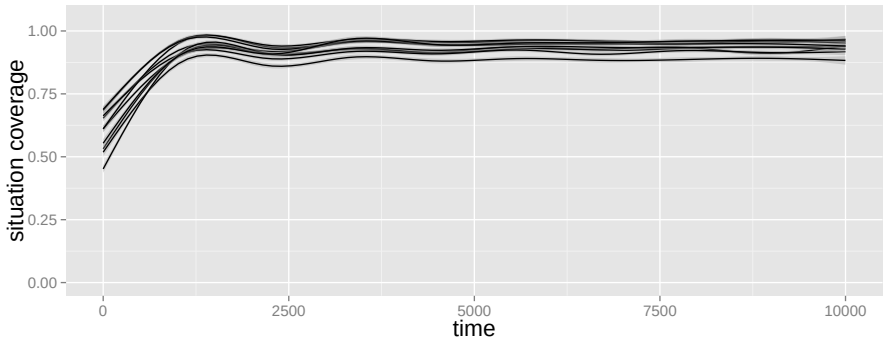
Finally, after another 300 input items, the model has an intransitive motion construction, as shown in example (32), which is combined with the known meanings of *go* and *out*. From this example, we can glean that the model eagerly applies abstract patterns to situations in which they lead to misinterpretations. These errors are overcome once a larger inventory of constructions is built up.

(29) *you go out*

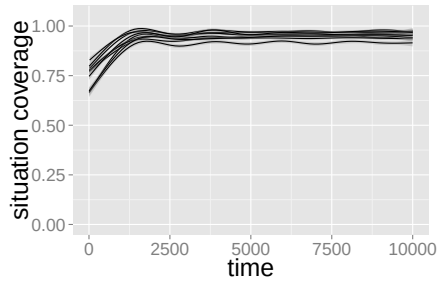
(30) [[ENTITY]→[HEARER / *you*] [EVENT]→(GO / *go*)]

(31) [[PERSON]→[HEARER / *you*] [CAUSE]→(CAUSE-MOVE / *go*) [OBJECT]→(ARTEFACT / *out*)]

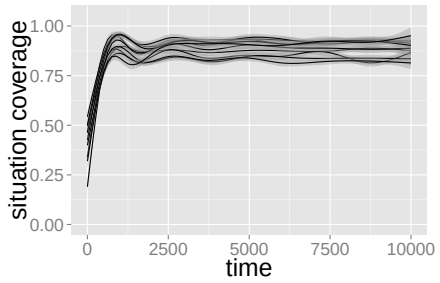
(32) [[PERSON]→[HEARER / *you*] [EVENT]→[CAUSE-MOVE / *go*] [ROLE]→ [DESTINATION-ROLE(LOCATION) / *out*]]



(a) Situation coverage for all input items.



(b) Situation coverage for correctly identified target situations.



(c) Situation coverage for incorrectly identified target situations.

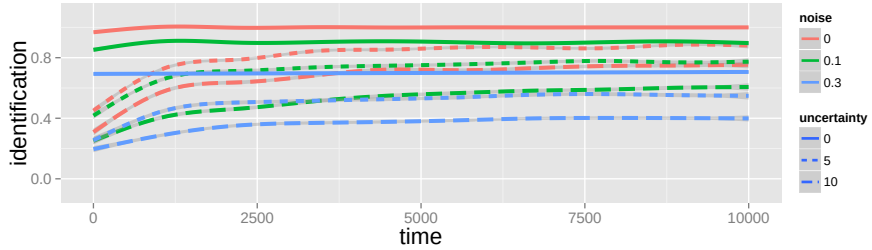
Figure 5.3: Situation coverage for 10 simulations over time.

5.2.3 Situation coverage

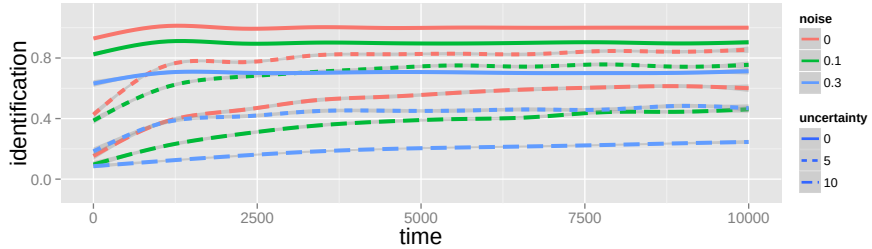
We find a highly similar pattern for the model's understanding of the parts of the situation that are being signified by the utterance, or the **situation coverage** in figure 5.3a. The model quickly achieves high levels of understanding of the situation, with a stabilization around 1500 input items.

Again, we see a difference between the correctly and incorrectly identified target utterances (figures 5.3b and 5.3c). For input items in which the model correctly identified the target situation, the situation coverage starts out relatively high (values around 0.75), whereas for input items with incorrectly identified target situations, the situation coverage starts out low (values between 0.25 and 0.50).

An interesting future step would be to have the utterance and situation coverage affect the reinforcement of the used constructions. Currently, the



(a) Identification scores for nine unique noise and uncertainty settings over time given $P_{\text{reset}} = 0.05$.



(b) Identification scores for nine unique noise and uncertainty settings over time given $P_{\text{reset}} = 1$.

Figure 5.4: Identification scores given various parameter settings.

probability of an analysis is simply penalized for not being able to parse parts of the utterance, but if an analysis involving ignored words and ignored parts of the situation is (despite this penalty) the best analysis, the used constructions receive as much of an increase as when the analysis covers all of the utterance and the identified target situation. If we allow the model to reinforce the construction proportionally to their utterance and situation coverage, erroneous analysis, and hence (often) erroneous constructions will receive less counts, and therefore be less likely to be re-used.

The important question is: would this merely be a 'hack', i.e., a trick to get the model to work better, or is it in some sense a cognitively motivated operation? I do not intend to give a definitive answer to that question, but it seems to me that the firmness of the belief that something is the right analysis is a feature that can be used by a model to 'bootstrap' itself. Furthermore, it is not a grammar-wide optimalization operation, but a local effect of the processing, and therefore still in line with desideratum D2-8 (learning-as-processing).

5.2.4 Robustness to uncertainty and noise

The parameters $noise$ and $uncertainty$ that I used in the experiments were set on the basis of the findings in chapter 4. It would nonetheless be interesting to see how the model behaves under different settings for these parameters. Furthermore, I set the probability of generating the next event without taking the previous one into account to 0.05. This means that subsequent frames are very likely to look alike. However, we may wonder how the model behaves if all frames are independently generated (i.e., $P_{\text{reset}} = 1$).

In figure 5.4a, the identification scores for nine unique parameter settings is given if we set $P_{\text{reset}} = 0.05$. For each unique parameter setting, three simulations were run. The $noise$ values were set to 0, 0.1, and 0.3, and the $uncertainty$ values to 0, 5, and 10.

Looking at $uncertainty$ first, we can see that the model trivially performs at ceiling level (given each $noise$ setting) if there are no non-target situations present. Adding uncertainty causes the model to misidentify the target situation more often. However, even with 10 non-target situations present, the model still identifies the target situation correctly in six out of ten cases under the no-noise condition (where randomly guessing would yield a score of 0.09). It might furthermore be that given high levels of uncertainty, more input items would be needed to arrive at some level of communicative competence: the slopes of the developmental curves for the settings $noise = 0.1, uncertainty = 10$ and $noise = 0.3, uncertainty = 10$ do not seem to have reached a point of convergence after $10,000$ input items (unlike the other curves).

With $noise$, we see a similar pattern. Adding more noise causes the model to learn erroneous representations and apply them, even in situations where the target situation is present. However, even with three out of ten target situations being absent, the model still identifies the target correctly, given $uncertainty = 5$, around 58% of the cases (where the ceiling level of the performance would be 0.70.

Setting the probability of reset P_{reset} to 1 causes a more variable performance of the model (see figure 5.4b). Whenever there is uncertainty present, this has a greater negative effect on the scores than when $P_{\text{reset}} = 0.05$. The reason for this is that, when the model misidentifies a situation under the condition $P_{\text{reset}} = 0.05$, it is very likely that it still has a correct partial identification: some referents, or the action given in the situation can be the same as the one in the target situation. Nonetheless, the model acquires some correct constructions even under the most dire settings for $noise$ and $uncertainty$, given that the performance with that setting (**identification** = 0.3 after $10,000$ input items) is still more than four times as high as the chance baseline for that setting (i.e., one out of eleven of seven of out ten situations, or ± 0.07).

This means that the model's performance decays gracefully under increasingly hard conditions. Even though we motivated the parameter settings, this result supports the idea that SPL is a robust learner.

5.3 Used representations

We have seen in section 5.2 that the model is able to comprehend sentences relatively well. In this section, we have a look at the kinds of representations the model uses in analyzing the input items. From a usage-based perspective, several topics are of interest: the use of unanalyzed chunk-like structures, the use of bootstrapping to analyze unseen words, the use of the concatenation operation, the abstractness of the used constructions, and the types of abstraction (over verbs, or over nouns). A computational model like SPL allows us to look at the representations used in the analyses.

5.3.1 The use of chunks

The usage-based approach claims that in many cases, language users operate with representations that could be further analyzed, but that *are* not further analyzed (Arnon 2010, McCauley & Christiansen 2014*a*). SPL learns lexical constructions without knowing what the word boundaries are. That is to say: it has the true word boundaries, but it may extract larger units as being a single word, both through bootstrapping and cross-situational learning. We can therefore expect the model to build up an inventory of such unanalyzed-but-analyzable lexical constructions. We furthermore expect the amount of chunks used in the analysis of input items to decay over time, as the more compositional constructions, used in a wider array of cases, will become stronger and outweigh the chunks. However, we can also expect the model to continue using some chunks, as even adult language users use unanalyzed-but-analyzable language material. It should be noted here that our definition of 'chunk' only covers a subset of what McCauley & Christiansen (2014*a*) consider chunks, namely the internally unanalyzed ones. As we will see, for the internally analyzed larger units (which I call 'lexically specific constructions'), we do find a behavior akin to the one reported in McCauley & Christiansen (2014*a*), viz. that their importance in use increases over time.

All of these expectations are found in the behavior of the model. I operationalize the notion of 'chunk' to be any construction in which there is at least one constituent consisting of more than one word. This includes constructions with more than one constituent, but for which (at least) one of the constituents has more than one word as their phonological constraint. Figure 5.5 shows the frequency of the use of chunks over time. For the first 750 input items, the model employs chunk-like constructions relatively frequently, after which the number drops, but remains stable at around 4 used chunks per 250 input items.

What are the chunks the model uses? Table 5.1 gives the most frequently used chunks for three simulations. We can see that in simulation 0, there are mainly many chunks with *play with*. The chunk has been syntagmatized with entity words like *matches* and *truck* to form grammatical constructions involving the chunk. Note that all chunks in simulation 0 are 'correct', in that they

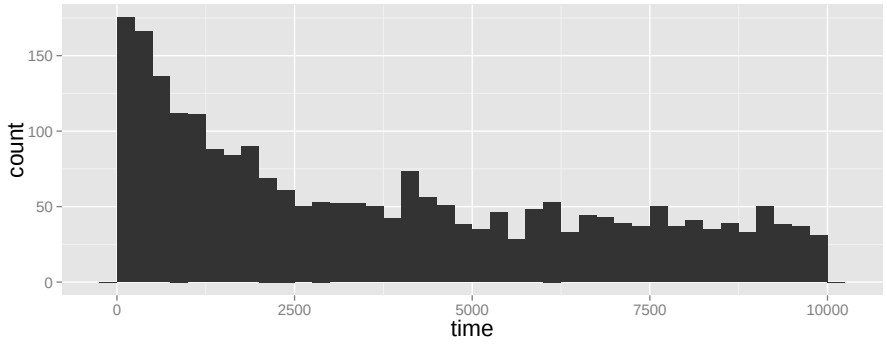


Figure 5.5: Frequency of the use of chunks over time (summed over simulations).

rank	simulation 0
1	[PLAY(AGENT,PATIENT) / *play with*] (265)
2	[PUT(AGENT(SPEAKER)) / *I put*] (33)
3	[COME(AGENT,DIRECTION-ROLE(LOCATION)) / *out come*] (25)
4	[[PLAY / *play with*] [MATCHES / *matches*]] (11)
5	[[PLAY / *play with*] [TRUCK / *truck*]] (8)
rank	simulation 4
1	[THEY / *here come*] (8)
2	[SPEAKER / *they make*] (7)
3	[BABY / *baby take*] (3)
4	[SPEAKER / *we go*] (1)
5	[[SPEAKER / *we go*] [SEE / *outside*]] (1)
rank	simulation 7
1	[[PUT / *put them*] [DESTINATION-ROLE / *in*] [ARTEFACT]] (78)
2	[GIVE / *she give*] (46)
3	[[SIT(SURFACE-LOCATION) / *sit on*] [ARTEFACT / *it*]] (44)
4	[PUT(AGENT(HEARER)) / *you put*] (29)
5	[[SIT / *sit on*] [SPEAKER / *me*]] (27)

Table 5.1: Most frequent used chunks for three simulations.

do seem to capture the meaning of the words they contain.

The extraction of *play with* as a chunk is interesting. SPL has acquired *play with* with the meaning PLAY(AGENT,INSTRUMENT). This can be considered to be an error, but the word *with* occurs in the input generation procedure only in one other, highly infrequent, construction, namely the [[ENTITY] [CREATE / *make*] [ENTITY] [SOURCE / *with*] [ENTITY]] construction (e.g., *I made a cookie with dough*). The meaning of *with* in this construction is furthermore different from that in *play with*. Therefore, the model 'decides' to use *play with* essentially as a bi-syllabic word denoting the action PLAY and its roles.

However, *play* is also used without *with*, in utterances like *you play game*. This gives the model the opportunity to learn the meaning of *play* by itself, which it does: it also has a [PLAY / *play*] construction. However, as the *play with*-construction covers more of the utterance, it is given preference over the lexical *play*-construction in the analysis of sentences containing the substring *play with*. We again find that the *play with*-chunk, and its syntagmatic extensions are used thoughout development.

Table 5.1 also shows us that there is massive variation between simulations. Both simulation 4 and 7 display less use of chunks than simulation 0. This is interesting, as it gives us the well-known difference between analytic and holistic learners (Bretherton, McNew, Snyder & Bates 1982) without parameters governing that particular behavior. That is: it is not due to a change in the model that different amounts of chunk use are found in different simulations. Rather, it is merely an effect of input order, and the chance of the subsequent co-occurrence of certain utterances. If two utterances with *play with* in it are found subsequently with different arguments, the model will extract a *play with*-construction. Perhaps this can be taken to mean that the difference in what seem like learning strategies (some learner learn many chunks, while others learn few), may (also) be an effect of input order and dispersion of the input items.

Note, finally, that there is variation in the kinds of chunks the model learns in different simulations. In simulation 4, the acquired chunks all refer to the wrong entity, and hence receive little reinforcement, whereas in simulation 7, the most frequent construction involving a chunk is a semi-open construction, with *put them* as its first constituent, followed by *in*, which has its own role, followed by any ARTEFACT. Again, this is an effect of the coincidental juxtaposition of input items and the subsequent build-up of the grammar through syntagmatization and paradigmatization, which to my mind is an exciting, albeit rather extreme hypothesis following from usage-based theory that can and should be further explored using both experiments and dense corpora.

5.3.2 The use of bootstrapping

The ability to bootstrap words into open constituents of constructions is a mechanism that allows the model to analyze utterances for which it does not know all the words. Suppose that the model encounters the utterance in ex-

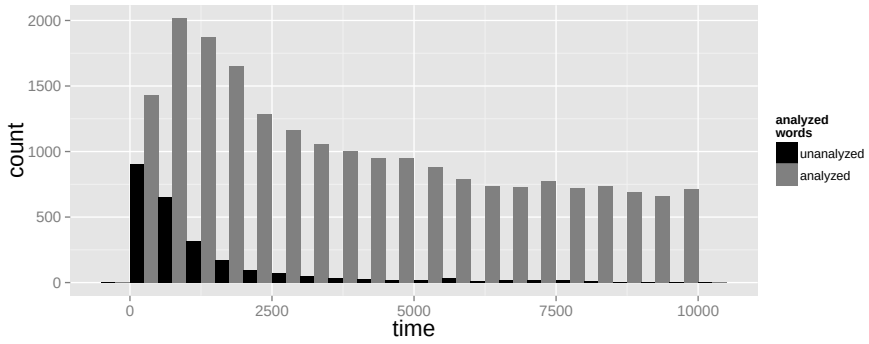


Figure 5.6: Frequency of the use of the bootstrapping operation over time (summed over simulations).

ample (33). Having access to a [[HEARER / *you*] [SIT / *sit*] [LOCATION / *on*] [OBJECT]] construction, all the model has to do is make the assumption that microphone refers to the OBJECT in the LOCATION-role of the sitting event, and it has learned a new word, as can be seen in example (34), which gives the best analysis of example (33).

(33) *you sit on microphone*

(34) [[HEARER / *you*] [SIT / *sit*] [LOCATION / *on*] [OBJECT] → (MICROPHONE / *microphone*)]

Bootstrapping provides a strong mechanism for interpreting and acquiring novel lexical constructions. However, the risk of allowing for an operation like bootstrapping is that the model will bootstrap too freely, assigning meanings to word forms that already have well-entrenched meanings associated with them. When we look at the number of bootstrapping operations over time (figure 5.6), several things can be observed. First of all: the number of bootstrapping operations decreases over time, consistent with the idea that the learner has increasingly many (lexical) constructions in her inventory. This can be expected, as the expected number of novel, unanalyzed words decreases over time (cf. figure 5.7). However, whenever a novel word type is encountered, it is most likely to be bootstrapped into a slot of a (semi-)open grammatical construction.

More interestingly, the amount of bootstrapping operations over words that have been analyzed before (i.e., for which there is a constructional representation in SPL's grammar) decreases less rapidly than the amount of bootstrapping operations over unanalyzed words. This means that the model bootstraps words for which it already has a representation. In some cases, this

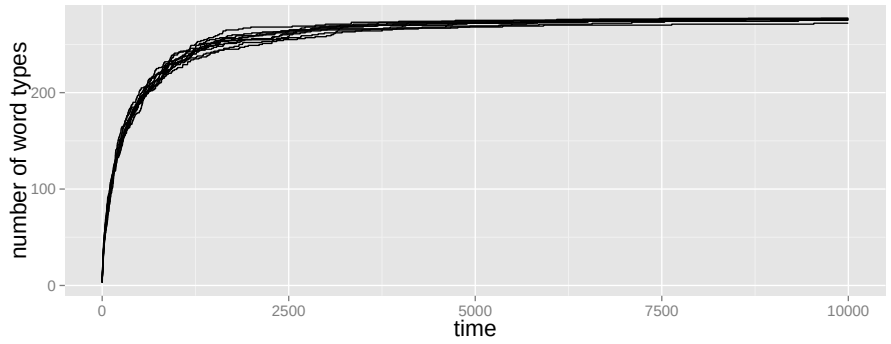


Figure 5.7: Number of observed word types for 10 simulations over time.

happens in noisy input items (i.e., items in which the target situation is absent). When encountering the utterance *you make picture*, but the meaning MAKE(MAKER(HEARER(YOU),MADE-THING(PICTURE)) is absent, but another situation MAKE(MAKER(HEARER),MADE-THING(COOKIE)) is present, the model will bootstrap the word *picture* as meaning COOKIE. This is not very problematic for the model, as the bootstrapped construction [COOKIE / *picture*] will rarely if ever be reinforced in other input items. However, this does point to a design feature of the model that might be too strict, namely that it is forced to to select a situation. If the model has a strong conviction the *picture* means PICTURE, having no situation present that contains that conceptual element should ideally force the model to consider the input item to be noisy and not consider the analysis in which *picture* refers to PICTURE to be the best one.

In other cases, the grammatical construction used to bootstrap the word is erroneous. In the same simulation, the model has acquired a construction [[PERCEIVE / *you look at*] [PERCEIVER-ROLE]], where the second constituent refers to the PERCEIVER role. When encountering *you look at picture*, the model considers *picture* to refer to that agent role (as if it were a nominative case marker, essentially), and bootstrap a construction [AGENT-ROLE / *picture*]. Again, this construction will be used in few subsequent analyses and therefore not be reinforced, but it does lead to an erroneous analysis of the mapping between the utterance and the identified situation. Here too, the fact that the model does not take the reinforcement of a [PICTURE / *picture*] construction can be considered a weakness in the design of the model.

This analysis gives us an insight in the complex interaction of mechanisms that must take place when a word is being bootstrapped. On the one hand, we have the selection preferences of a slot of a construction and the number of other elements that can fill that slot. The higher this number is, the lower the

probability of bootstrapping something new in the slot. Interestingly, this idea runs counter to Bybee's (2006) ideas about high type frequencies (many other constructions being able to fill a slot) making a slot *more* extendable. However this works, there is a top-down effect of the slot of the construction. On the other hand, there are bottom-up effects of the word. If the hearer knows with a lot of certainty that a word already refers to very different meanings, he would find it very unlikely that the speaker uses it now to refer to this particular concept. It would be as if a speaker and a hearer are looking at a painting, and the speaker says *what a nice book*. The hearer would not, in this situation, bootstrap *book* in the open slot of a *what a nice*-X construction, because the word form *book* is already used in lexical constructions referring to BOOKs. In this case, the hearer would come to the conclusion that the speaker is an uncooperative communication partner. However, if the speaker said *what a nice fammer*, the hearer would be prone to bootstrap the meaning of the word *fammer* as relating to something concerning the painting or maybe an object depicted in it. Finally, there are cases where the use of a word seems like an extension of the meaning. When the speaker says *what a nice Vermeer*, and the hearer does not know that one can use the name of an artist metonymically for the product of their artistry, the hearer can still make the inferential step that *Vermeer* refers to a product of Johannes Vermeer. A new lexical representation is then added, linking *Vermeer* to the concept PAINTING-BY-VERMEER. The bottom-up effects of the bootstrapping thus also concern the closeness of the bootstrapped meaning to one of the known meanings, but this is likely the way radial concepts in lexical meanings emerge (cf. Lakoff 1987).

Concluding, the bootstrapping operation as implemented in SPL is a naïve one, that does what it should do, namely learn new words, but that also applies too frequently and in an underconstrained way. A possible solution is to not only take into consideration the top-down preferences of the slots of the grammatical constructions, but also the bottom-up knowledge concerning the other constructions in which the word form is already used.

5.3.3 The use of concatenation

Concatenation is the processing mechanism that allows the model to form a more encompassing interpretation of an utterance on the basis of partial analyses. SPL explicitly frames concatenation as a back-off device for cases when no better (i.e., construction-based) analysis can be found: the probability of the rule leading to a concatenation is a small, smoothed probability depending on the number of constructional analyses that can be given. As such, we can expect its use to decrease over time. Figure 5.8 shows that this is indeed the case: the number of concatenations decreases over time.

A successful case of concatenation is given in the analysis in example (35). When processing the utterance *you put animal in it*, SPL uses a construction [[HEARER / *you*] [PUT / *put*] [ENTITY] [GOAL-LOCATION / *in*]], which is combined with the lexical [ANIMAL / *animal*] construction. This derivation

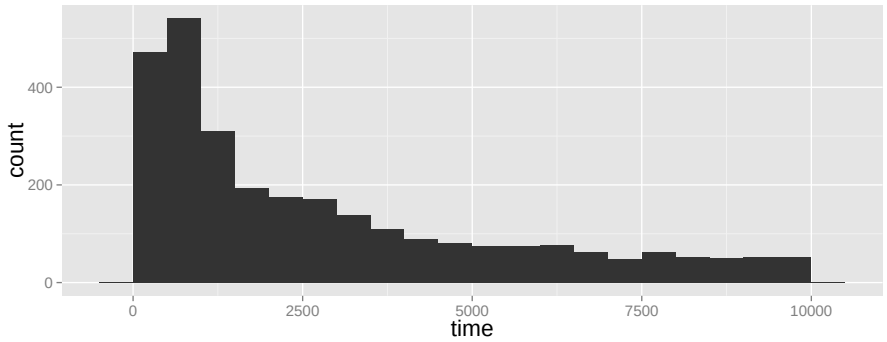


Figure 5.8: Frequency of concatenation operations over time (summed over simulations).

is finally concatenated with the [THING / *it*] construction. Note that the concatenation has meaning beyond the sum of the elements: the THING-meaning is bound to the referent filling the GOAL-LOCATION-role.

(35) ([[HEARER / *you*] [PUT / *put*] [ENTITY]→[ANIMAL / *animal*] [GOAL-LOCATION / *in*]] [THING / *it*]) |
PUT(PUTTER(HEARER),PUT-THING(ANIMAL),
GOAL-LOCATION(THING))

The design feature of concatenation as a back-off device can be doubted, however. Perhaps using something akin to concatenation is a regular way of processing utterances (cf. Frank et al. 2012), in which case the probability model would have to be adjusted. However, there still seems to be a difference between non-conventional concatenation, as implemented in SPL, and conventional but non-hierarchical processing. I leave it to the proponents of the strong non-hierarchicality thesis to develop a working model that involves meaning.

5.3.4 The length and abstraction of the used representations

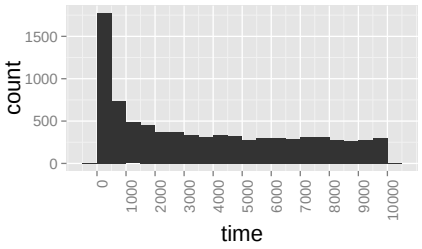
At the heart of the usage-based perspective on the acquisition of grammar is the claim that grammatical representations are built up in a gradual, bottom-up fashion. As I argued earlier, this is a cognitive take on Brown's (1973) (observational) law of cumulative complexity, which holds that more complex representations emerge in development after, and on the basis of, simpler ones. For grammatical representations, we can take this to mean that the representations become increasingly long and increasingly abstract. However,

we can wonder if this implies that the *used* representations become more abstract. After all, the language-learning child also encounters more concrete instances of grammatical patterns, which, under the usage-based perspective, leave traces in the mind as well. Here, the old pair 'competence and performance' (Chomsky 1965) comes in handy. Even within a usage-based model, the potential of a model may differ from what is doing most of the time. Whereas a model may have acquired the representational potential to make all sorts of generalizations, it may be the case that the more abstract ones are only needed in few cases, because the more concrete representations pre-empt the use of the more abstract ones in use. The competence of the model is then, of course, something derived from, or immanent in the processing involved in the performance, but conceptually, we can describe the learner's global competence distinctly from its performance in specific cases. Again we see a case of a conceptual or analytical distinction that is ontologically non-distinct, but may methodologically or analytically be separated. Applied to a usage-based perspective, it furthermore corresponds to the distinction between a static and a dynamic take, where the competence describes the state of the language user's potential and the performance the actual use in processing of that competence.[1]

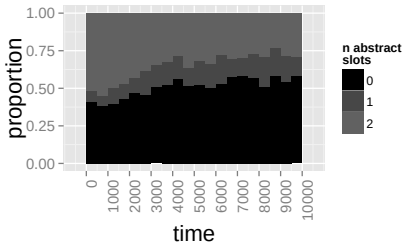
Let us start with SPL's performance first. In chapter 6, we explore the competence side in more depth, but here, we look primarily at the nature of the constructions that the model *uses* to analyze the utterances. In figure 5.9, the frequency of constructions of various length and abstraction over time is given. The first thing worth noticing is that the longer representations are only used to the full extent in the $1500 - 2000$ bin for length-4 constructions and the $3000 - 3500$ bin for length-5 constructions. Length-2 constructions are used 'too much' over the first 2000 input items, which is when they are used to analyze utterances for which length-3, 4 or 5 constructions would be most suited. We see a similar pattern for length-3 constructions, being overused in the $500 - 1000$ bin. All in all, this means that early on, SPL analyzes input items with longer utterances by means of shorter representations, concatenation, and ignoring words, and that the development to higher-arity constructions depends on the use of these lower-arity constructions.

For the abstraction of the used representations, the analysis is slightly more complex. The end state, after $10,000$ input items, is that the longer the used representation is, the higher the chance of it being a more abstract one. About

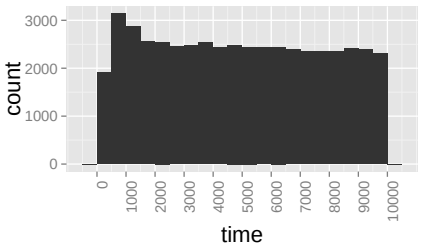
[1] Allowing myself a small digression: the idea that competence and performance are 'implemented' in the minds of language users as distinct 'things' can be seen as a case of the reification of an analytical distinction into an ontological one whereas the distinction may equally well be viewed as two perspectives on the same object. As such, it constitutes a case of Gigerenzer's (1991) tools-to-theories heuristic, in which tools of analysis shape the conception of the objects of study. Vice versa, going from the denial of this ontological distinction to a strict what-you-see-is-what-you-get approach (more formally: the analyst's inference of the most likely grammar on the basis of behavioral patterns) is equally fallacious as it misses the logical possibility that the learner has a more abstract representational potential, but simply not uses it because more concrete constructions pre-empt the abstract ones in all but few cases.



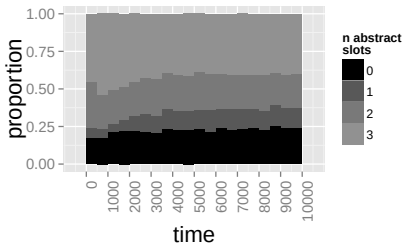
(a) Frequency of length-2 constructions.



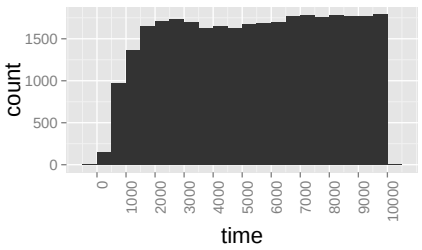
(b) Abstraction of length-2 constructions.



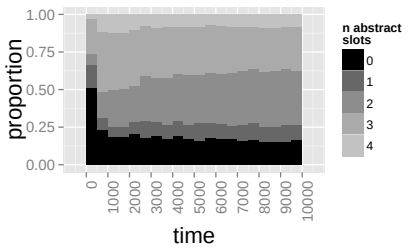
(c) Frequency of length-3 constructions.



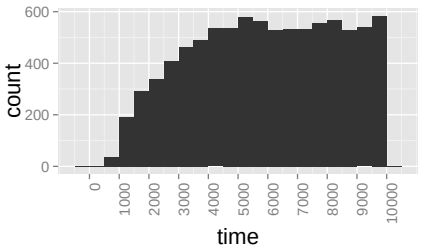
(d) Abstraction of length-3 constructions.



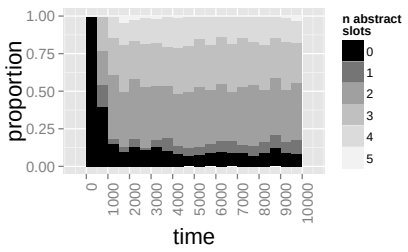
(e) Frequency of length-4 constructions.



(f) Abstraction of length-4 constructions.



(g) Frequency of length-5 constructions.



(h) Abstraction of length-5 constructions.

Figure 5.9: Frequency and abstraction of constructions of various length used in comprehension over time (summed over simulations).

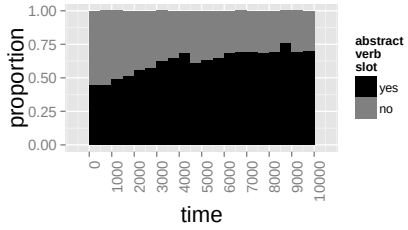
half of the length-2 constructions have no open slots, whereas for length-5 constructions this figure is around 20%. This, of course, is an effect of the kinds of utterances they are employed for. There are simply fewer unique long utterances than there are unique short ones. Nonetheless, if this effect is realistic, it has interesting consequences for the nature of the representational system. It means that the longer a construction is, the higher the likelihood of it being more abstract, all other things being equal. Perhaps this can be taken to mean that caused-motion constructions and prepositional datives have abstract representations that are more reinforced than intransitives and transitives. With the latter two being researched less intensely than the former, this question cannot be straightforwardly answered, but it would be an interesting research avenue.

Turning to the development over time, we can see that the length-4 and length-5 constructions used early on are mostly very concrete, and that they become more abstract over time. SPL employs them, despite building up an ever-growing inventory of more concrete patterns that can be re-used. To give an example, the model encountered the utterance *you put her in here* after some 9700 input items. The model has relatively concrete constructions available to analyze this utterance (e.g., [*you put* ENTITY *in* ENTITY], and even [*you put* ENTITY *in here*]), but it analyzes the utterance using the abstract construction in example (36), in which only *put* is lexically specified.

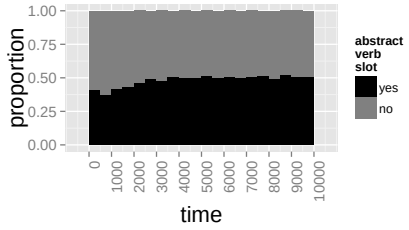
(36) [[PERSON] [PUT / *put*] [OBJECT] [LOCATION-ROLE] [ENTITY]] |
PUT(PUTTER(PERSON), MOVED-OBJECT(OBJECT),
LOCATION-ROLE(ENTITY))

Why does the model do so? Given the high diversity in sentences expressing a caused-motion event, we can expect the more abstract constructions to be the most-concrete used construction, and therefore get reinforced relatively frequently. Furthermore, the words used in these slots are also seen in many other contexts, and are therefore also well entrenched. With the high count, and hence high probability, of the abstract construction, and the well-entrenched lexical items, the analysis involving a more abstract representation thus becomes more likely than ones involving less abstract representations.The effect here is due to a dynamic version of Bybee's (2006) notion of type frequency: the more the abstract representation is actually used to analyze unseen utterances (i.e., utterances with novel word types – at least in that slot), the more it gets entrenched, and can therefore be used to analyze utterances for which in principle more concrete representations can be used.

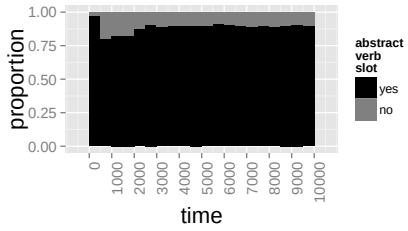
Given that the model uses a variety of concrete and abstract constructions, what are the kinds of abstraction that are useful in language comprehension? Recall that under Tomasello's (1992) hypothesis, young learners operate with verb-island constructions, consisting of verbs and their highly-specific roles. Dodson & Tomasello (1998) added the possibility of learners using argument-frame constructions, in which the arguments but not the verb is specified.



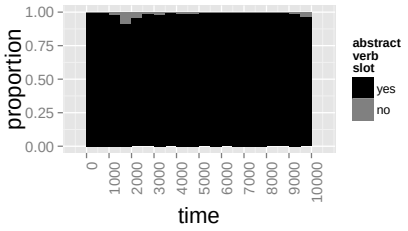
(a) Verb-islands among length-2 constructions.



(b) Verb-islands among length-3 constructions.



(c) Verb-islands among length-5 constructions.



(d) Verb-islands among length-5 constructions.

Figure 5.10: Frequency verb-island and non-verb-island constructions of various length used in comprehension over time (summed over simulations).

This especially happens with pronouns, in hypothesized constructions such as [[SPEAKER / *I*] [ACTION] [OBJECT / *it*]].

The model gives peculiar results when we look at the amount of constructions with lexically-specific verbs slots being used (figure 5.10). The length-4 and length-5 constructions that the model uses initially all have verbs specified. Afterwards, the model discovers that there is regularity in the variation (e.g., you can PUT a ball on the table, but also TAKE it from the box), and some constructions with abstract verb slots are used. However, with the increasing accrual of more concrete patterns that (crucially) involve a specific verb, the model reverts to using verb-island-like constructions. These verb-island constructions potentially have all other slots of the construction being abstract (as in example (36) above), but the verb is fixed. Before jumping to conclusions, it should be said that the model has a low number of verbs occurring in length-4 and length-5 constructions, and the type frequency of the verbs in this slot therefore is rather low. Perhaps if the model were exposed to a wider array of verbs in these slots, it would use constructions with abstract verb slots more often.

When we look at length-3 constructions, next, we see that the abstract-verb constructions form a majority. Again, I believe this is an effect of the nature of the distribution of verbs. As many verbs occur in length-3 constructions, and as several more are at least at some point *used* in length-3 constructions in comprehension, the constructions with abstract verb slot receive more reinforcement, and are hence more likely to be used later (again, despite there being more concrete patterns in the models representational potential as well).

Length-2 constructions, finally, look more like length-4 and length-5 constructions than length-3 constructions. Initially, the model uses some abstract-verb constructions, but these are given up in favor of verb-island constructions later. Here too, I believe this effect is due to the nature of the distribution of the verbs in the input: there are few verbs that occur in the intransitive pattern, and therefore the model finds little use for a general intransitive.

5.4 Desiderata and explananda

In chapter 2, I set out several desiderata for a usage-based computational model. In chapter 3, I presented the Syntagmatic-Paradigmatic Learner that was intended to meet these desiderata. Using the parameter settings obtained through the study in chapter 4, the present chapter constitutes a first evaluation of the model in terms of its behavior in comprehending utterances. These parameter settings are both stricter than most models' (there is more noise and uncertainty; cf. the comparison in section 4.2.1), and allow for more informative sets of candidate situations because of the overlap between situations. The overlap between situations constitutes a problem in identifying the correct target situation, but also makes failing to do so less problematic – when the model identifies the wrong situation, it still gets some mappings between the utterance and conceptual elements right.

Given this input procedure, we have seen in this chapter how SPL is increasingly able to understand the input items it processes. Not only does it correctly identify the target situation more frequently (around $70-80\%$ of the cases after $10,000$ input items, given a baseline of 15% and a ceiling of 90%), it also is increasingly able to analyze the full utterance and map it to many elements of the situation. Looking at the mechanisms and representations used by the models provides insight in the way SPL achieves this. By recognizing multiple words and concatenating them, the model is able to understand larger parts of the utterance. The trace these analyses leave, via syntagmatization, leads to the first grammatical constructions, which are then abstracted over if multiple similar ones have been seen. The (semi-)abstract constructions further bolster the potential for analyzing input items by allowing for novel combinations of constructions, but also by enabling the model to interpret unseen words through bootstrapping.

We are now in the position to evaluate the model against several of the desiderata and explananda. In chapter 3, I argued how the model *in princi-*

ple satisfies these, but we would like to know if that promise is made true by the behavior of the model. Concerning desideratum D2, being able to do both comprehension and production, we have seen that the model performs well in the comprehension experiment given a level of noise and uncertainty that is higher than that of most models, but with a set of candidate target situations that consists of highly similar situations, thereby aiding the model as well. This points to an often overlooked aspect in the discussion of referential uncertainty: even if the child picks out the wrong situation, or the wrong conception of a situation (as in multiply perspectivizable events, e.g., chase/flee), it will get many other things right, which eventually helps the language-learning child in verbally getting off the ground.

We have seen some remarkable effects of the quantitative grounding (D4-2) of the representational system in the usage events. Besides the obvious effects of entrenchment of more frequently processed representations, we found that the used length-3 constructions tend to be more abstract than the other constructions. I identified two reasons for this. First, the number of verb types in the 'transitive' construction is simply higher than that of the other constructions. Second, many length-4 and length-5 constructions develop from the length-3 constructions (using a 'transitive' construction, a lexical item and concatenation). The effect of this is that even more verb types are observed in length-3 constructions, thus reinforcing the abstract representation of this construction further.

This brings us to the cumulative complexity observed in the model (D6-1). We have seen that the longer constructions emerge later in development and are formed on the basis of shorter representations with concatenation. Abstract constructions show up rather early in development, but their use becomes increasingly constrained by more concrete ones, unless the abstract construction and the lexical constructions filling the slots are reinforced to such an extent that they outweigh the use of more concrete representations. As such, SPL is an avid generalizer (cf. Naigles et al. 2009), but I do not consider this property to be contrary to the usage-based perspective. It may be the case that language-learning children are not conservative in forming abstractions, but rather that their use of abstractions becomes increasingly constrained by the growing inventory of more concrete constructions. At the stage where the model has abstractions, but not many concrete constructions 'pre-empting' them, abstract patterns are used. This may also be the stage where overgeneralizations are found, a topic to which we will return in chapter 7. It seems that the distinction between a learner's (usage-based) competence and her performance is a relevant conceptual distinction: a potential for abstraction does not entail its use. This view, again, is not at odds with the usage-based perspective: all representations and their degree of entrenchment is still grounded in the experienced usage events.

A further property of SPL in which it differs from the other usage-based computational models, but similar to Kwiatkowski's (2011) model is that it acquires an inventory of both lexical and grammatical constructions at the same

time (D3). Unlike in Kwiatkowski's model, the set of grammatical representations is unconstrained, but SPL fares well in solving this daunting task. All processing and learning mechanisms involved are needed for this task: cross-situational learning to get the model started, various forms of reinforcement to find out which representations are the most useful, concatenation to build up the grammatical constructions, abstraction to generalize, and bootstrapping to acquire lexical constructions quickly.

Despite the availability of all mechanisms at all times, some are used more in early development than others. When we look at the processing mechanisms, we can observe that the bootstrapping operation peaks in frequency early, but not at the beginning of development, and that the use of concatenation decreases over time. In this sense, the model reflects the insights of Hollich et al. (2000), who argue that various cues and various mechanisms may be at work at various times in development. Again, the competence-performance distinction is insightful: all mechanisms discussed are in principle available, but it is their *use* that varies over time (D6-4).

Two of the explananda are partially satisfied in the comprehension experiment. We have observed that verbs behave conservatively in the fact that most constructions used in the comprehension process are verb-island constructions (E1). The model has the potential for using other, non-verb-specific, constructions, but does not do so, suggesting that SPL finds it more useful to structure its comprehension around verb-island constructions rather than more general verb-argument constructions. However, the model does have constructions available in which the verb is not lexically specified.

Obviously, SPL does not get everything right, and several aspects of the model are worth reconsidering in future research. One that should be pointed out here is that the current implementation is highly inefficient when an utterance is analyzed without any situational context present. We should expect the model to be able to do so at some point. Simulating this artificial situation requires some changes to the model (e.g., allowing it to build up situational representations within the space of all possible situations, rather than 2 or 5 or 10). Practically, this would be highly inefficient. Perhaps the design choice present in Chang's (2008) model, viz. to have a layer of constructional meaning first being inferred, which is only then resolved against the situational context (which, in the case of this hypothetical experiment would be absent), resolves this, and it does not seem like a major step to change the model to share this design feature. A downside of that feature, however, is that the pragmatic resolution only takes place post-hoc, that is: after the semantic analysis has been completed. A realistic analyzer would do this online. That is to say: after hearing *the man*, an analyzer would have to resolve it already in the situational context ("what likely definite reference to an instance of the category *man* can be given here?"), rather than it having to wait until the full utterance is processed.

SPL functions well as a model of acquiring communicative competence when comprehension is concerned. However, we would also like to know

how well it satisfies the second half of desideratum D2, namely production. I discuss several aspects of production in chapter 7. Before we go there, I would like to dwell on the structure of the representational knowledge of the model for a bit in chapter 6. In the present chapter, a revised take on the competence-performance distinction came to light. As one of the goals of construction grammar, or any cognitive theory of language, is to understand the representational knowledge or competence of a language user, it may be insightful to take a 'look under the hood'.

CHAPTER 6

Entering the black box

In chapter 5, we looked at the behavior of the model in understanding the input items that it processes. At several points, I referred to the idea that SPL's potential for analyzing utterances may go beyond the behavior that it shows in comprehending input items. Unlike with human subjects, a computational model such as SPL allows us to 'take a look under the hood', and find out what the inner workings of the model are. In this chapter, I explore two of these. these. First, it is interesting to inspect the frequency with which the learning operations are applied. Despite their availability thoughout ontogenetic development (cf. desideratum D6-4), their actual use may vary. What does this tell us about the actual use of the model's processing competence? Second, we look at the representations learned by the model. Recall from chapter 5 that it may be that the model uses only a limited subset of all representations it has acquired. In that chapter, I suggested that this be taken as the usage-based instantiation of the (representational) competence-performance distinction. In this chapter, we look at the representational competence of the model.

6.1 Learning mechanisms

We can inspect how frequently the various learning mechanisms are applied by the model. A first reason to do so, is that it provides us with further insight in the way the model works. Can the application of learning mechanisms for instance be linked to the law of cumulative complexity? Furthermore, any patterns we detect in the application of the learning mechanisms can inspire novel hypotheses about the course of language acquisition in the child.

6.1.1 Lexical learning

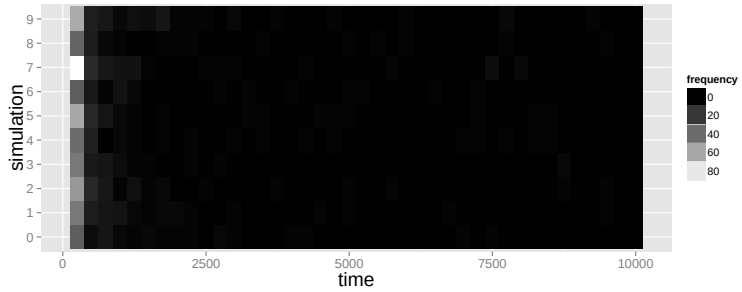
The hypothesis that the available mechanisms vary in their importance has been framed most clearly by Lila Gleitman in various publications (Gleitman 1990, Gleitman et al. 2005). Although cast within a nativist framework, the idea can be easily transferred to a usage-based one. In Gleitman's account, simple associative learning is a capacity available at any time in ontogeny, but its use may be restricted to early development. Afterwards, after all, the learner has acquired several grammatical representations that it may use in a top-down way to analyze a substring of the utterance for which it does not have a lexical representation yet. Gleitman calls this 'syntactic bootstrapping', and the process is instantiated in SPL as the bootstrapping operator of rule **vi**, whereby any phonological string can be fit into an non-phonologically specified slot of a construction. If the analysis involving the application of bootstrapping turns out to be the best one, a lexical construction containing the bootstrapped phonological string is added to the grammar.

When we look at the relative importance of the various operations involved in the acquisition and reinforcement of lexical constructions (figure 6.1), we can see a very similar picture to Gleitman's emerging. Light-colored cells depict a high amount of applications of the learning mechanism, and dark-colored cells a low amount. I counted an application of cross-situational learning, bootstrapping and adding a most-concrete construction only if the representation with which the grammar was updated was not already in the grammar. In other words: I counted the first three mechanisms only if they gave rise to a novel representation.

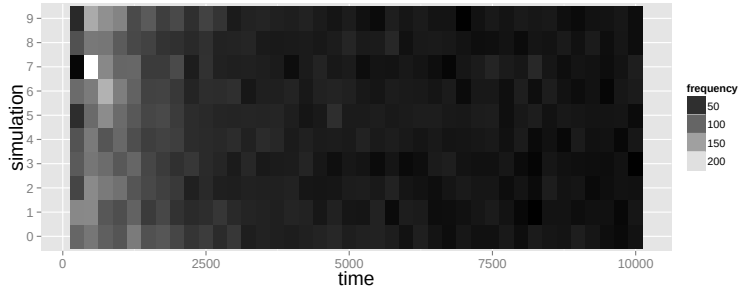
Simple, associative cross-situational learning is used only in the very early stages, up until about 250 input items, after which it completely falls out of use. After having processed very few input items, the model seems to have built up a repertoire of grammatical constructions allowing it to bootstrap novel lexical constructions. This mechanism remains being used by the model to obtain novel lexical constructions throughout development, although less frequently (recall that the model has seen almost all word types after some 1500 input items). This means that over the whole of development, most lexical constructions are obtained by bootstrapping them on the basis of the linguistic knowledge applied to the rest of the utterance rather than by a form of cross-situational learning.

The mechanism whereby the model adds a new representation on the basis of the most-concrete construction given an existing lexical item rarely occurs. This does not come unexpected: most words have a fixed set of semantic features, and hence abstractions over words are typically not very useful to the model. Hence, these abstractions are few, and so are any novel most-concrete construction **mcc**s learned on the basis of analyses involving these abstractions.

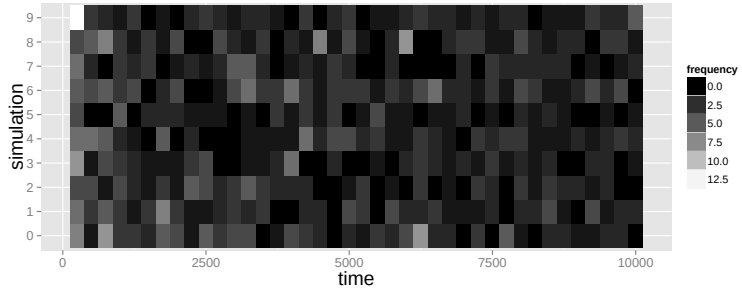
Of course, one caveat here is that I only implemented one form of cross-situational learning. Nonetheless, I believe this result provides us with an in-



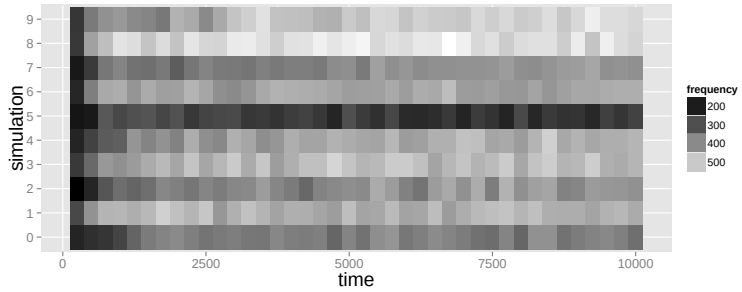
(a) Cross-situational learning.



(b) Bootstrapping.



(c) Update of a lexical most-concrete construction.



(d) Reinforcement of a lexical most-concrete used construction.

Figure 6.1: Frequency of learning mechanisms involved in the acquisition of lexical constructions over the first 1000 input items.

teresting line of further study, namely the exploration of the ways in which lexical constructions, or words and their meanings are acquired and the question which sources of information are used *over developmental time*. The results from SPL, in line with Gleitman's idea, suggest that a combination of knowledge of the rest of the linguistic structure with some form of top-down processing, may be dominant in later development, whereas associative learning may prevail earlier on.

An interesting pattern, finally, that we can glean from these graphs, is that in simulation 5, relatively few reinforcements of the most-concrete used construction are made. As we will see, this is because the model reinforces most-concrete used *grammatical* constructions instead. I postpone the analysis of this observation to the next section.

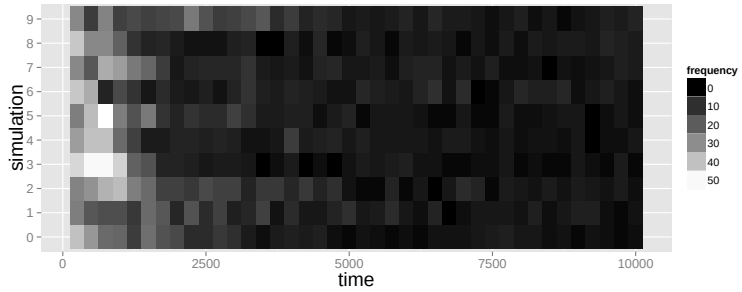
6.1.2 Grammatical learning

As for the acquisition of lexical constructions, we find variation in the frequency of use of various learning mechanisms for grammatical constructions over time (figure 6.2). Syntagmatization is mainly found in early development, after which SPL starts abstracting over the obtained grammatical representations. Later syntagmatization operations likely involve the extension of three-argument to four-argument patterns, and we will look at this more closely in the latter two sections of this chapter.

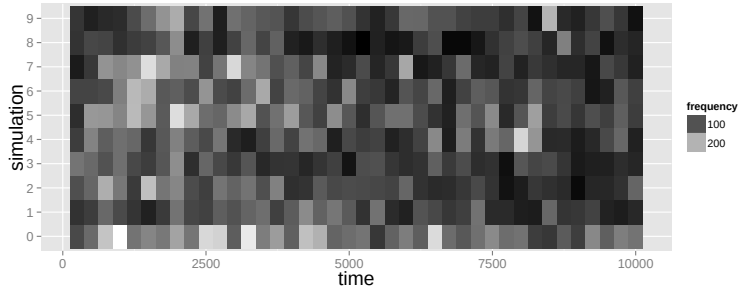
Learning from most-concrete constructions is also a learning mechanism that takes place mostly early in development, but its use over time decays slower than that of syntagmatization. Recall that with the addition of a most-concrete construction **mcc**, the model creates a trace of the processed exemplar. As novel input items (i.e., input items that – as a whole – have not been seen before) will be presented to the model throughout development, adding a trace of the analysis of that novel input item is something the model will keep doing. Of course, the number of novel utterances will decay over time, and because of that, the amount of **mcc**s.

An interesting finding for abstraction is that, unlike the other mechanisms, its application is not smoothly distributed over time. Syntagmatization and the acquisition of novel representations by most-concrete constructions are frequent early on, and gradually decay over time. Abstraction, however, seems to take place in bursts. What happens here, is that when SPL encounters an analysis with a novel grammatical construction, for instance through adding an **mcc**, this pattern may trigger a number of abstractions, with various other constructions. These bursts are suggestive of a developmental pattern Kwiatkowski (2011) models, namely, the non-gradual development of the learner's production. Similar bursts in the model's potential will be seen in chapter 7, where we discuss how the model generates utterances on the basis of a situation.

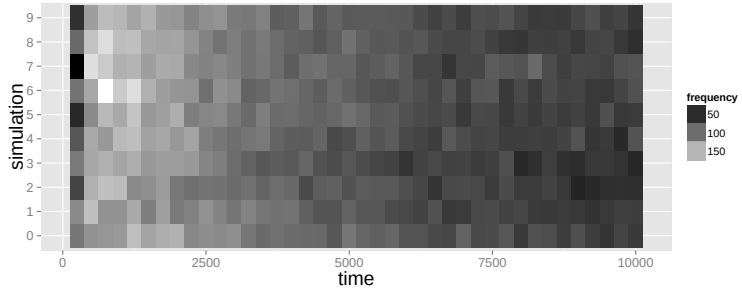
Reinforcement of the most-concrete used constructions (the **mcuc**s) is something that takes place continuously. Recall that we observed that for sim-



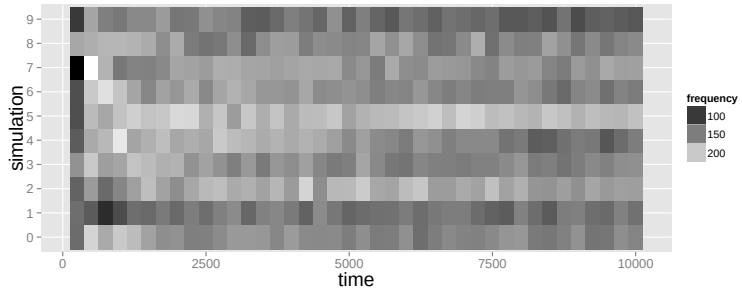
(a) Syntagmatization.



(b) Abstraction.



(c) Update of a grammatical most-concrete construction.



(d) Reinforcement of a grammatical most-concrete used construction.

Figure 6.2: Frequency of learning mechanisms involved in the acquisition of lexical constructions over time.

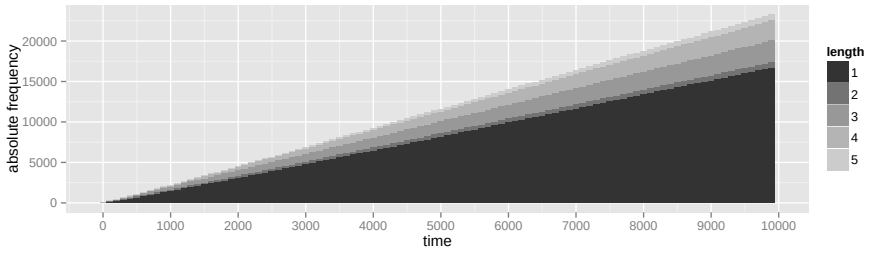
ulation 5, SPL performed fewer updates of lexical **mcuc**s than for the other simulations. Interestingly, we find the reverse for the grammatical **mcuc**s, namely that there are more reinforcements of grammatical **mcuc**s in simulation 5 than for the other simulations. What happens in simulation 5, is that the model relies more on lexically specific grammatical constructions than in the other simulations. This is merely an effect of the order of the first hundreds of input items, but it raises the interesting possibility that the order and temporal distribution of the input items may affect the kinds of representational categories that are used and reinforced, thus allowing for individual variation despite the same mechanisms and sensitivities (or parameters) of the mechanisms. Crucially, in all simulations adequate behavioral performance is achieved: the model is able to identify the target situation, analyze the full utterance and understand to what parts of the identified situation the elements of the utterance refer. This finding supports the recent insight that it may be the case that, despite behavioral near-identity in everyday behavior, language users' internal grammars may vary (e.g., Dąbrowska 2012). However, they do so through a different route: whereas in the case of Dąbrowska's results, the differences between individuals are likely a product of differences in the quantity and quality of experience, in the case of this modeling experiment, the quantity and (to a large extent) the quality of the linguistic experience are the same between simulations. This raises the interesting suggestion that the order of input items may affect the representations learned by a language user.

6.2 The representational potential

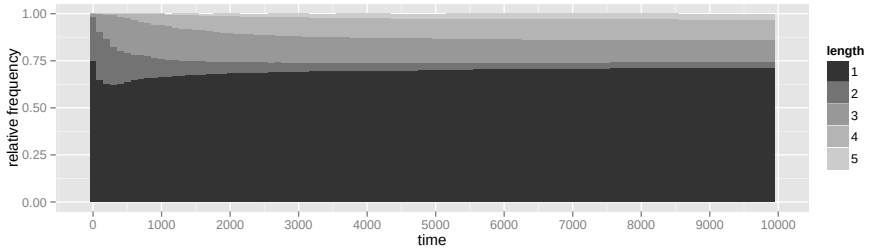
In section 5.3.4 we looked at how often constructions of various length and abstraction are used by SPL in comprehending utterances. At that point, I remarked that there may be a difference between the constructions used by the model and the potential the model has. The internal state of the model can be compared with the behavior of the model (in comprehension, for instance). This way, we can arrive at an understanding how distant the model's constructional potential is from the behavior it produces. Such insights are important, given that in many usage-based corpus studies a strong what-you-see-is-what-you-get perspective is taken, assuming that the behavior as given in a corpus does not provide evidence for a more abstract representational system, but it may be that the typically highly limited behavior of children is produced by a richer (i.e., more abstract) representational system in which, for instance, the abstract patterns never surface in behavior because they are always preempted by more concrete, slightly worse-fitting but better-entrenched ones.

6.2.1 Length of the acquired constructions

Before we look at specific cases, let us inspect some general properties of the model. Figure 6.3a illustrates how the constructional knowledge is monoto-



(a) Unsmoothed absolute frequency of constructions of various length over time.



(b) Unsmoothed relative frequency of constructions of various length over time.

Figure 6.3: Unsmoothed frequency of constructions of various length over time.

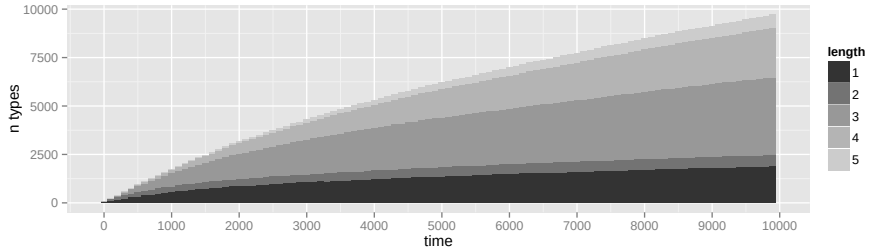


Figure 6.4: Number of unique construction types of various length over time.

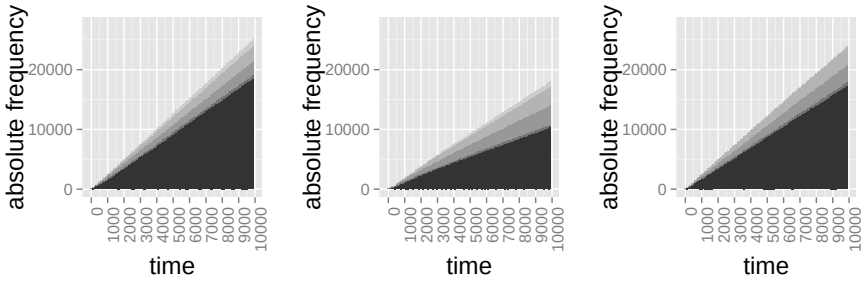
nously increasing. The height of the bars reflect the total amount of reinforcements the constructions of various length have received as **mcuc**s. As the frequency is unsmoothed, constructions with a count of zero (i.e., those that have been acquired through bootstrapping, cross-situational learning, syntagmatization, paradigmatization, or as an **mcc**, but that have never been reinforced), do not count towards the global frequency.

The figure that depicts the same data, but then as proportions of the total grammatical knowledge (figure 6.3b) shows a slightly different picture. In it, we can see that in the early stages, most of the counts are divided over lexical constructions and length-2 grammatical constructions. One by one, length-3, length-4 and length-5 constructions enter the constructicon and become reinforced.

Many constructions may be present as representations without ever having received any reinforcements, and as such figures 6.3a and 6.3b give a slightly distorted image. After all, in actual use, the counts of these constructions are smoothed, so that their probability is non-zero. An alternative way of conceiving of the absolute and relative strength of the various representations is by looking at the number of unique construction types at each time. Figure 6.4 gives this information.

One striking aspect of the number of types, when compared to the absolute or relative frequencies, is that there are many constructions of length 3 and greater that have not been reinforced. This is an effect of the blind application of the paradigmatization operation, where any and all abstractions are added to the grammar. It also points to the clear way in which SPL instantiates the idea of immanence: any overlap between any two patterns is part of the model's potential for analyzing novel utterances.

Looking at the variation between simulations, next, we can first observe that there is a difference in the absolute number of reinforcements divided over the grammar. Whereas in simulation 3, the total number of reinforcements after $10,000$ input items is around $23,000$, the number of reinforcements



(a) Simulation 3 (b) Simulation 5. (c) Simulation 6.

Figure 6.5: Absolute frequencies of constructions over various lengths over time for three simulations. Legend is the same as in figure 6.3.

in simulation 5 lies around $18,000$. Interestingly, simulation 5 also performs slightly worse on the identification of the target situation as well as the situation coverage (cf. figures 5.1 and 5.3). As we will see in the next chapter, the model also behaves slightly differently in simulation 5 than in the other simulations. Nonetheless, even in simulation 5, SPL is a relatively successful communicative agent, correctly identifying over 70% of the target situations.

Furthermore, an interesting pattern in the comparison between the simulations surfaces. Whereas simulations 3 and 5 (and all others) have constructions of length-5, in simulation 6, reinforced constructions of that length are not in the representational system most of the time, with a few emerging only at the end. The model is in this simulation nonetheless as successful as in the other simulations. What happens in simulation 5, is that various length-4 constructions of the type given in example (37) are acquired. These constructions become reinforced both by sentences of the type *you put ball in basket* as well as cases of *you put ball there*, where the model analyzes *there* as referring to the LOCATION. At some point in simulation 6, constructions of the type in (37) have been reinforced to such an extent that the final word may even be known (e.g., [SPEAKER / *me*]), but this word cannot be concatenated with the construction, as it refers to the LOCATION as well. Combining them with concatenation would constitute a violation of the isomorphy principle, and is therefore excluded. Alternative analyses (e.g., using a length-3 construction and concatenating that with the well-known word) turn out less likely than the ones on the basis of the types of constructions exemplified in (37).

(37) [[ENTITY] [PUT / *put*] [ENTITY] [LOCATION]]

(38) [[ENTITY] [PUT / *put*] [ENTITY] [LOCATION-ROLE] [LOCATION]]

At around 9000 input items, SPL has started to acquire the caused-motion

construction as exemplified in (38). Upon encountering further instances of sentences like *you put ball in basket*, the model is now able to parse them with a length-5 construction, and it is likely that this construction will continue being reinforced over time. All in all, nothing is lost, but the model is simply a late learner with respect to the length-5 constructions.

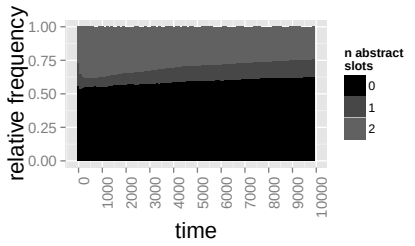
However, not all utterances that *can* be covered with length-5 constructions *are* covered with such constructions. The model has mistakenly taken up the prepositional dative construction (the construction behind sentences such as *he gave the book to Mary*) as a length-4 construction:

(39) [[ANIMATE$_i$] [GIVE / *give*] [OBJECT] [ANIMATE$_j$ / *to*]] |
GIVE(GIVER(ANIMATE$_i$),GIVEN-OBJECT(OBJECT),
RECIPIENT(ANIMATE$_j$))

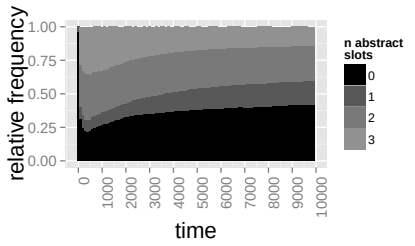
This construction involves (correctly) an animate entity in the giver-role, the verb, and a given object. It has mistakenly learned *to* to refer to the recipient entity, but only in the context of this constructions: SPL is able to analyze sentences such as *you go to school* or *you take ball to table* with a construction that involves *to* as a marker of direction. As with the earlier erroneous caused-motion construction in (37), the fact that *to* refers to the entity filling the recipient role blocks the pattern from being concatenated with a noun or pronoun following it, even if that noun or pronoun is well known.

To ascertain that this is not an effect that can be overcome with more data, I let simulation 6 continue processing input items after it was done. Even after $20,000$ input items, the model still analyzes prepositional datives with constructions such as (39). We can take this to mean that the model got stuck in a local optimum. This means that it has acquired a construction (i.e., the one in example (39)), that allows it to identify the situation correctly in most cases, but that does not cover all of the utterance and the situation. Of course, real language-learning children would never find themselves 'stuck' in such a situation: the functional relatedness of *to* in the prepositional dative to that in several motion constructions (underlying such utterances as *you go to school* and *you take ball to table*), and the fact that the application of the construction of (39) always leaves one word of the utterance unanalyzed, even if that word may be well known, should, at some point, convince the learner that the construction in (39) is not a conventional pattern of the language.

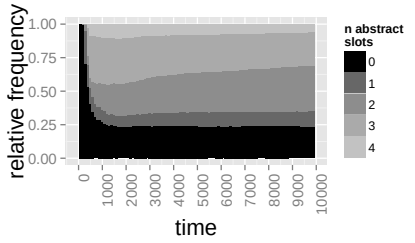
This points to a point of weakness of the model: it is not able to overcome these local optima. This constitutes a kind of brittleness that we would like a model to be able to overcome. To my mind, a crucial change in the model might be to make the 'penalty' for ignoring words proportional to how well these words are known. If the learner encounters *you give it to me*, and knows that *me* refers to the speaker, it should penalize analyses in which *me* is taken to be noise more severely than analyses in which *to* is taken to be noise (as that word likely has little reinforcements outside of the constructions in which it constitutes a fixed element).



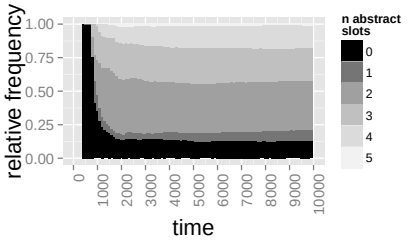
(a) Abstraction among length-2 constructions.



(b) Abstraction among length-3 constructions.



(c) Abstraction among length-5 constructions.



(d) Abstraction among length-5 constructions.

Figure 6.6: Relative frequencies of the various degrees of abstraction, per length, over time.

6.2.2 Abstraction in the representational potential

In section 5.3.4, I discussed the use of constructions of various length and degrees of abstraction over time. Being the constructions that are used in finding the best analysis, these constructions are also the ones that are reinforced over time. We can interpret the effects on the abstraction of the constructions of various lengths by looking at how much reinforcement each of these levels of abstraction has accrued over time. Figure 6.6 shows the normalized frequencies of each level of abstraction, per length, over time.

What the various figures show, is that the potential for generalization is quickly obtained by the model (somewhat later in the length-4 and length-5 constructions than the length-2 and length-3 constructions). After having found this potential, more and more more concrete patterns are learned that take up increasingly much of the relative frequency. That is: the potential for having a fitting representation for each situation becomes greater over time. Note that, unlike SPL's use of unanalyzed lexical chunks, it's increasing use of analyzed but phonologically specific constructions is in line with the findings reported by McCauley & Christiansen (2014*a*).

What are the semi-abstract longer constructions that are well-entrenched? If we look at simulation 5, and inspect the most-frequently used length-5 constructions, the following five constructions constitute the top-5:

(40) [[SPEAKER / *you*] [PUT / *put*] [PLURAL-PERSON / *them*] [CONTAINMENT-ROLE / *in*] [ENTITY]] ($count = 94$)

(41) [[PERSON$_i$] [GIVE / *give*] [THING / *it*] [GIVER-ROLE / *to*] [PERSON$_j$]] ($count = 93$)

(42) [[PERSON] [GIVE / *give*] [THING / *it*] [GIVER-ROLE / *to*] [WOMAN]] ($count = 80$)

(43) [[PERSON$_i$] [GIVE / *give*] [THING] [GIVER-ROLE / *to*] [PERSON$_j$]] ($count = 53$)

(44) [[SPEAKER / *you*] [PUT / *put*] [THING] [CONTAINMENT-ROLE / *in*] [ENTITY]] ($count = 30$)

We see both the caused-motion pattern and the prepositional dative in various degrees of abstraction among the five most-frequently used constructions. These semi-open constructions function as composite multi-word units in comprehension: multi-word units because they capture frequently occurring lexical patterns, composite because each of the parts of the construction specifies a certain role in the more global meaning. As such, these patterns are distinct from true 'chunks', that are internally not analyzed.

In all of the five most-frequently used length-5 constructions, the verb is fixed. In fact, in none of the length-5 constructions in this simulation, a pattern in which a generalization over caused-motion constructions and prepositional dative construction is made. This is a direct effect of the fact that the model has erroneously acquired *to* in the prepositional dative to refer to the GIVER, or AGENT, role. Because of this, the model cannot form an abstraction over the meaning representations of the two constructions.

As we can glean from figure 5.9h in chapter 5, there are some simulations in which the abstraction over caused-motion constructions and prepositional datives is made, judging by the small, but non-zero amount of length-5 constructions with 5 abstract slots. In simulation 2, for instance, the model has acquired a construction, given in (45), that only has a fixed subject, but no other lexically specified roles. The reason this abstraction could be made, is that in simulation 2 the model did correctly acquire the meaning of *to* in the prepositional datives as referring to the RECIPIENT role (unlike in simulation 7, where it is analyzed as denoting the PATIENT or GIVEN-THING role, and simulation 5, where it is analyzed as marking the RECIPIENT referent).

(45) [[HEARER / *you*] [CAUSE] [OBJECT] [ROLE] [ENTITY]] |
CAUSE(CAUSER(HEARER),AFFECTED(OBJECT),ROLE(ENTITY))

This construction, however, is used only between 1500 and 4900 input items, and only to analyze prepositional datives. What happens here, is that

the model extracts the construction in (45), and finds it to be part of the most likely analyses of prepositional datives with *give* as the verb. These analyses then are added to the grammar as maximally-concrete constructions (**mccs**), and after 4900 input items, SPL has acquired a range of these more concrete patterns to the extent that the abstraction in (45) is no longer needed.

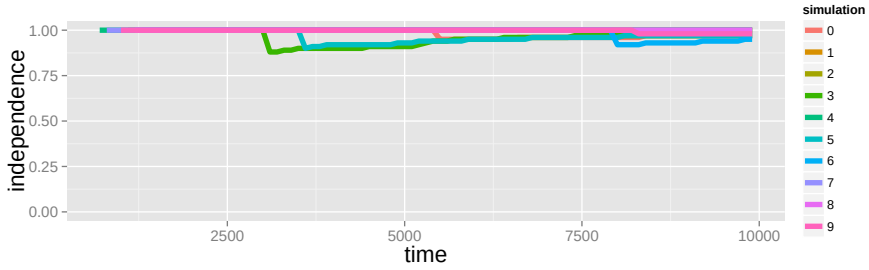
Returning to *to*, it seems that the various simulations differ in how they analyze *to* in the prepositional dative. Out of ten, only four assign the correct RECIPIENT role to the word, whereas in five cases the RECIPIENT referent is taken to be the meaning of *to*, and in one case, as we have seen, the PATIENT role. Of course, more than 40% of children acquiring English get the meaning of *to* correct (although it may be a preposition for which semantic errors could be expected).

Several aspects prevent the model from being like a child for this phenomenon. First, the input is more scarce in types of verbs and prepositions than a child receives. If various verbs and prepositions are heard, the chance of acquiring the right meaning of *to*, which contrasts with other prepositions in that position, becomes greater. Suppose various verbs are heard in length-5 constructions. An abstraction of the type (45) is then quickly acquired. Even if the meaning of *to* is erroneously acquired, a construction with an open verb slot, like the one in (45), may 'overrule' the more specific, but erroneous pattern with *give* and *to* lexically specified. As we have seen in chapter 5 that more abstract constructions may 'overrule' (which we can take to be the antonym of 'pre-empt') the use of more concrete ones if the abstract ones are well-entrenched and the lexical units filling the slots are well-entrenched as well.

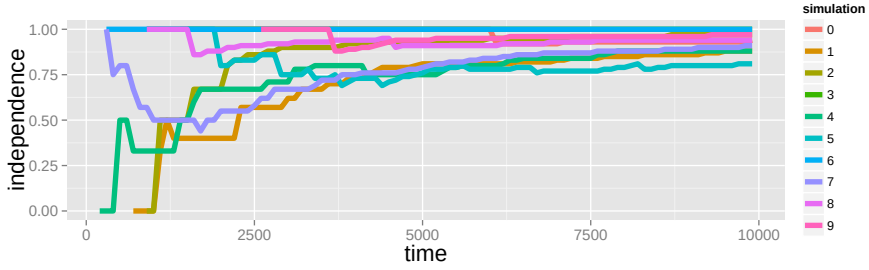
6.3 The independence of morphemes

We saw that *to* was acquired, with the correct or incorrect meaning, as part of a larger construction. In none of the simulations, a lexical construction of the type [RECIPIENT-ROLE / *to*] is well-entrenched. Of course, the representation is there, but it never gets reinforced, because it is always the larger construction in which *to* is used that is reinforced. This raises an interesting issue, namely whether the model can tell us something about the independence of the smallest units. This is an issue that touches on both (what are traditionally called) morphology and syntax: a unit is bound if it can be only used in combination with other units, whereas it is free if it can be used independently. Can the model give us any insight in the degree of independence of the various words?

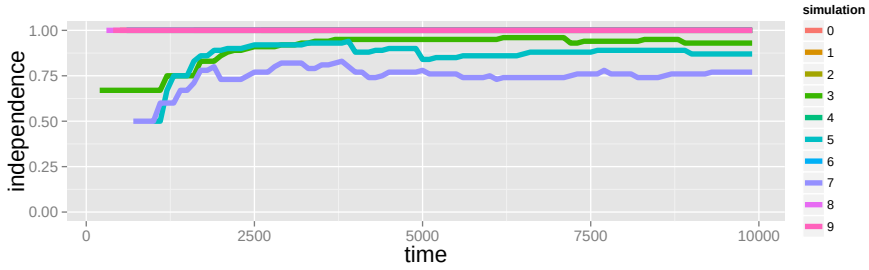
The issue of independence finds its theoretical relevance for the usage-based approach in the programmatic article on language acquisition by Langacker (2009), who argues that the independence of a unit (construction) depends on the variety of contexts it occurs in, in interplay with the frequency of the unit itself, both inside and outside of particular constructional contexts. Certain words, such as determiners, may never obtain strong independent



(a) Independence of the word *cereal* over time.



(b) Independence of the word *animal* over time.



(c) Independence of the word *aunt* over time.

Figure 6.7: Independence of various entity words over time.

status, whereas others, occurring over a variety of contexts, do get reinforced as independent units. I would like to add one aspect to Langacker's conceptual analysis, namely that, besides the token frequency and the dispersion of a word over various constructions, also the type frequency of the constructional slot (i.e., the amount of types of units filling it), plays a role in establishing the degree of independence of a unit. This last point is taken from Bybee's (2006) analysis of constructional productivity. In this section, I will show how all these effects can be seen in the model when we look at the strength of the representation of lexical constructions as opposed to grammatical constructions containing those lexical constructions.

In the following paragraphs, we look at five groups of words, corresponding roughly to nouns, adjectives, pronouns, verbs, and prepositions/spatial adverbs. We can expect the degree of independence to vary between them, as they have different quantitative values for the three properties mentioned above.

As a simple measure to operationalize the independence of a word form w, I take the relative frequency of lexical constructions out of all constructions in which a word form w is lexically specified (cf. equation 6.1, where Γ_w is defined as the subset of the constructicon Γ consisting of all constructions in which w occurs as the phonological specification of a constituent). This tells us how often the word form w is analyzed with a lexical construction. The more frequently this happens, the more we can claim that w and its meaning are free units. We call this value the **independence** score, ranging between 0 and 1.

$$\mathbf{independence}(w) = \frac{\sum\limits_{c \in \Gamma_w \wedge c = \mathbf{lexical}} c.count}{\sum\limits_{c' \in \Gamma_w} c'.count} \tag{6.1}$$

6.3.1 Entity words

Words referring to entities, typically called 'nouns', can be expected to be among the most independent words. After all, they occur as the arguments of multiple action words ('verbs') in the input generation procedure, and many other entity words fit these slots as well, making it likely that the optimal analysis involves the lexical construction involving the entity word and a grammatical construction with an open slot where the entity word is fit in. Figure 6.7 shows, for three entity words, that this is indeed the case. After $10,000$ input items, constructions involving the phonological strings *cereal*, *animal*, and *aunt* are mostly lexical.

When focusing on the developmental path, we see *cereal* being used mainly in lexical constructions from the onset of the simulations, whereas *aunt*, and especially *animal* start out as often being part of a grammatical construction early on, and gradually being used more as an independent word, and hence

receiving more reinforcement as a lexical construction. The string *animal* occurs in all but a few cases as the theme argument of a caused-motion construction (in utterances such as *you put animal on table*. Because of the restricted variability, the model does not have to use the lexical construction [ANIMAL / *animal*] in any other context, and in the context of caused-motion sentences, the model has a semi-open construction of the type in example (46). Whereas the semi-open constructions in example (46) will receive reinforcement over time, the lexical construction will not. This pattern of pre-emption, however, is gradually overturned, as constructions such as (47) also receive much reinforcement. In this construction, the theme argument is open, and because many different theme arguments are encountered, this kind of construction receives much reinforcement. Over time, the best analysis is increasingly likely to involve the construction with an open theme-argument slot (example (47)) and the independent lexical construction [ANIMAL / *animal*], and more reinforcement is given to the lexical unit. The same happens, to a weaker degree, for *aunt*.

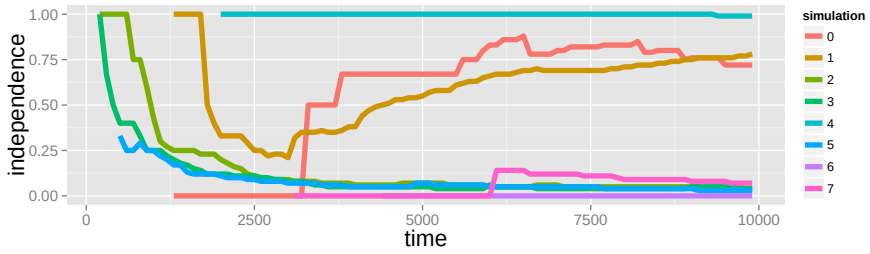
(46) [[HEARER / *you*] [PUT / *put*] [ANIMAL / *animal*] [SURFACE-ROLE / *on*] [ENTITY]]

(47) [[HEARER / *you*] [PUT / *put*] [OBJECT] [SURFACE-ROLE / *on*] [ENTITY]]

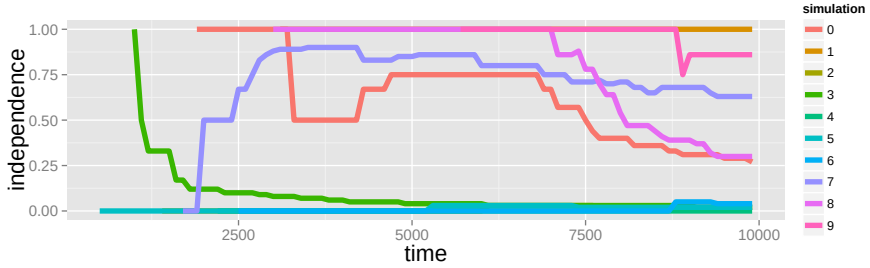
6.3.2 Attribute words

Unlike the entity words, the attribute words ('adjectives') are not used in many different constructions and the verbs that have them as arguments have a fairly restricted set of attribute words in the input generation procedure. Especially in the case of the construction [[ENTITY] [BECOME / *get*] [ATTRIBUTE]], the model moreover often acquires chunks consisting of *get* and the attribute word. Whenever attribute words are acquired, they vary in whether they are learned as a lexical construction or as part of a grammatical construction. For all three words we find the tendency that they become increasingly associated with a construction in which they are lexically specific (decreasing values on the y-axis). However, in some simulations (e.g., simulation 1 for the word *dirty*), the word starts out being used most often in lexical constructions, after which it is used as an element of a grammatical construction, and finally it is dissociated from that construction again. This effect is caused by the interaction of the fact that *dirty* is only used in the [[ENTITY] [BECOME / *get*] [ATTRIBUTE]] construction, but that the ATTRIBUTE slot of that construction is extended with new types over time, leading to increased reinforcement, and hence a greater likelihood of combining the lexical [DIRTY / *dirty*] construction with the construction in which the ATTRIBUTE slot is phonologically open.

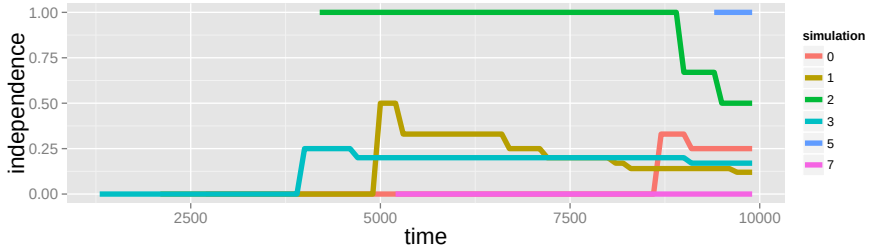
The fact that these attribute words gravitate towards being used in grammatical constructions may be partially due to the fact that there are no copula



(a) Independence of the word *dirty* over time.



(b) Independence of the word *closer* over time.



(c) Independence of the word *pretty* over time.

Figure 6.8: Independence of various attribute words over time.

constructions in the input generation procedure. If there were, the attribute words would also be used in those cases, and their reinforcement as lexical constructions would be greater.

6.3.3 Pronouns

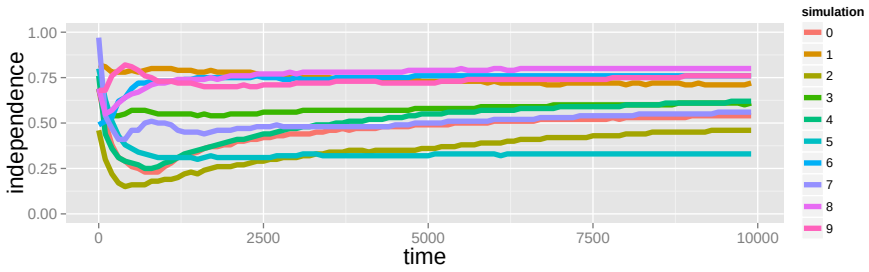
Pronouns constitute an interesting case for the test of independence. Because of their high frequency, they are expected, on a usage-based account, to be part of argument-frame constructions. On the other hand, their varying distribution (especially in a language like English where the pronouns only express two grammatical cases) makes the reinforcement of their independent forms to be expected. Figure 6.9 shows the independence for three pronouns, *you*, *I*, and *we*. As we can see, their degree of independence varies dramatically among them and between simulations. *You* is acquired in all cases both as part of a lexical construction and as part of a grammatical construction (i.e., the learner has both a [HEARER / *you*] construction, and various grammatical constructions in which *you* is used, and, crucially, reinforces all of them regularly (otherwise the relatively stable, horizontal lines of figure 6.9a would not be maintained). The variation ranges between independence scores per simulation of 0.3 and 0.8.

For *I*, the picture is different. Here, we see that there is a significant amount of between-simulation variation, but the stable state of the model in various simulations seems to be more 'polar': either *I* is most strongly represented as an independent construction, or the grammatical constructions in which [SPEAKER / *I*] is a constituent are the primary locus of the knowledge about *I*.

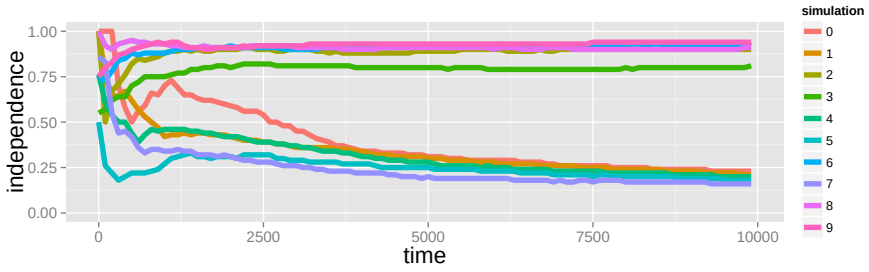
We, finally, is primarily acquired as the phonological constraint on an independent lexical construction. Unlike for the entity and attribute words, it is not easy to find an explanation for this high amount of variation: all three words are used in various constructional slots, and these slots are typically highly productive (i.e., many other items can fit in them). This difficulty of explanation, however, does point to the insight that the degree of independence of a word may be an effect of many interacting factors.

6.3.4 Event words

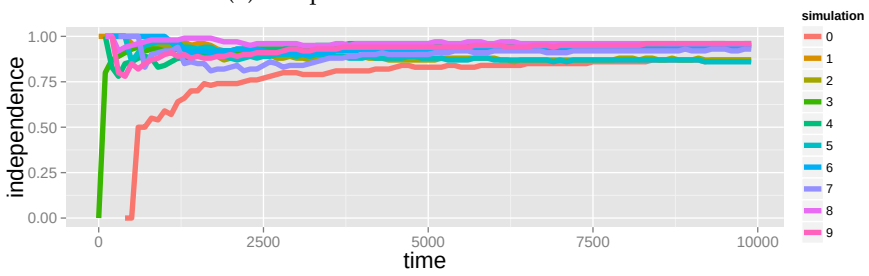
As with the pronouns, the picture of the **independence** of the event words ('verbs') is rather diverse (figure 6.10). The word *eat* is most strongly represented as an independent construction in most of the simulations. This does not come as a surprise if we bring to mind that the non-lexically-specific transitive construction (i.e., the transitive construction with an open EVENT slot) is strongly reinforced. *Put*, on the other hand, is only processed in the context of the caused-motion construction, and this construction allows for no other verbs in it in the input generation procedure. Furthermore, the abstraction over the various caused-motion constructions and the prepositional datives is



(a) Independence of the word *you* over time.

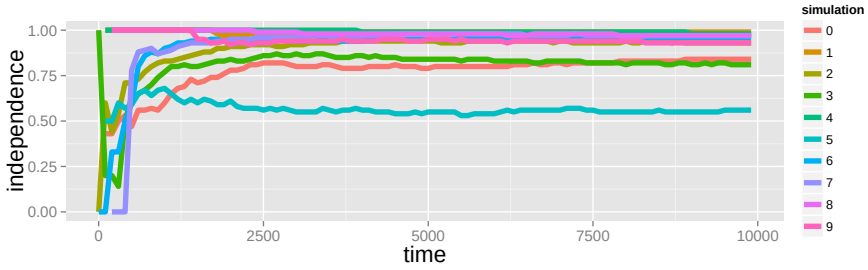


(b) Independence of the word *I* over time.

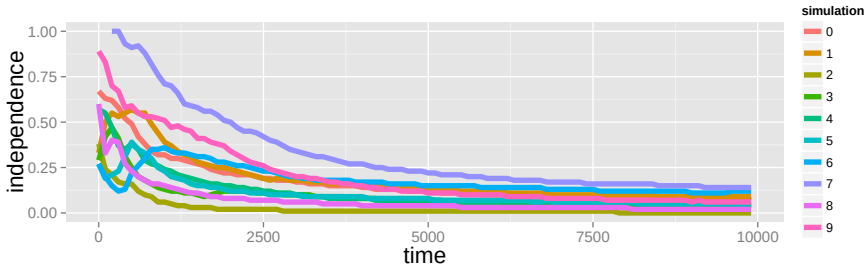


(c) Independence of the word *we* over time.

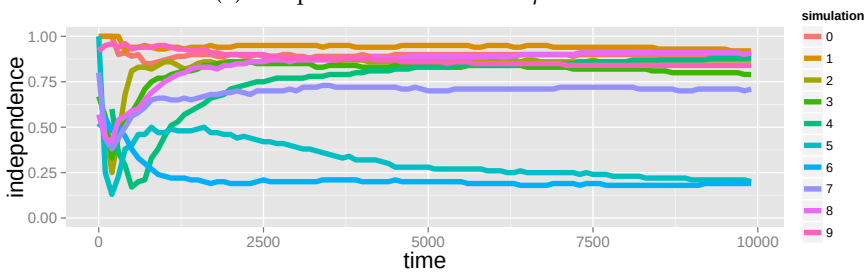
Figure 6.9: Independence of various pronouns over time.



(a) Independence of the word *eat* over time.



(b) Independence of the word *put* over time.



(c) Independence of the word *make* over time.

Figure 6.10: Independence of various event words over time.

rarely made, so that there is neither a non-verb-specific length-5 construction available. *Make*, finally, and like *you*, varies between simulations. In some simulations, the string is primarily used as the sole phonological constraint on a lexically specific construction, whereas in others, *make* is the phonological constraint of the [MAKE / *make*] slot of a larger, grammatical, construction.

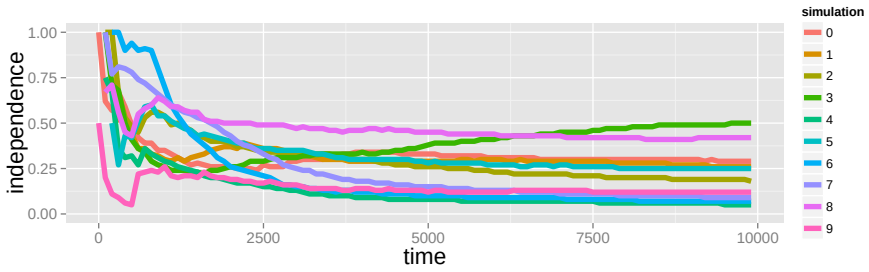
A curious phenomenon for the verbs (and to some extent for the pronouns as well) is that in some simulations, the curve displays a dip in **independence**, after which the value goes up again. The effect that causes this is again the interplay of the productivity of the EVENT slot of various grammatical constructions and the variety in grammatical constructions the verb can occur in. In some simulations, it seems to be the case that the event word starts out as an independent word (it is bootstrapped, or learned by means of cross-situational learning), after which the semi-open constructions in which it is specified amass reinforcement. As the amount of variation in the input data grows, the abstractions over the various semi-open constructions begin accruing reinforcement as well, and at a certain point the most likely analyses involving these event words consist of a grammatical construction with an open EVENT slot, combined with a lexical construction containing the event word. This moment is at the bottom of the dip: afterwards, the **independence** score starts rising again, because the lexical construction gets reinforced, but the grammatical constructions with lexically specified EVENT slots do not.

One could tentatively associate this effect with the idea that children are conservative in the generalization of early verbs (McClure et al. 2006). This very finding has been questioned (Naigles et al. 2009), but it may be that there is a lot of variation between learners, between verbs, and that the periods in which the learner behaves conservatively, or, oppositely, too progressively, may vary as well.

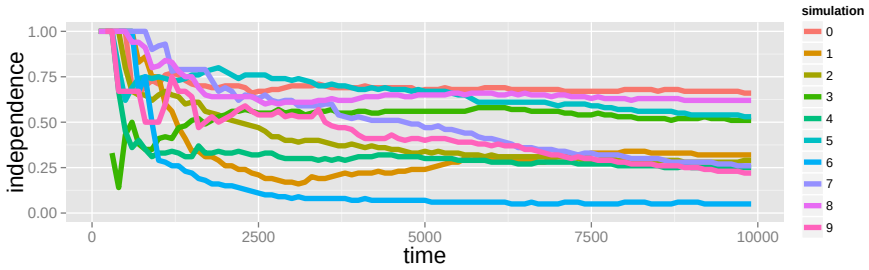
6.3.5 Role-marking words

Role-marking words, traditionally known as prepositions, are expected to be fixed elements of the grammatical constructions they occur in. However, as figure 6.11 shows, this does not seem to be (fully) the case: both *on* and *in* have a relatively strong representation as the phonological constraints on independent lexical constructions. This is due to the fact that these words do occur in multiple constructions, and contrast with other role-marking words (e.g., *to* and *out of*). Nonetheless, in most simulations, the **independence** scores are decreasing over time, meaning that more and more, the words are only used as parts of grammatical constructions.

The difference between *on* and *in* is especially striking. After all, both words occur in exactly the same constructional environments. I believe the difference is due to their varying token frequencies. We can expect, in line with Bybee (2006), that, all other things being equal, words with higher token frequencies (in particular environments) will be more entrenched in those environments. After $10,000$ input items, the average counts of all constructions



(a) Independence of the word *in* over time.



(b) Independence of the word *on* over time.

Figure 6.11: Independence of various role-marking words over time.

containing *in* and *on* is 879 and 350, respectively. This means that *in* is simply more frequent than *on* in the input generation procedure. The effect of this difference is that the more frequent word, *in*, is associated more strongly with the constructional environments it is used in, and hence that lexically specific constructions containing *in* receive more reinforcement than those containing *on*. We see here that SPL not only captures the effect of type frequency on productivity, but also the effect of token frequency on entrenchment.

6.3.6 Comparing the classes

Finally, let us take a more global look. If we group all words for which the model has any representation in at least one of the simulations according to the five-way distinction presented above, and subsequently average over all simulations and all words, per semantic class, we obtain the average **independence** values presented in figure 6.12.

The pronouns and entity words clearly have the strongest independent representations. Attribute words are mostly reinforced as part of the grammatical construction they occur in, and event and role-marking words start out relatively independent, but become more and more associated with par-

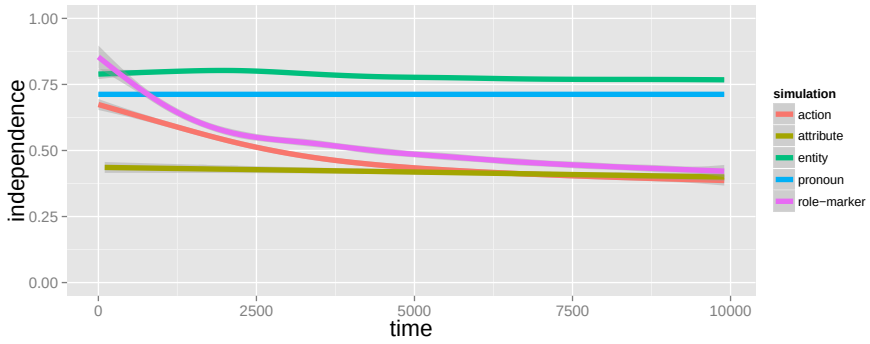


Figure 6.12: Mean independence of the five classes of words over time.

ticular grammatical constructions. This last finding is at odds with the verb-island hypothesis (Tomasello 1992), but I believe it constitutes a viable hypothesis within a usage-based framework in which the starting-small perspective is re-emphasized. Once the learner (i.e., the model) has bootstrapped or cross-situationally learned the meaning of an event word, the build-up phase consists in finding the arguments with which this word can occur. Once those have been found, abstractions are made over that event word and others occurring in similar argument-structure constructions. These abstractions only receive reinforcement if they are extended to novel event words. If that does not happen, the event words will increasingly become associated with the argument-structure constructions, leading to more lexically-specific constructions. This hypothesized developmental pathway also suggests that, in the long run, the representational knowledge of a speaker that is actually used becomes more concrete over time (cf. McCauley & Christiansen 2014*a*), whereas the potential for generalization to novel cases remains stable. Speakers get better at what they do most, without forgetting the tricks for handling novel grammatical situations.

6.3.7 Discussion

The analysis of words in various semantic classes shows us how the degree of independence, as measured by the **independence** score varies on the basis of (1) the type frequencies of the constructional slots of grammatical constructions they occur in, (2) the amount of different constructional environments they occur in, and (3) their token frequencies, much in line with Bybee's (2006) and Langacker's (2009) characterization of notion like productivity and independence. The cases discussed thus provide insight in the subtlety of the

notion of productivity when applied to grammatical, as opposed to morphological constructions. Nonetheless, I believe that through careful analysis and interpretation, we can identify the factors involved in the productivity of the construction. Notably, this is not merely a study of the corpus frequencies of the words: we have to take into account that we are dealing with a learner selectively reinforcing patterns over ontogenetic time. An important difference with Langacker's account is that a short phase of independence may, in SPL, precede a higher degree of dependence. Whether this is an artefact of the model, or an actual developmental phenomenon that becomes visible once we re-evaluate aspects of the starting-small conception of language acquisition as they apply to the usage-based theory, remains to be seen. I find the latter option not inconceivable.

An important insight from the various cases is that the model, in a way, does engage in whole-to-part learning besides part-to-whole learning (D6-3), but in a quantitative way. Qualitatively, after all, the word has been established as a lexical unit. The 'dips' discussed for the event words suggest that after this establishment, the word may go through a phase of being bound to the grammatical constructions it occurs in, after which it re-establishes independence. Part-to-whole and (quantitative) whole-to-part learning thus interact in an interesting way.

A second insight from this analysis, is that SPL displays a tremendous amount of variation between the simulations. The internal representational states of the various 'speakers' differ in the independence of various words. Nonetheless, they all perform very similarly on the comprehension experiments described in the previous section, as well as, as we will see, on the production task. It seems that there is more than one representational way to Rome when grammatical behavior is concerned, a finding in line with the recent experiments of Dąbrowska (2012).

Finally, I believe this exercise supports the recent reanalysis of some old conceptions of language. Most of linguistics, even within the constructivist take on it, is committed to a perspective in which words are the atomic primitives of languages, to be combined with grammar.

The words-as-atomic-primitives perspective has led, within functional linguistics especially, to debates about the nature of word-meaning. It has long been recognized that words can have multiple related senses, a property especially true of function words, such as adpositions, auxiliary verbs and discourse particles. The discussion about word meaning mainly concerns the question whether words have a single, highly abstract meaning (monosemy), the details of which are filled in by the pragmatics, or multiple concrete and related meanings (polysemy). The Croftian perspective, in which the constructions are the primitives (but not necessarily the atoms), allows us to question the central assumption underlying this debate: the word as the locus of meaning. If we take the perspective that constructions are the non-atomic primitives of linguistic knowledge, words (as we normally conceive them as linguists) become secondary, derived realities. A word, by this token, is simply

a phonological and conceptual similarity relation between the parts of various constructions. In some cases, these constructions may coincide with the word (which is what we expect for many nouns, for instance), but in others, the 'word' is the potential emanating from the use of a phonological structure and several similar functional structures across several constructions.

This perspective is much in line with suggestions of Verhagen (2006) and Boogaart (2009). Boogaart argues, for modal verbs, that there may be a third option, resolving the discussion, namely that words have certain meaning within certain constructions. Polysemy becomes, under Boogaart's analysis, a superficial effect of the same word form occurring in multiple constructions. This analysis is supported by the results of the analysis in this section: words that are strongly associated with a particular construction have weak independent representations as lexical constructions. It can be expected that modal verbs, Boogaart's case study, are strongly associated with particular constructional frames (after all, they are fairly restricted in their use across constructions, there is only a small set of them, and they have high token frequencies). If that is the case, it may well be that the lexical representation of a Dutch modal verb like *kunnen* is very weak and that the primary locus of representational strength of the word is in various constructions, each with their own meaning (e.g., deontic vs. epistemological modality).

6.4 The growth of the caused-motion construction

Besides these more quantitatively-oriented explorations of the representational potential of the model, it may also be insightful, especially for those used to doing grammatical analysis within the construction grammar framework, to see how the 'network' of constructions grows over time. Because the grammars after $10,000$ input items contain about the same number of constructions, it is not feasible to look at all of them. Therefore, we focus on a part of this network, namely where it involves events in which motion is expressed, with an external cause for that motion being presented. These are the constructions underlying such utterances as *you put it on table*. As even for this small part of the network, the number of constructions is too vast[1] to represent graphically, I focus on some interesting ones.

Figure 6.13 displays a part of the network after 100 input items. The thickness of the lines is indicative for the counts of the constructions, and constructions in grey have not been reinforced. We can see that the model has learned the action word *put*, and syntagmatized it with two entity words, *you* and *Sarah*, to form two very simple grammatical constructions. Over these constructions, furthermore, the abstraction [[PERSON] [PUT / *put*]] is made,

[1] Note that this way of framing it ('number') presents the constructions as discrete units, which they are in the implementation. As I argued earlier, we can equally well regard these as the potential for generalization the model has – a vast number of constructions in the discrete conception corresponds to a wide potential on a 'immanent perspective'.

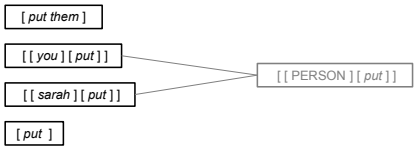


Figure 6.13: Part of the network of caused-motion constructions after 100 input items.

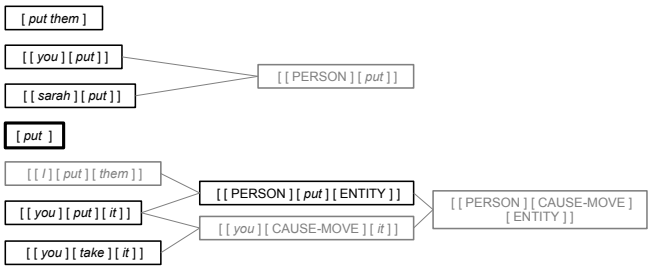


Figure 6.14: Part of the network of caused-motion constructions after 500 input items.

but this construction has not been reinforced yet. We can see that a chunk has been extracted as well, viz. [PUT(AFFECTED(THEM)) / *put them*], and this chunk has been reinforced several times.

Four hundred input items later (figure 6.14), the lexical construction [PUT / *put*] has been further reinforced. Furthermore, several length-3 constructions have been added. The various fully lexically-specific ones give rise to a small network of abstractions, even though many of the fully lexically-specific ones may not have been reinforced (not all constructions are shown here, as there are already dozens of length-3 constructions at this point). The 'old' constructions remain at the same level of reinforcement: as there is now a more useful length-3 construction, the various length-2 constructions no longer lead to optimal analyses.

Moving to the state of the constructicon after 1000 input items, we can seen that length-4 and length-5 constructions now entered the scene. For length-4 constructions, a small, but generalizable network has been built up, including a well-reinforced, highly abstract construction in which only the word *put* is specified. Note here that this abstract construction, [[PERSON] [PUT / *put*] [OBJECT] [LOCATION-ROLE]], has received more reinforcement than its daughter nodes. This is because it is the abstract construction, rather than

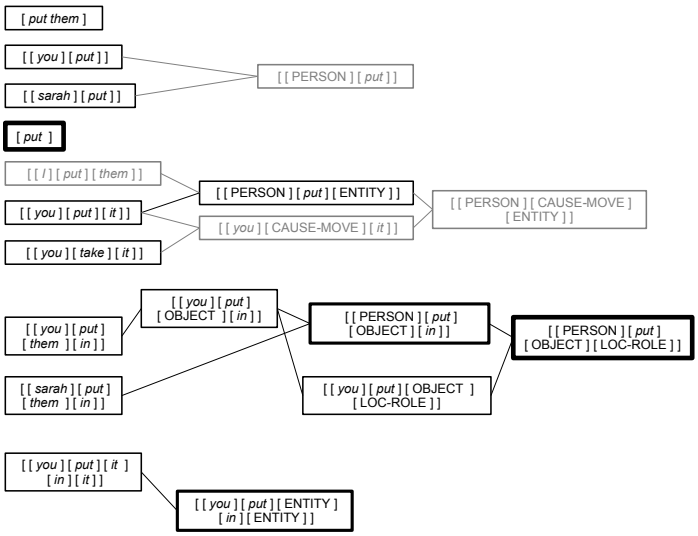


Figure 6.15: Part of the network of caused-motion constructions after 1000 input items.

its daughters that is used in processing the input items. The effect here is akin to the effect of abstract constructions obtaining unit status without the more concrete ones doing so, as described in Langacker. For the length-5 constructions, a relatively lexically-specific construction [[HEARER / *you*] [PUT / *put*] [ENTITY] [CONTAINMENT-ROLE / *in*] [ENTITY]] has been extracted, but the model has not seen any evidence for abstractions beyond this level.

After $10,000$ input items, it has seen evidence for more abstract length-5 constructions, as can be seen in figure 6.16. The network now even contains a construction in which the action word is not specified, abstracting over the constructions with *put* and those with *take* (the other verb occurring in the caused-motion construction in the input generation procedure). This maximally abstract construction has even been reinforced several times, but its more concrete daughter construction involving a phonologically specified ACTION slot (with [PUT / *put*]) has received the most reinforcement, and constitutes the prototype of this network. As we have seen in the previous chapter, it is this construction that sometimes trumps the use of more concrete constructions, because of the many different types of arguments it occurs with.

Interestingly, between 1000 and $10,000$ input items, another length-4 construction emerged as well. The pattern with a lexically specific LOCATION-ROLE, i.e., [LOCATION-ROLE(LOCATION) / *there*], is used frequently enough

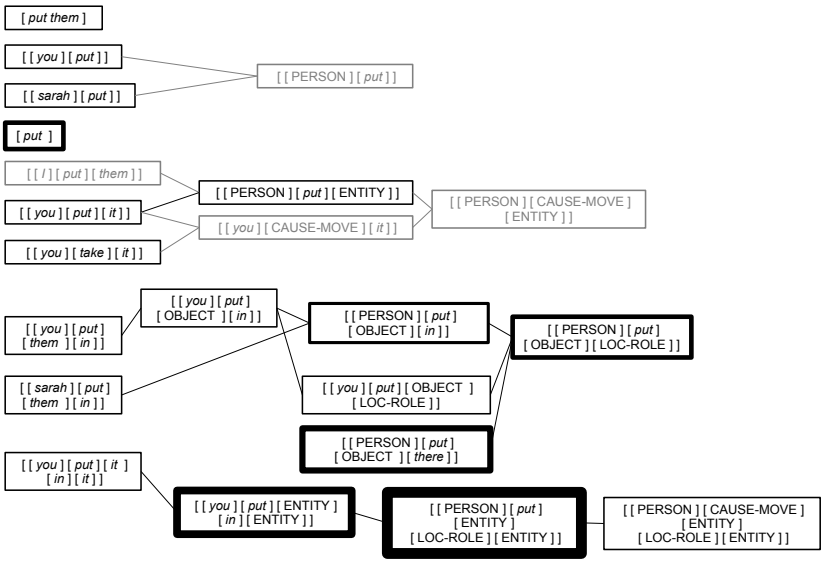


Figure 6.16: Part of the network of caused-motion constructions after $10,000$ input items.

to be strongly reinforced. A more abstract construction exists as well, but that one is less strongly reinforced. Note, finally, that all the older constructions have not received any reinforcement in the meantime. If the model involved some sort of decay function, these constructions would, by now, have withered away.

The visualization of the development of the network illustrates several important aspects of SPL. First of all, the constructions grow in length and abstraction, with each next step of length and abstraction depending on what, at that point, is available to the model. Second, we have seen how abstract units may obtain unit status, or (at least) become strongly reinforced 'prototypes' in the network. Third, the temporal, or dynamic dimension of the model becomes clear: old constructions fall out of use, while novel, and more useful, ones, take over.

6.5 Discussion

In this chapter, I looked at the learning mechanism the model employs and the representations resulting from these. I made several observations that all follow from a rigorous application of usage-based theory to the development

of a model, but that may be at odds with some conceived ways of thinking.

First, we saw how the learning mechanisms are applied in section 6.1). Whereas all mechanisms are available to SPL throughout development (D6-4), the frequency of their application varies over time. Notably, the acquisition of lexical constructions is primarily done by means of bootstrapping rather than by cross-situational learning. Whereas the latter is used to get an initial set of lexical constructions, the former makes for a more reliable way of acquiring word meanings as the abstraction in the representational potential grows. In the learning mechanisms for grammatical constructions, we found that syntagmatization is applied only early on, after which the reinforcement of most-concrete used constructions and the addition of most-concrete constructions become the primary means of learning. If language is, as I suggested earlier, a set of old tools (evolutionarily speaking), used for novel purposes (i.e., language), the tools are of various use at different moments in time. A final point of interest is that paradigmatization, the process whereby novel, more abstract constructions are acquired, takes place in bursts. This observation may bring the usage-based conception in harmony with the finding that not all development is gradual.

Next, I discussed the length and abstraction of the acquired representation (section 6.2). I found that the length of the constructions in the representational potential of the model grows over time, in line with the law of cumulative complexity (D6-1). For abstraction, the first main finding was that for longer constructions, the model goes through a phase of abstraction before building up an ever growing inventory of more concrete constructions. This suggests that adult language users may operate with a large number of semi-open constructions, and that the abstractions are merely kept as a failsafe device in case the more concrete constructions cannot be applied. Nonetheless, an answer to the question when it is better to use a more concrete construction than the combination of a more abstract one and a lexical construction, depends on various quantities, viz. the degrees of reinforcement of the two grammatical constructions as well as the lexical one. As we have seen in the previous chapter, a more abstract construction may lead to a more likely analysis than a more concrete one.

The second main finding concerning abstraction was that length-3 constructions (i.e., transitives) were generally more abstract than constructions of other lengths. I argued that this effect is due to the type frequency on the EVENT slot: as many different words occur in it, the more abstract version of the transitive construction accrues more reinforcement as compared to constructions of other length.

Thirdly, I looked at the degree of independence of lexical constructions (section 6.3). Word forms may be strongly associated with lexical constructions, or with parts of grammatical constructions. In the former case, they constitute independent units, whereas in the latter, they should be considered dependent on the grammatical construction they occur in. We found that, for some items, the independence of word forms varies enormously between

words, semantic word classes and even simulations. The main factors I identified were (1) the type frequency of the slot of the grammatical construction, (2) the number of constructions a word occurs in, and (3) the token frequency of the word. High values for the former two create more independent lexical constructions, whereas high values for the latter create more dependent word forms. The effect of this is that words in semantic classes that combine freely and have relatively few tokens, such as entity words, or nouns, display stronger independent representations than words in semantic classes that occur in a fixed set of environments, where the environments themselves display little variation, and the token frequencies are high, such as event words, or verbs.

An interesting development over time was found for the event words and pronouns. For both cases we saw that, in some simulations, the word was first used mainly as part of a grammatical construction, then as a free unit, and finally as part of a grammatical construction again. In other simulations, we observed only the second and third stage. Especially these latter cases are at odds with the general conception of learning in a usage-based framework, which states that the learner starts with larger units, which are decomposed over time. However, I argued that these findings do follow from the insights of a starting-small approach as applied to usage-based theory.

Despite this finding, the model does engage in some sort of whole-to-parts learning. When a word form is used mainly as a part of grammatical constructions early in development and later on, by developing strong abstract representations, the model comes to understand the word form as an independent entity, it has effectively performed part-to-whole learning, albeit in a quantitative sense. Qualitatively, the word form has already been established as an independent unit, because the blame assignment (i.e., the creation of a symbolic link to the meaning of the word form) has already been done.

The exploration of the development of the network in section 6.4 highlights several important aspects of SPL. First, the law of cumulative complexity is illustrated with the increase of length and abstraction in the network. Second, we saw how more abstract units may receive strong reinforcement despite their more concrete daughter constructions being less strongly reinforced. Finally, the temporal dimension of SPL becomes clear: some constructions may play an important role early on, but become obsolete as longer and more encompassing constructions enter the scene.

In all of the first three sections, I discussed the between-simulation variation. As a mere effect of the input, I found that (1) some learners rely more on lexically-specific grammatical constructions than others, and (2) that the degree of independence of lexical constructions varies between simulations. Despite this variation, all simulations perform similarly in the comprehension experiment, as well as, as we will see, on the production experiment. This suggests that, even without differing sensitivities to the input data, the order and dispersion of the input items may have an effect on the representations that are built up.

What the analyses in this chapter finally show, is that the models potential for linguistic behavior cannot be directly equated with its behavior itself. We could consider this a re-appreciation of the competence-performance distinction, where the competence is, of course, one that is built up through language use. Just as the strict division of competence and performance may be a false reification of an analytic principle in generative approaches to language acquisition, so may the all-too-strong reliance on behavior to understand the representational system in usage-based approaches constitute a case of the reverse. The fact that, in the usage-based framework, the potential and the use of that potential are considered to be one thing ontologically, does not imply that we can make a direct inference from the use of that potential to the potential itself. This point will be further supported by the production experiments presented in the next chapter.

CHAPTER 7

Production experiments

Having seen the behavior of the model (chapter 5) and its inner workings (chapter 6), we now turn to the last topic: the production of language. Desideratum D2 holds that a computational model of language acquisition not only has to account for comprehension, but also for production. In this chapter, we look at the capacity of the model to produce utterances on the basis of a situation, as well as how its behavior develops over time.

7.1 Global development of production

7.1.1 Evaluation

How do we evaluate the accuracy of the produced utterances? Recall that the input generation procedure of Alishahi & Stevenson (2010) generates utterance-situation pairs. In the first production experiment, we generate a test set of 100 utterance-situation pairs at random. Importantly, we are interested in SPL's grammatical behavior, and giving it situations it has seen before would result in simple 'recall' of the analysis of an utterance paired with that situation. For that reason, the 100 utterance-situation pairs in the test set are held out from the input generation procedure for the input items in the simulation as reported in chapter 5.[1]

[1] This works as follows: SPL first generates 100 unique utterance-situation pairs. When generating novel input items for the simulation, it checks for every input item if it can be found in this set of test items. If it is found, a new input item is generated. This procedure is repeated until the new input item is no longer one of the test items.

After every 100 input items, we give the model the situations, but not the utterances of the test set, and ask it to generate the most likely utterance on the basis of the situation (as defined in section 3.7). The resulting utterance U_{gen} can then be compared with the utterance Uwhich was generated by the input generation procedure.

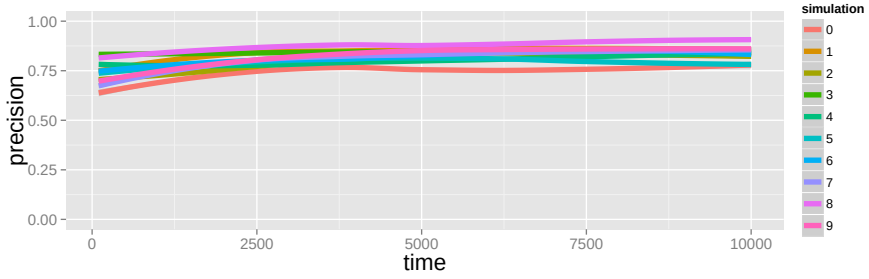
Two aspects of the comparison between U and U_{gen} are central to the evaluation. First, what proportion of U_{gen} is correct? That is: if we generate an utterance, does the model produce words that are part of U. If it produces different words, it has learned erroneous representations. Moreover, we want the model to produce the correct words *in the correct order*. When generating an utterance for the situation in which the father gets the ball, we do not want the model to generate *ball daddy get* or *ball get daddy*. To measure the proportion of words of U_{gen} being produced in the right order, we take the length of the maximal, potentially discontiguous substring shared between U and U_{gen} and divide it by the length of U_{gen}. We call this measure **precision** and errors on the precision correspond to errors of comission: SPL produces things that it should not produce. To give an example of the precision calculation: if the model produced *daddy give ball*, and U consists of the string *daddy give me ball*, the precision is $\frac{3}{3} = 1$ as all words in U_{gen} are found in U in the right order (but *me* is missing from U_{gen}). If the model, however, produced *give ball daddy*, the maximally shared discontiguous substring is *give ball*, and the **precision** is $\frac{2}{3} \approx 0.67$.

The complementary measure of evaluation is the **recall**. This measure captures what proportion of U is present in U_{gen}, again in the correct order. To calculate the recall, we again take the length of the maximal, potentially discontiguous substring shared between U and U_{gen}, but now divide it by the length of U rather than that of U_{gen}. **Recall** measures the amount of errors of omission: the score is penalized for words that are left out of U_{gen} but are present in U. For $U =$ *daddy give me ball* and $U_{\text{gen}} =$ *daddy give ball*, the maximal shared substring is *daddy give ball*, and the recall would be $\frac{3}{4} = 0.75$. For $U_{\text{gen}} =$ *give ball daddy*, the **recall** would be **recall** $= \frac{2}{4} = 0.5$.

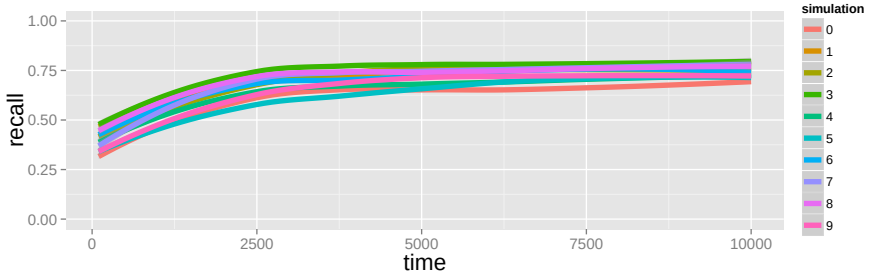
Two other numbers are of interest. Besides **precision** and **recall**, it is insightful to see how long the productions in U_{gen} are, compared to the actual utterance U. This figure tells us whether produced utterances become longer over developmental time regardless of their correctness. **Relative length** is calculated by dividing the length of U_{gen} by the length of U. Finally, as with the comprehension experiment, we would like to know what parts of the situation the model expresses with its production. To this end, we calculate the **situation coverage** for the best analysis (see equation (5.2)).

7.1.2 Results

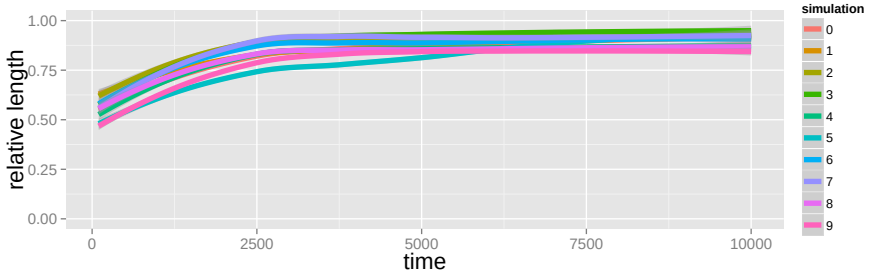
Figure 7.1 gives the values over time for the four measures. After $10,000$ input items, the **precision** scores for the ten simulations range between 0.75 and 0.9 (0.84 on average), whereas the **recall** scores at the end of the simulation range



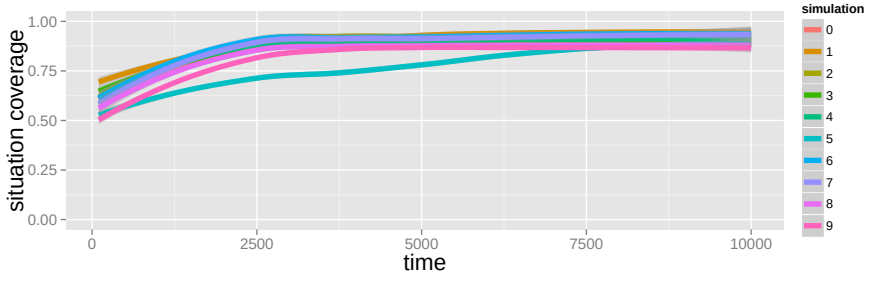
(a) Precision.



(b) Recall.



(c) Relative length.



(d) Situation coverage.

Figure 7.1: Evaluation of production results.

between 0.7 and 0.8 (0.75 on average). Alhough far from perfect, the model does produce utterances that are relatively close to what an adult (i.c., the actual utterances from input generation procedure) would have said. I will analyze the errors the model makes in section 7.2.

Comparing **precision** and **recall**, it is remarkable to see how the **precision** starts out high, goes through a small dip in some simulations, and then goes up again, whereas **recall** starts low (0.2 to 0.35), and rises over the first 3000 input items to its final values (with the exception of simulation 5, the lowest line in **recall**, **relative length**, and **situation coverage**, to which we will return below). This observation is in line with the general observation that children's errors of commission are few, whereas they frequently make errors of omission.

Turning to the **relative length** now (figure 7.1c), we can see that the length of the produced utterances when compared to the actual utterances approaches its ceiling level after 4000 input items for most simulations (and some 6000 for simulation 5). The relative length at the end of the simulation is between 0.85 and 0.95, meaning that the utterances produced by the model are on average 0.85 to 0.95 times as long as the actual utterances.

Finally, the **situation coverage** of the model converges to an almost full expressivity relatively quickly, reaching values of around 0.90 and higher after some 2500 input items, again with simulation 5 lagging behind and reaching full expressivity after some 7000 input items.

Concluding: the model is relatively well able to produce utterances for novel situations, expressing the largest part of the situation. The **precision** and **recall** scores never reach, or even approach the full 1.00. We turn to the sources of this effect in the next section.

7.1.3 An example

Suppose you want to express a state of affairs in which an entity who can be categorized as a father enables the change of possession of a piece of gum. An adult speaker could say something like *father gives me gum* in such a case. Aften 900 input items, the model does so as well (example (52)), producing the utterance *father give me gum*. When we look at the best analysis leading to this utterance, we can see that SPL uses a maximally abstract ditransitive construction, combined with lexical constructions for every word.

The road to this production is one of a gradual build-up of the full utterance when looking at the utterances produced. As we can see in example (48) through (51), the model subsequently produces *give*, *me give*, and *father give me* before arriving at *father give me gum*. This is in line with the observation that over time more and more arguments of a verb are expressed (Tomasello 1992). When looking at the best analyses leading to these generated utterances, we find an interesting pattern. First, only a lexical construction leading to the word *give* is used, after which the model employs a maximally abstract intransitive construction to combine *me* with *give*. The intransitive construction

only specifies that the first constituent fulfills a participant role in the event, and so the recipient *me* fits that slot. Combining this constellation with the lexical *give*-construction, the model arrives at a richer semantic interpretation: the role filled by *me* is now specified to be the RECIPIENT. It is interesting that the model makes a word-order error because of this: it takes the 'pre-verbal' slot to allow any semantic argument, and as such the model overextends a construction. Note that this kind of generation is allowed by the model in subsequent generation turns as well, but from $t = 300$ onwards, there are already analyses that are more likely and have a better coverage of the meaning. Overgeneralization does not go away, it is just outcompeted.

Looking at the generations at $t = 300$ and $t = 500$ (examples (50) and (51)), we can see that the model generates the string *father give me*, but does so with different means. In the former case, SPL uses a fully lexicalized construction, whereas in the latter, a verb-island construction [[PERSON] [GIVE / *give*] [ENTITY]] is used, combined with lexical constructions for *father* and *me*. This means that by 500 input items, the slightly more abstract construction has become reinforced to a greater extent than the fully lexicalized construction.

(48) [GIVE(GIVER,GIVEN,RECIPIENT) / *give*]

(49) [[PERSON]→[SPEAKER / *me*] [EVENT]→[GIVE(GIVER,GIVEN,RECIPIENT) / *give*] |
GIVE(GIVER,GIVEN,RECIPIENT(SPEAKER))

(50) [[FATHER / *father*] [GIVE / *give*] [SPEAKER / *me*]] |
GIVE(GIVER(FATHER),GIVEN,BENEFICIARY(SPEAKER))

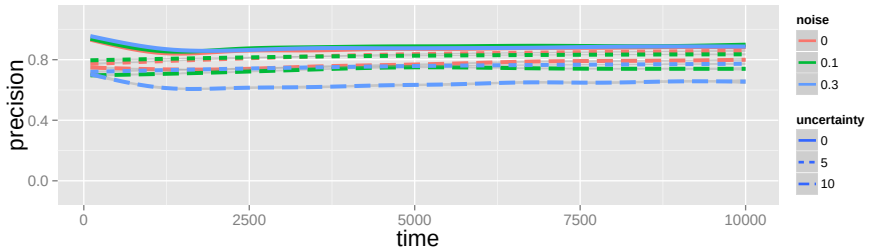
(51) [[PERSON]→[FATHER / *father*] [GIVE / *give*] [ENTITY]→[SPEAKER / *me*]] |
GIVE(GIVER(FATHER),AFFECTED-ROLE(SPEAKER))

(52) [[PERSON]→[FATHER / *father*] [CAUSE]→[GIVE / *give*] [OBJECT]→[SPEAKER / me] [ENTITY]→[GUM / *gum*]] |
GIVE(GIVER(FATHER),GIVEN(GUM),RECIPIENT(SPEAKER)

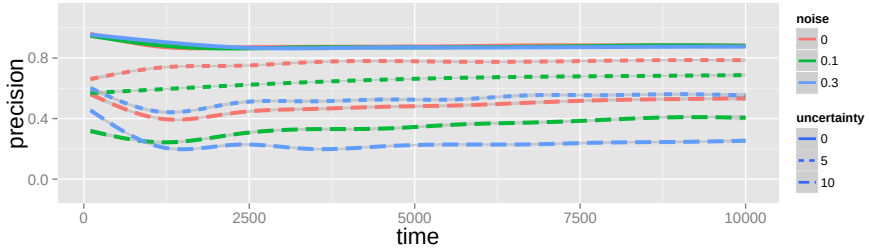
7.1.4 Robustness to uncertainty and noise

As in section 5.2.4, we can look at the model's performance given various settings for P_{noise}, $uncertainty$ and P_{reset}. If we make the conditions harder, does the model perform much worse on the generation task, or does its performance degrade gracefully? Again, we take values $noise = \{0.0, 0.1, 0.3\}$, $uncertainty = \{0, 5, 10\}$, $P_{\text{reset}} = \{0.05, 1\}$, and we run three simulations for every setting.

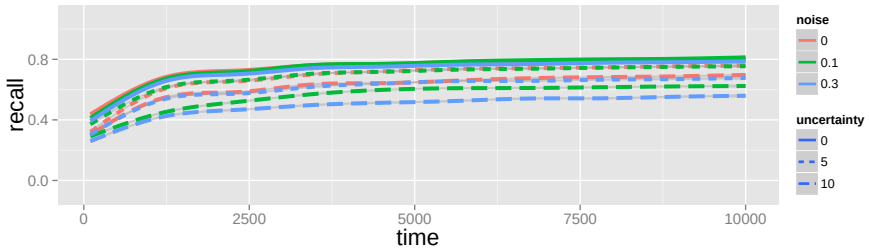
Looking at **precision** first, we can see that with $P_{\text{reset}} = 0.05$, increasing the levels of noise and uncertainty does not have a strong effect on the model's performance (figure 7.2a). Under the hardest condition, $uncertainty = 10$, $noise = 0.3$, the **precision** score after $10,000$ input items is 0.68, meaning that more than two thirds of the words the model produces are still correct. For



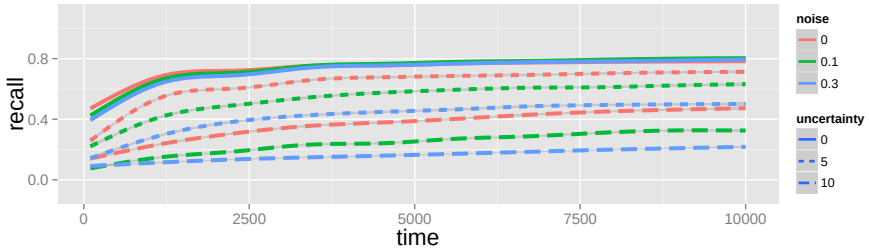
(a) Precision scores for nine unique noise and uncertainty settings over time given $P_{\text{reset}} = 0.05$.



(b) Precision scores for nine unique noise and uncertainty settings over time given $P_{\text{reset}} = 1$.



(c) Recall scores for nine unique noise and uncertainty settings over time given $P_{\text{reset}} = 0.05$.



(d) Recall scores for nine unique noise and uncertainty settings over time given $P_{\text{reset}} = 1$.

Figure 7.2: Precision and recall scores given various parameter settings.

all other settings, the PRECISION scores range between 0.75 and 0.85. That is: the model reasonably picks up the right representations from the noisy and uncertain sets of situations.

Setting P_{reset} to 1 makes SPL less robust to uncertainty and noise, as we have seen for the **identification** scores in section 5.2.4. Under the hardest conditions, the productions of the model are now only correct for some 20%, meaning that SPL has acquired many erroneous representations that, moreover, have been reinforced over time (figure 7.2b).

The variation between the various P_{noise} and $uncertainty$ settings is somewhat greater for the recall, meaning that, despite primarily producing utterances that are correct, they become less complete if the model faces higher levels of noise and uncertainty (figure 7.2c). The latter parameters seems to have a stronger effect than the former here: the lowest two scores after $10,000$ input items are for the setting $uncertainty = 10$. Again, the effect of setting the P_{reset} to 1 is dramatic (figure 7.2d): SPL acquires many erroneous representations, especially in the situation sets with high uncertainty, and subsequently fails to produce the correct target utterances.

Summarizing these findings, we could say that SPL is a robust learner given relatively high levels of noise and uncertainty (at least: higher levels than reported in other modeling experiments), but the chain of situations has to be 'coherent': if situations do not resemble each other, the robustness of the model fades away. However, I believe the uncertainty faced by actual learners is rather like the one given $P_{\text{reset}} = 0.05$ than $P_{\text{reset}} = 1$, as I argued in chapter 4. Asking the model to perform well given $P_{\text{reset}} = 1$ presents the experiential world of the child as an incoherent, haphazard sequence of events which we know it is not.

7.2 Error analysis

More interesting than the cases that are learned correctly are the ones where the model fails. Studying them provides us with more insight in the aspects of the model that cause this behavior, and thus constitute stepping stones towards even more comprehensive models. When the model omits words that are part of the actual utterance U or when it adds words that are not part of U, what are the kinds of errors the model makes? Some errors are more interesting than others: if the model simply has not acquired a lexical construction yet, and is hence unable to produce a certain word, it is simply a matter of time before the model encounters the word and (hopefully) acquires it. If we find errors in the grammatical patterns, for instance in the omission of arguments or displaying a different order, there is a more interesting story to be told. We will have a look at several cases in this section.

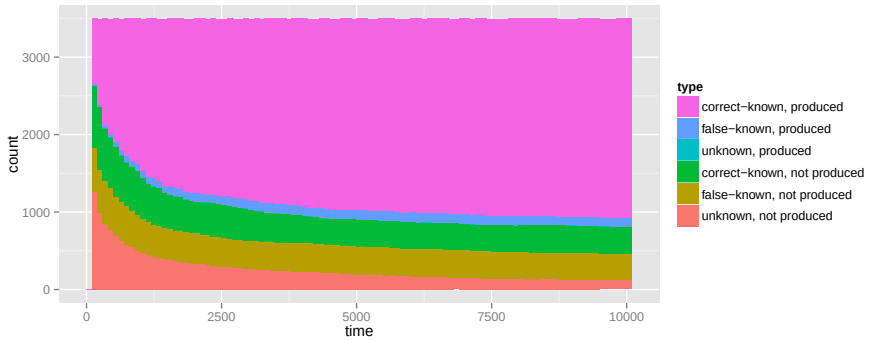


Figure 7.3: Lexical production over time, summed over 10 simulations.

7.2.1 Lexical errors

When a word in U is not produced in the generated utterance U_{gen}, there are several possibilities. First of all, SPL may simply not know the word, in which case it will either use another word or not express the meaning. Second, it may also be that the model has acquired the word. In that case, the acquisition may be correct (the word is learned with the right meaning) or incorrect (the word is learned with the wrong meaning). In the case of unknown words, and incorrectly acquired words, the story is relatively simple: SPL does not have the adequate representation, and hence does not produce the correct word. The case of correctly known, but not produced words is more interesting. Why would SPL not produce known and correct words when they are called for?

We can divide up the words of the various actual target utterances (the Us) in several groups: there are words that are produced, and words that are not produced (i.e., words that are or are not in U_{gen}. Both produced and non-produced words can be known as a word or not known as a word (i.e., at time t, there is a construction in Γ^t that has exactly one constituent with that word form as its phonological constraint). The known words can be further subdivided into correctly learned words and incorrectly learned ones (according to the input generation procedure). We count a word as correctly learned if there is at least one construction in Γ^t that has the meaning assigned to it in the input generation procedure as its meaning.

The counts of the six groups over time are given in figure 7.3. After $10,000$ input items, about 5 out of 7 words in all Us are correctly learned and produced in the generated utterance U_{gen}s. Initially, many words are simply not known and hence not produced, but the count of this group drops rapidly (recall that most word types have been seen after 1500 input items, cf. figure 5.7). Several words are simply acquired with the wrong meaning, and are therefore

mostly not produced.

Outcompeted words

An interesting group are the cases in which the meaning has been acquired correctly, but that are still not produced (the green bin in figure 7.3). There are several reasons why cases like these exist. In the generation in example (53) below, the model tries to express the event in which a boy plays with a pen. It involves a semi-open construction involving the chunk *play with* and two open constituents for the participants. The two participant roles, however, are both filled with the word *worse* instead of *boy* and *pen*.

(53) U_{gen}: *worse play with worse*
U: *boy play with pen*

The expression of BOY with *worse* is easily explained: there are no lexical constructions involving *boy*. There are, however, several (erroneous) lexical constructions involving *worse* as the phonological specification. The abstraction over these, (i.e., the lowest common denominator) is the maximally abstract semantic feature ENTITY. The model now faces two choices: either not expressing BOY at all, or expressing it with the highly abstract [ENTITY / *worse*] construction. Because the model has acquired many grammatical constructions with the agent-role expressed as the first constituent and few without it, it will prefer the generation in which it can use a 'transitive'-like pattern combined with *worse* over a verb-patient construction without any word.

Roughly the same happens for the patient role. SPL has, at this point, acquired a [PEN / *pen*] construction, with a count of 1. Why does the model not combine this construction with the third constituent of the grammatical construction? The reason here is that *worse* has also been bootstrapped once (and erroneously) as meaning PEN. The count, however, is 0. There are, nonetheless, many lexical constructions with *worse* as their phonological form, and a meaning like ENTITY or ARTEFACT. These abstractions, as well as the [PEN / *worse*] 'gang up' (being equivalent derivations) and outweigh the [PEN / *pen*] construction.

This type of error can be considered to be a flaw in the design of the model, but resolving it on principled grounds is harder, and as such poses more of a theoretical challenge than an implementational issue. The problem is in the abstraction over lexical constructions: if a word is erroneously acquired and reinforced, and correctly learned and reinforced (e.g., [FATHER / *father*] and [PEN / *father*]), the lowest common denominator between the two is abstracted (e.g., [OBJECT / *father*]). We know this is unrealistic, but constraining the paradigmatization learning operation to apply in a more limited way would have to apply across the board. This is what, for instance Chang (2008) does in her model: the two constructions over which an abstraction is made, have to be sufficiently similar according to some metric. It is likely that this would work, but to what extent can it be justified as a cognitive operation? If

abstraction is immanent, any shared structure is – in principle – an immanent abstraction. Restricting the amount of abstraction seems to me to impose an unprincipled constraint on immanence. If, however, such restrictions can be motivated, there is nothing barring us from implementing such a feature in a model.

Grammatical restrictions

The second case is constituted by words that are correctly learned, but not produced because there is no grammatical construction facilitating them or because the grammatical construction is less likely than another grammatical construction that does not facilitate that word.

In the former case, there simply is no grammatical construction to accomodate the production of the word. We can see an example of that in the best analysis of a situation in which Sarah puts a finger in her mouth (U = *Sarah put finger in mouth*), represented in (54) below.

(54) [[PERSON]→[SARAH / *sarah*] [EVENT]→[PUT / *put*] [OBJECT]→ [MOUTH / *mouth*]]

What happens in this case is that the best grammatical construction, the one that captures most of the situation and is most likely, is a transitive, and *Sarah* and *mouth* are expressed as the two arguments of that transitive. Nonetheless, at this point, the model does have two lexical constructions [IN / *in*] and [FINGER / *finger*], but it does not have the means to produce them under a single grammatical constellation.

These cases are interesting, because they are in line with the claim that errors of omission in early stages of language production do not depend on the vocabulary size, but that it is really a matter of grammar (Berk & Lillo-Martin 2012). Although Berk & Lillo-Martin (2012) argue for a different conception of grammar, their point can be easily transferred to a constructivist framework: all lexical constructions for producing a caused-motion pattern are present, it is just the caused-motion construction that is missing. This kind of analysis also provides a hint at a constructivist solution to Berk & Lillo-Martin's (2012) puzzle: if one-and-a-half-year-olds and six-year-olds that otherwise developed normally, go through the same phase of argument omission, the reason must be a grammatical one. A usage-based explanation of this phenomenon that, crucially, involves syntagmatization would be that the more abstact and longer grammatical patterns have not been 'constructed' yet.

The second case, where the grammatical pattern is available, but outcompeted, happens for an item in simulation 9 where the target utterance is *she play with toy* and the target situation PLAY(PLAYER(FEMALE-PERSON),TOOL-ROLE(TOY)). In the interval between 700 and 1400 input items, the model produces *she play with toy*, correctly, as the generated utterance, and does so on the basis of the analysis in example (55). This analysis involves a highly abstract transitive construction being combined with the chunk *play with* and the two

participants. However, after 1400 input items, the model erroneously learns that *with* refers to the entity filling the TOOL-ROLE of the PLAY event, and acquires a construction given in example (56), in which the word *with* is taken to refer to the TOY. On the basis of an analysis combining this construction with the lexical construction [FEMALE-PERSON / *she*], the model produces the incomplete utterance *she play with*. This case is illustrative of a lexical error that is made despite the word being known: the model considers another construction 'better' for these purposes, despite even having a construction to express more aspects of the meaning.

(55) [[PERSON]→[FEMALE-PERSON / *she*] [EVENT]→[PLAY(PLAYER,TOOL-ROLE) / *play with*] [OBJECT]→[TOY / *toy*]]

(56) [[PERSON]→[FEMALE-PERSON / *she*] [PLAY / *play*] [OBJECT / *with*]]

7.2.2 Argument structure errors

Argument structure errors come in various sorts in the generations of the model. A first one is the case of a caused-motion event with a causer, and a object undergoing a falling action. The target utterance for such a sentence would be an intransitive utterance involving the undergoing object and the word *fall*, for instance *ball fall*. However, the meaning does steer towards a transitive expression. Note that the model does not have any alternative expressions available for expressing the causation of a falling event (e.g., the suppletive verb *drop* in *I dropped the ball* or a periphrastic causative like *I made the ball fall*. What happens in the model is that, after producing the sole word *fall* for a number of test moments, the model starts producing *fall* in the transitive frame, basically combining a maximally open transitive construction with the words for the causer and the undergoing object, and *fall*. This could be seen as a case of overgeneralization: the model wants to be expressive, but has no better means to do so than to use a transitive. However, the model never 'recovers' from this overgeneralization, as it has, as I mentioned, no alternative ways of expressing it and it has the built-in desire to trade off maximal expressivity with likelihood of the constructions.

The same pattern is found with caused motion events that involve, in the actual utterances, verbs like *go* and *come*, but are produced in a transitive frame (*you go it* for 'you made it move'). Here, again, there is no competing construction and the model relies on a highly general transitive construction despite never having heard *go* or *come* used in this frame. Here, however, it seems that the model does have a competing construction, viz. the caused-motion construction. However, both situations with *come* and *go* have a semantic feature COME and GO associated with them that clashes with the feature PUT associated with *put*, and hence the model is not able to use *put*. We will return to overgeneralizations in section 7.3.

7.2.3 Argument omission

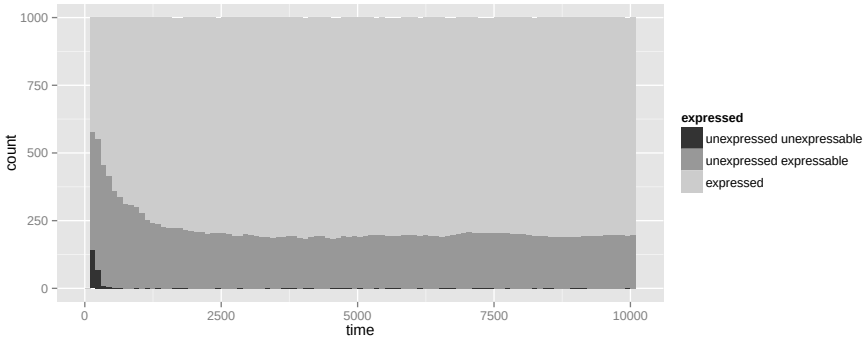
Recall that two of the explananda for a usage-based theory, E1 and E2, held that a computational model of language acquisition has to account for the increasing length of utterances, as well as explain why subject omission is more prevalent than the omission of other arguments. The data in figure 7.4 already suggests that the first explanandum is met: utterances become longer over time. The question, however, is whether this is actually an effect of more arguments being expressed or whether it is done for some other reason.

The three graphs in figure 7.4 show, over time, how often certain arguments are expressed. I grouped the arguments into three bins: 'first' arguments, such as agents and intransitive subjects, 'second' arguments, which are always undergoers, and 'third' arguments, encompassing recipients and locations. What we find, first, is that, in line with explanandum E1, more arguments are expressed over time.

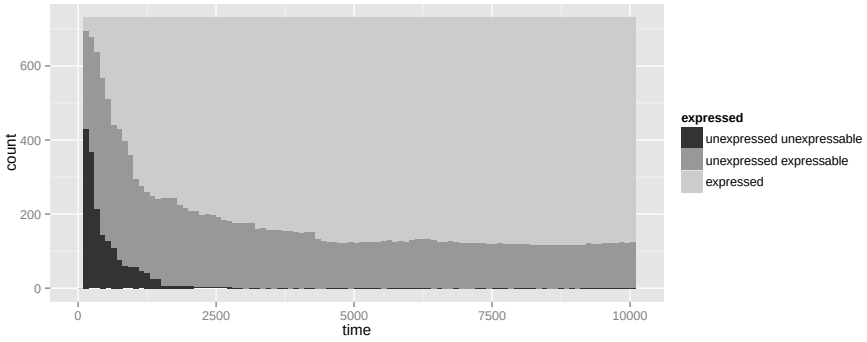
This is not simply a factor of the growing vocabulary, as one may argue. The red bars in figure 7.4 display the 'unexpressed unexpressables', i.e., those meanings for which there is no construction in the grammar at that moment expressing them, whereas the green bars represent the 'unexpressed expressables' (i.e., those meanings that can be, but are not expressed). The former case is 'excusable': SPL simply has no means of expressing that concept. The latter group, the unexpressed expressables, is more interesting: here, SPL has a means of expressing that meaning, but cannot do so, because the grammatical constructions do not allow for it. As we can see for all three groups of arguments, the number of unexpressed unexpressables diminishes rapidly, whereas the number of unexpressed expressables diminishes more gradually. A main factor, according to this analysis, in early argument omission, is the availability of grammatical constructions for expressing arguments, in line with the findings of Berk & Lillo-Martin (2012), who excluded vocabulary size as a factor for the two-word phase (as discussed in chapter 2).

Turning to explanandum E2, the prevalence of subject omission, we can see that the model fares less well. For the first few hundreds of iterations, almost all second and third arguments are omitted, but only about half of the first arguments (i.e., subjects). One explanation for this could be that the model has no notion of information structure. As I discussed in chapter 2, Graf et al. (2015) found that children are more likely to omit old information. As subjects typically contain old information (Du Bois 1987), it is more likely that they are omitted. However, this explanation does not say how this is done representationally: are the subject arguments present in the grammatical representation and omitted, or are they simply not part of the linguistic construct? This is an issue that has been discussed extensively in various generative approaches (see chapter 2), but for which there is no clear answer yet within the usage-based framework.

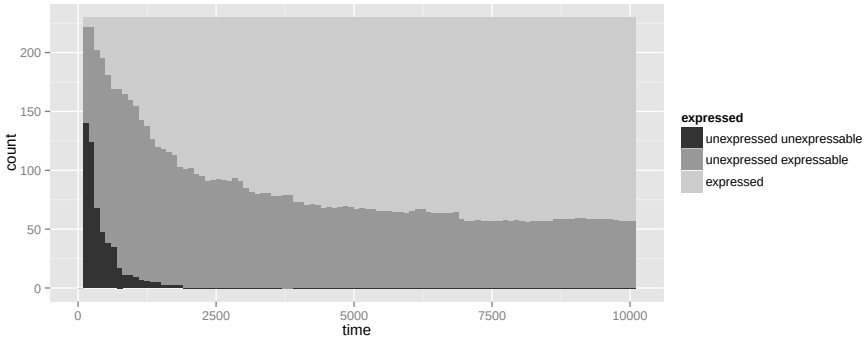
A second explanation would be that learners have a right-edge bias in processing, in line with, for instance the MOSAIC model. If this is the case, it is



(a) The expression of 'first' arguments over time.



(b) The expression of 'second' arguments over time.



(c) The expression of 'third' arguments over time.

Figure 7.4: The expression of arguments over time, summed over 10 simulations.

likely that the model will start picking up [[EVENT] [ENTITY]] patterns earlier than [[ENTITY] [EVENT]] patterns, and that, hence, first arguments will be omitted more frequently. Similarly, one could imagine adding information structure to the comprehension: the more an argument is expected, the less salient it is, the less likely it is to be incorporated in the grammatical analysis, and hence the less likely it is to syntagmatize patterns involving the expected argument.

7.3 Overgeneralization

7.3.1 Motivation and Experimental set-up

In the previous section, we have seen that the model overgeneralizes the transitive construction to the verb *fall*, and does not overcome this overgeneralization. The reason it does not learn that *fall* is not to be used in a transitive frame, as adult speakers of English know, is that it has no alternative that prevents (or: pre-empts) this production. The existence of alternatives opens up the question under what conditions pre-emption takes place. The studies on overgeneralization by Ambridge and colleagues, as discussed in 2.4.3 present several factors involved in this process.

Statistical pre-emption, first, takes place when a competing form to the overgeneralization has been frequently encountered. Second, children seem to understand that if a verb is more frequently seen in a fixed set of constructions, their expectation of the occurrence of that verb in other argument-structure constructions becomes lower (entrenchment). Third, children are increasingly sensitive to the narrow verb classes for the various constructions: verbs of sound emission cannot be transitivized without a periphrastic causative (*I made him scream* vs. **I screamed him*) whereas verbs of manner of motion can be transitivized both with and without a periphrastic causative (*I rolled it* and *I made it roll*). Finally, Ambridge and colleagues suggest that the frequency of the various argument-structure constructions involved may have an effect as well: the more frequently an argument-structure construction occurs, irrespective of its relative frequency to the competing construction, the more entrenched it will be, and hence the more accessible.

All of these effects seem to follow from Alishahi & Stevenson's (2008) model. Can we, similarly, find them in the parsing approach taken with SPL? To investigate this, we adapt the input generation procedure slightly. The verb *fall* is part of the input generation procedure. It is produced either with a moving object as the first argument, in which case the situational event meaning is {EVENT,MOVE,FALL} and the underlying construction is [[ENTITY] [FALL / *fall*]] | FALL(MOVER(ENTITY)). The second construction in which *fall* has a moved object as the first argument, in which case the event meaning in the situation is {EVENT,CAUSE,MOVE,FALL} and the construction underlying it is [[ENTITY] [CAUSE-FALL / *fall*]] | CAUSE-FALL(MOVED(ENTITY)).

Recall that SPL overgeneralizes the transitive construction to generate cases like *you fall it* for the last type. Recall furthermore that the model does not overcome this overgeneralization for a lack of an alternative. For this experiment, I added another verb, *drop*, which has the same meaning as the second type of *fall* (viz. {EVENT,CAUSE,MOVE,FALL}), but also occurs in the transitive construction (i.e., [[ENTITY$_i$] [CAUSE-FALL / *drop*] [ENTITY$_j$]] | CAUSE-FALL(CAUSER(ENTITY$_i$),MOVER(ENTITY$_j$))). Will this alternative pre-empt the use of *fall* in the transitive construction?

Using this additional verb, we can manipulate the frequencies of the two verbs and the constructions they occur in to see if effects of entrenchment and pre-emption are found. The three frequencies we manipulate are:

1. The frequency of *fall* in the non-causative meaning. We expect that the higher the frequency of *fall* in this construction is, the more it will be entrenched, and the less likely it is that it will be extended to other argument frames. Within SPL, this expectation arises through the effect of independence, as discussed in chapter 6: the more a word will be seen in a particular construction, the more it will be associated with that construction, and the less autonomous it will be. We set the frequencies of *fall* in the non-causative frame to 750 (its original frequency) or 75.

2. The frequency of *fall* given a causative meaning. We expect that the higher the frequency of *fall* given this meaning, the more entrenched it is in the intransitive construction (but with a causative meaning), and the less frequent the overgeneralization will be.

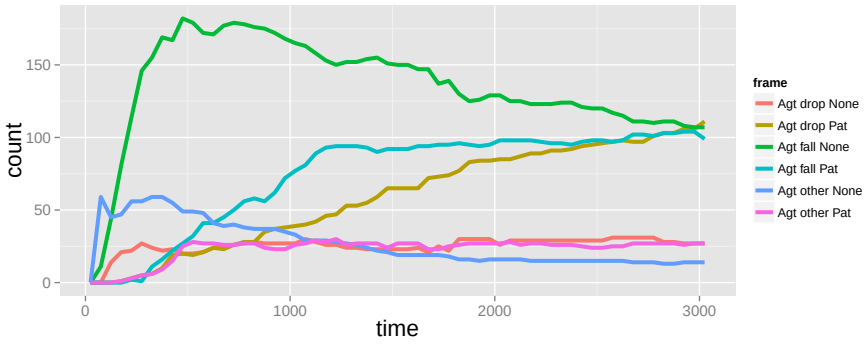
3. The frequency of *drop*. If *drop* is rare, its reinforcement will be weaker, and the chance of overgeneralizations will be higher. We set the frequencies of *drop* to 10 or 100.

I test these hypotheses by running 10 simulations of $3,000$ input items for each of the 8 unique combinations of frequency settings. Every 50 input items, the model will receive 10 frames with a CAUSE-FALL event and two participants and is asked to generate utterances for each of them. I scored the produced generations as follows: the CAUSER-role can be expressed (Agt) or left unexpressed (None). The CAUSE-FALL event can be expressed with *drop*, *fall*, or another word, or left unexpressed. The MOVER-role, finally, can be expressed (Pat) or left unexpressed (None).

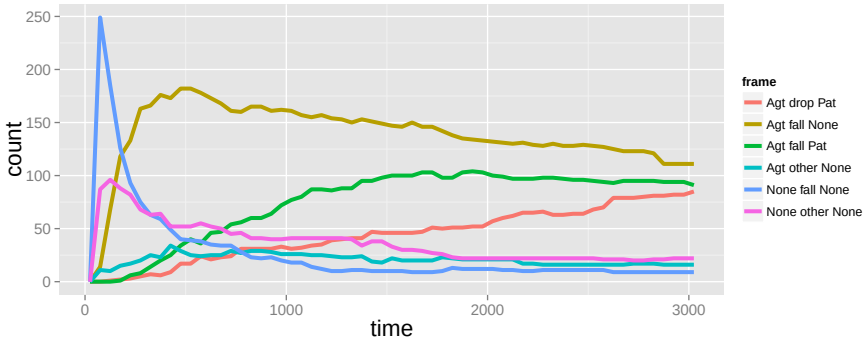
7.3.2 Results

Frequency of non-causative *fall*

Figure 7.5 displays the various types of generations for a CAUSE-FALL situation with two participants. For both frequency settings of non-causative *fall*, we can see that the majority of generations involves a causer and a



(a) Produced frames for caused-falling events over time, given a frequency of *fall* with non-causative meaning = 750.



(b) Produced frames for caused-falling events over time, given a frequency of *fall* with non-causative meaning = 1500.

Figure 7.5: Produced frames for caused-falling events over time with the frequency of *fall* with non-causative meaning as a dependent variable.

word expressing the event, represented as 'Agt fall' (e.g., *You fall* for CAUSE-FALL(CAUSER(HEARER),MOVED(BALL))). Over time, however, both the transitive use of *fall* ('Agt fall Pat', e.g., *you fall ball*) and the transitive use of *drop* ('Agt drop Pat', e.g., *you drop ball*) are on the rise.

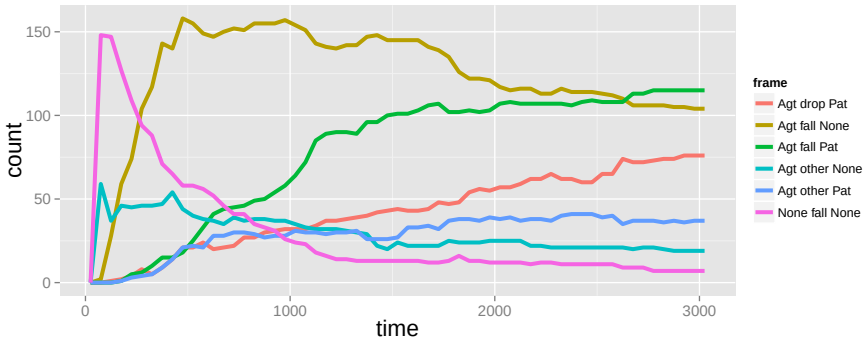
The difference between the two settings is that with the frequency of non-causative *fall* set to 750, the use of transitive *drop* surpasses that of both agentive-intransitive *fall* and transitive *fall* around $3,000$ input items, whereas it remains lower than these two erroneous production types if we set the frequency of non-causative *fall* to 1500. This means that we do not find an entrenchment effect of *fall*: given the pure entrenchment hypothesis, we would expect that the more *fall* is seen in one grammatical construction, the less likely it would be to use it in other grammatical construction. Of course, this is an effect of the fact that SPL only positively reinforces verb-construction associations (with most-concrete constructions), but does not inhibit the non-occurrence of non-observed grammatical constructions.

Frequency of causative *fall*

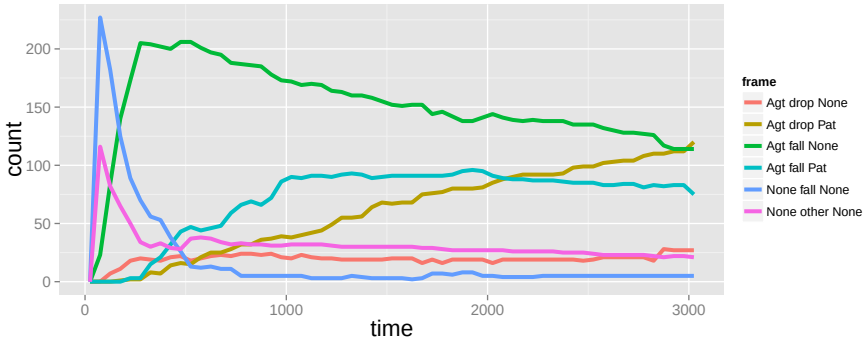
Interestingly, for the frequency of *fall* in the causative, but intransitive, frame, we do see an entrenchment effect (figure 7.6). Again, we find 'Agt drop None' being used most frequently early on, with 'Agt fall Pat' and 'Agt drop Pat' rising in frequency over time. However, here the higher frequency of *fall* given a causative meaning makes the correct use of *drop* being acquired faster, with its use surpassing that of 'Agt fall' and 'Agt fall Pat' at aroun 3000 input items (figure 7.6b). This means that we do find an entrenchment effect here: the more the model has seen *fall* with a causative meaning in the intransitive construction only, the quicker it arrives at productions with *drop* as a suppletive verb. SPL behaves like this because the representation of the constructions underlying the intransitive-*fall* utterances with a causative meaning are more reinforced, thus allowing the model to produce 'Pat fall' constructions. These constructions are, however, never produced, because the model finds the 'Agt fall Pat' and 'Agt drop Pat' patterns more expressive, and the 'Agt fall' pattern better entrenched and hence more likely.

Frequency of *drop*

The frequency setting for *drop* has the greatest effect. If we set the frequency of *drop* to 10, as in figure 7.7a, the verb is simply not reinforced enough to compete with *fall*, which has a frequency summed over both frames it occurs in of 775. Setting the frequency of *drop* to 100 remedies this and makes *drop* a viable competitor to the use of *fall*: the use of *drop* in a transitive construction surpasses both the 'Agt fall' and 'Agt fall Pat' patterns around 1800 input items, despite *drop* still being around 8 times as infrequent as *fall*.

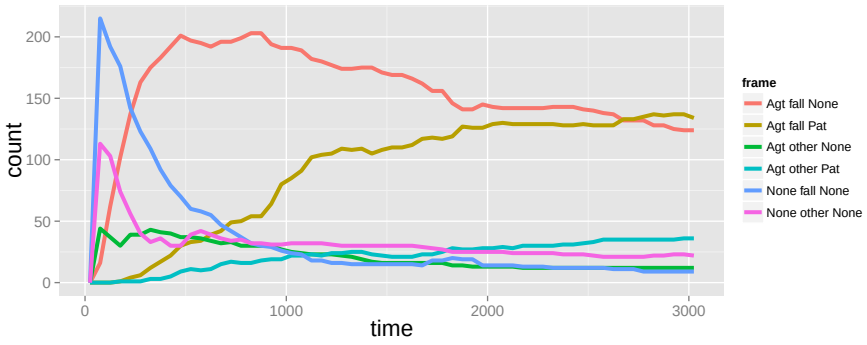


(a) Produced frames for caused-falling events over time, given a frequency of *fall* with causative meaning = 25.

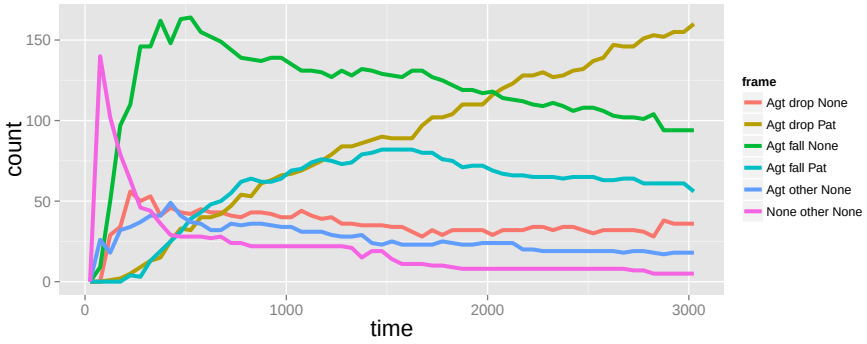


(b) Produced frames for caused-falling events over time, given a frequency of *fall* with non-causative meaning = 250.

Figure 7.6: Produced frames for caused-falling events over time with the frequency of *fall* with causative meaning as a dependent variable.



(a) Produced frames for caused-falling events over time, given a frequency of *drop* = 10.



(b) Produced frames for caused-falling events over time, given a frequency of *drop* = 100.

Figure 7.7: Produced frames for caused-falling events over time with the frequency of *drop* as a dependent variable.

7.3.3 Factors in the overgeneralization and retreat

As we saw in the inspection of the various settings, the model overgeneralizes *fall* to a transitive frame in all cases. This is not strange given the design of SPL: the model rewards expressiveness strongly, and if no alternative to a transitive construction with *fall* as the word expressing the EVENT is present, SPL will simply use that pattern. Alternatively, it uses the less expressive, but very well entrenched 'Agt fall' pattern, in which an intransitive is combined with the word *fall*. Overgeneralization is, as it were, the default state of the model: in its desire to be expressive, it will use whatever means it has available to express as much of the conceptualization of the situation as possible.

We have also seen that the model can overcome this overgeneralization, but that the alternative has to be frequent enough to outcompete *fall*. When *drop* is highly infrequent (77.5 times as infrequent as *fall*), it will not outcompete *fall*, but when it is less infrequent ('only' 7.8 times as infrequent), it will. This opens the interesting possibility that we can model the regularizations of linguistic systems through usage processes with SPL: as a diachronic model, SPL would predict that *drop* would fall out of use if its frequency were 10, and *fall* would become a transitive verb. If *drop* has a frequency of 100, however, it would remain stable in the language.

The other interesting effect is that of the frequency of *fall* with a causative meaning. If this pattern is seen often, the model is quicker to use *drop* as the expression of the causative meaning. This is remarkable, given that SPL does not negatively reinforce (or: inhibit) non-observed grammatical constructions for words. Why, then, does the frequency of causative-but-intransitive *fall* matter? It seems to me that the causative-but-intransitive *fall*-construction is acquired more readily given this setting. This construction prevents a more generic construction (with any role as the first constituent and *fall* as the second constituent) to be acquired. It is this latter, generic-intransitive-*fall* construction that causes the model to overgeneralize, and if it is 'latently pre-empted'[2] by the 'Pat fall' patterns, the 'Agt drop Pat' patterns have more of a chance of being produced.

The two factors involved in the retreat from overgeneralization show that SPL can account for explananda E4 and E5: the model overgeneralizes and retreats from it, and we can study how the frequencies of the various constructions play a role in this. A high frequency of *fall* with a causative meaning 'latently pre-empts' the use of transitive *fall*, and a high frequency of *drop* straightforwardly pre-empts the use of transitive *fall*. This suggests that pure entrenchment has no role to play and is a mere epiphenomenon. Given the various findings in experimental studies, I will leave this suggestion to future research.

The fact that SPL never produces 'Pat fall' patterns (i.e., patterns with the patient of a CAUSE-FALL event as the subject) may indicate that the expressiv-

[2] I say latently because the 'Pat fall' patterns are never produced – they do, however, take reinforcement mass away from the 'Any-Role fall' pattern.

ity constraint on generation is too strong: the model finds both the erroneous and correct patterns with two expressed arguments more likely in all cases, because they express more of the situation. It may be that taking the discourse salience of the participants into account remedies this.

7.4 Discussion

In the production experiment, SPL proves to perform reasonably well on the various tasks, making the model fully satisfy desideratum D2 (comprehensiveness) now. We have seen that the model omits increasingly less arguments over time (explanandum E1), but does not simulate the prevalence of subject omission (E2). I argued that this latter effect is due to either the model having no notion of discourse salience or its lack of a right-edge biased, a notion well established by models such as MOSAIC (Freudenthal et al. 2010).

One may wonder why I made such an effort at analyzing the errors the model makes. I believe it is in the things that the model does not do 'right', according to the target utterance, that we see how it works. The error analysis revealed the fact that lexical abstraction and grammatical abstraction seem to work differently; whereas it does not hurt to abstract any and all abstractions over grammatical constructions (they are pre-empted by more concrete ones anyway), abstracting over lexical constructions is problematic, because overly abstract word meanings emerge. This has theoretical consequences. Does it, for instance, mean that they are, despite the constructivist axiom of 'everything is a construction', different beasts? I would not be willing to draw that conclusion yet, but this is an issue that is definitely in want of further attention.

Similarly, I found that many overgeneralizations were not overcome given the set-up (maximally concrete features such as FALL and no suppletive cases for verbs like *fall*). The addition of the latter, when *drop* is defined as CAUSE-FALL, surely helps, as we have seen in section 7.3, but then the question remains: how do we implement a system in which the violation of some of the conceptual properties of the situation is allowed in a highly restricted way. Again, like the condition on expressivity, we could argue that the model has to be able to produce analyses for a situation that include features not present in the situation, at the cost of some penalty. This would allow the model to produce argument-structure patterns that match the situation better, but that also are overly specific in their features (and therefore penalized).

CHAPTER 8

Concluding remarks

Understanding how children acquire the language of their community within a limited amount of time is a central question in linguistics. The usage-based constructivist approach to language acquisition holds that children do so by using domain-general learning mechanisms such as social cognition and pattern recognizing mechanisms. Computational modeling, that is: simulating a child's behavior by formalizing and implementing important pieces of our favorite hypotheses as software, is becoming an increasingly important method in the field of language acquisition. I hope to have contributed to both the field of language acquisition and computational cognitive modeling with this dissertation by addressing four major points I presented at the outset:

- Achieving greater comprehensiveness of computational cognitive models
- Achieving greater naturalism in the computational modeling of the acquisition of meaning
- A reappreciation of the starting-small hypothesis within the usage-based framework
- A reassessment of proposed learning mechanisms (cognitive) and algorithms (computational).

I believe I have done so with the Syntagmatic-Paradigmatic Learner (SPL), a computational model of the acquisition of linguistic representations that aims to implement various aspects of a usage-based theory of language acquisition. Crucially, SPL starts off with no linguistic-representational content,

and learns to comprehend as well as produce utterances. SPL processes utterances in a context of situations (the properties of which were derived from an empirical study presented in chapter 4), and in doing so, gradually builds up a constructicon, an inventory of both lexical and grammatical constructions. The 'learning mechanisms' involved in the learning process are best thought of as mere traces of processing operations, rather than actual hypothesis-testing operations (which is the metaphor, grounded in deductivist thought, that is often used to describe the acquistion of linguistic representations).

8.1 Recapitulating SPL

Let us briefly go over the main properties of SPL once more. The model uses the representational format of the construction, a pairing of signifying elements and a signified conceptualization. Starting with no representations, it tries to parse novel input items, pairings of an utterance and a set of situations to which the utterance possibly refers.

The set of situations was generated by the input generation procedure of Alishahi & Stevenson (2010). I modified this procedure to reflect the actual properties of the situational contexts of linguistic used events, as studied in chapter 4. In that chapter, we found that the levels for noise (the absence of some conceptual target from experience) and uncertainty (the overwhelming presence of conceptual non-targets in experience) typically used in computational modeling studies are low compared to the ones we find in actual caregiver-child interaction. I studied the latter by looking at a corpus of videotaped caregiver-child interaction and annotated the corpus for all conceptual elements reasonably thought to be present in the situation around the speech situation. Another insight following from this study was that chains of events are highly dependent on each other: if the mother engages in an action with a ball, it is very likely that she will engage in another action with the ball afterwards, or perhaps in the same action with another object. Given the tediousness of hand-coding the data, this method did not prove scalable to the demands of a computational model. The study of these properties of interaction 'in the wild', however, did lead to an adaptation of Alishahi & Stevenson's (2010) input generation procedure. In this adapted procedure, we generate pairs of an utterance and the situational context in which the utterance occurs, with the latter consisting of a set of situations, one of which is the target situation, unless the target situation is absent. Notably, the similarity of the situations within the situational context, and between subsequent situational contexts to each other is given by the similarity we found in the caregiver-child interaction. Furthermore, the setting of the parameters for noise and uncertainty was derived from the video data as well.

For every processed input item, the model arrives at an optimal analysis, and does so without engaging in utterance-wide optimization. That is: SPL processes the utterance linearly and while keeping track of only the most

likely analysis up to that point. The best analysis constitutes the input for SPL's learning mechanisms. Through a set of learning mechanisms, SPL gradually builds up an inventory of constructions allowing it to comprehend and produce utterances. The learning mechanisms constitute the central innovation of the model in the aim to stay close to the usage-based approach as set out by Langacker (1988). I believe this aim has been fulfilled in the design of the model in several ways. Crucially, all of the learning mechanisms, with perhaps the exception of cross-situational learning, are online mechanisms. That is: they do not constitute post-hoc operations on the constructicon (the inventory of constructions), but rather reflect the traces left by the processing of the input item. These traces are found at several levels.

First, a trace of the most concrete representations of the utterances the processes is left in the representational system of SPL through the use of most-concrete constructions. This operation has the effect that highly concrete representations, if they are reinforced often enough, can become stronger over time. We can interpret this as the formation of category prototypes: the well-reinforced, highly-concrete representations are readily available to the model in analyzing and generating utterances.

Second, the mechanism of reinforcing the most-concrete used constructions, i.e. the most-concrete constructions, allows the model to accrue reinforcement mass for those constructions that are used frequently. The effect of this operation is that abstract constructions may obtain reinforcement if they are used to analyze utterances. Because the model only reinforces the most-concrete used construction, the reinforcement operation rewards patterns that are actually used. The usefulness of a construction is therefore determined by its frequency of use. Notably, this design feature implements Bybee's (2006) notion of type frequency. An abstract construction will typically only be reinforced once for each unique usage event for which it is used in an analysis. If the same usage event is encountered again, it is very likely that the more concrete construction blocks the use of the more abstract one. Routinization through high token frequency follows from the same learning operation: if a construction is used frequently, it is more readily available for subsequent analyses. If this construction happens to be a highly concrete one (i.e., one with many constituents lexically specified) the model will acquire such a construction as a routine.

Third, the model builds up increasingly long constructions through the use of the syntagmatization operation. Syntagmatization is the trace left by the processing of multiple, smaller, constructions for which the model has found no analysis in which they are connected to each other with a grammatical construction. These smaller constructions then form the constituents of a novel, wider, construction. Syntagmatization is the primary means through which SPL builds up grammatical constructions.

Finally, paradigmatization allows the model its potential to generalize to unseen usage events. By taking the joint structure of any two constructions that have been reinforced, the paradigmatization 'extracts' abstractions from

more concrete constructions. These abstractions, however, are only extracted in the implementational sense: as no selection over them takes place, they can be considered immanent in the more concrete constructions from which they are abstracted, by simply restating their overlap. However, through the reinforcement of the most-concrete used construction, they can be reinforced themselves, in a way akin to Langacker's (2009) description of how abstractions may obtain unit status without the more concrete patterns doing so. This way, selection of 'good' or 'useful' abstractions takes place, but without any selection mechanism performing a global evaluation of the usefulness of a novel abstraction.

The model gets off the ground by the cross-situational learning mechanism, which compares recent usage events and extracts any reliable overlap as initial lexical constructions. Another way of obtaining lexical constructions is through the bootstrap operation. Bootstrapping is a property of the utterance analysis mechanism that fills a non-phonologically-specified slot of a construction with a substring of the utterance, by assuming that substring is an actual word filling that slot.

Both cross-situational learning and bootstrapping allow for the extraction of chunks: lexical constructions that are larger than a single word in the 'adult' language. These chunks, unlike what many within the usage-based framework assume, are not broken down by the paradigmatization operation. This would require the model to engage in a post-hoc re-analysis of the chunks, which was an operation I wanted to avoid, as it makes learning more than a mere by-product of processing.

8.2 The behavior of SPL

I evaluated SPL's behavior both in a comprehension (chapter 5) and a production (chapter 7) experiment. In the comprehension experiment, I looked at the performance of the model in identifying the correct situation out of all possible situations the utterance could refer to, as well as the coverage of the utterance and the situation with the best analysis. On all three measures, SPL gradually becomes a more competent language user over time. Similarly, for production, SPL was tested by having it generate utterances on the basis of a situation and its constructicon at that point in time. The generated utterances become longer over time, and increasingly capture the linguistic material found in the utterance that would have been produced by the input generation procedure. Interestingly, the model displayed high scores of precision, or correctness, from the outset: whatever it produced was mostly correct. This is in line with the finding that children mainly make errors of omission (leaving out elements present in adults' speech), but few errors of comission (producing linguistic elements an adult would not produce).

Next, I looked at the robustness of the model. Recall that we set the parameters for the similarity of the situations in the situational context, as well

as the noise and uncertainty of the situational context on the basis of the empirical study of caregiver-child interaction. We may, however, ask how the model performs given different values for these parameters. I found that if the situations are similar to each other, the model is relatively robust to higher levels of noise and uncertainty (on the measures discussed above). Generating each situation independently of the previous one creates a situational context in which the situations are more dissimilar from each other, and in that condition, noise and uncertainty do affect the model's performance negatively. This suggest that the coherence of the situational contexts in which children have their early linguistic experiences plays an important role in bootstrapping a linguistic system: even if the child misidentifies the precise situation, the erroneously identified situation likely contains many elements that are correct.

It is, however, at a more detailed level that the interesting behavioral patterns can be seen, and especially from the failure of the model to behave as we expect, we learn important things about how the mechanisms work. In the two experimental chapters, I studied several behavioral patterns of the model in qualitative detail, to try to understand why the model behaves in certain ways.

In the production experiments, we observed that the number of expressed arguments grew over time as an effect of an increasing number of syntagmatized and subsequently paradigmatized constructions being acquired. I was not able to simulate the prevalence of subject omissions, but argued that this is likely due to a lack of pragmatics and of a right-edge processing bias, as, for instance, MOSAIC (Freudenthal et al. 2010) incorporates. What I did find was that the omission of early arguments was not only a matter of a small vocabulary: for many aspects of the situation the model had to express, it had a lexical construction available, but it simply did not have a grammatical construction ready to fit the lexical construction in. With this analysis, I provided a usage-based analysis of Berk & Lillo-Martin's (2012) finding that older children who have been deprived of linguistic input but are otherwise normally functioning, go through a two-word stage while having a far more extensive vocabulary than a eighteen-month old. An important caveat here is that the higher frequency of subject omissions over other argument omissions was not predicted by the model. Here, the model is somewhat more remote from reality. I argued that the most likely reason for this phenomenon is the information structure of discourse and the salience of the participants: if subjects typically denote less salient and discourse-given participants, we can expect them to be learned (through comprehension) and produced less frequently. An interesting extension of the current model would be to include a discourse model. This seems a relatively small step, since the current input generation procedure already involves chains of events and utterances, on the basis of which we can change the salience of certain referents and words.

A central question in language acquisition is why children sometimes overgeneralize argument-structure (and other) constructions and how they retreat from this overgeneralization. The overgeneralization of argument struc-

ture constructions and the subsequent retreat were modeled in chapter 7. The answer of SPL to these two questions is that it quickly builds up an inventory of abstract, generalizable, grammatical constructions (which it, however, hardly uses in comprehension) that it combines with verbs that cannot occur in these constructions (e.g., *you fall ball*). The presence of an alternative construction pre-empts this overgeneralization after a phase of overgeneralization. I argued that pre-emption works in two ways. First, the more entrenched this alternative construction is, the quicker the model retreats from overgeneralization. Second, we find an entrenchment effect of the 'correct' construction: when the model experiences more cases of *ball fall* with a causative meaning (someone dropping a ball), the constructions underlying such utterances are reinforced more, and because of this, highly general constructions allowing for the overgeneralization become less entrenched. I argued that, rather than describing this as entrenchment per se, we could better regard this effect as 'latent pre-emption', that is: as a pre-emption effect that is not seen in the behavior (the model does not produce *ball fall*, as it is less expressive than *you drop ball*), but that does block the use of a novel, erroneous, combination of an abstract construction and a verb.

8.3 The representations acquired by SPL

One interesting property of computational models is that we can study their representations independently of the model's behavior. I did so in chapter 6. A first finding reported there is that, even though all learning mechanisms are available over time, their use varies over time. For the acquisition of lexical constructions we found that cross-situational learning, the naïve method by means of which the model extracts similarities across linguistic usage events, is only used for the first few hundreds of input items. Afterwards, the model has built up an inventory of semi-open and open grammatical constructions that it can use to bootstrap the meaning of words it has not seen. The paradigmatization operation, secondly, displays interesting 'bursts' of activity over time, meaning that the model does not arrive at abstractions gradually, but encounters exemplars that 'unlock' new subspaces of the design space of linguistic representations.

The abstractions learned by SPL display the interesting property that they are not directly obvious from the behavior of the model in comprehension and production. If we would not have looked under the hood of the model, we might have arrived at the erroneous conclusion that its representational system is very concrete. This is a false line of reasoning: given the usage-based tenet that language users prefer the use of more concrete constructions over more abstract ones (as implemented in the probability model of SPL), we expect the highly concrete constructions to show up most of the time. However, representationally, the model has great potential for making generalizations. In fact, generalizations are found rather early, and the model spends the later

iterations mainly by adding more relatively concrete constructions to the abstract ones that pre-empt the latter. This is not strange, given the overgeneralization behavior we observe in both children and SPL: once abstraction is available, the model will use it for expressivity's sake, unless it has something more concrete that is equally expressive.

An interesting feature of the abstractions found in the model is that they clearly reflect the type frequencies of the items occurring in them (cf. Bybee 2006): the transitive construction is strongly reinforced as a non-verb-specific construction, because many verbs occur in it, whereas the caused-motion construction is only seen with two verbs, and hence reinforced in verb-island-like constructions rather than as constructions that abstract over verbs.

Reversing the perspective, we furthermore saw how certain words are more readily learned as independent lexical constructions whereas others are primarily learned as the constituents of grammatical constructions. Notably, words referring to entities ('nouns'), are typically learned as independent entities. For the other kinds of words, there was more variation, both between the words and between simulations. Pronouns are used in a lot of different contexts, hence boosting the likelihood of their independent acquisition, but they are also used frequently *within* particular constructions. What we find for pronouns, as well as for prepositions and verbs displaying similar distributions, is that they are acquired independently in some simulations, but as 'bound' elements of constructions in others. I identified three possible factors that determined a word's independence. First, the more different elements occur in a slot, the more likely it is that the abstraction over them will be used in comprehension and production, and the more likely it is that the filler word will be acquired independently. Second, the frequency of the word in the slot: the higher this value is, the more likely it is that it will not be acquired independently, as it will be reinforced as part of a grammatical construction often. Finally, the word's 'promiscuity' matters: if a word occurs across the slots of many grammatical constructions, it is more likely that it will be acquired independently.

On several aspects of the representations, we found high degrees of 'individual' variation between the simulations: the abstraction of the representations as well as the relative independence of various words varied between simulations. This is interesting, as the various simulations display grossly the same behavior – they perform equally well on the global tasks in comprehension and production. I will return to this issue in section 8.5.

8.4 Desiderata and explananda

In chapter 2, I set out a list of theoretical desiderata and empirical explananda the model has to satisfy. Previous models have made important contributions by focussing on parts of this list and my aim was to bring all insights together. I believe SPL reasonably succeeds in doing so: table 8.1 displays the list and

desideratum/explanandum	(Chang 2008)	(Alishahi & Stevenson 2008)	(Kwiatkowski 2011)	(Beekhuizen & Bod 2014)	(Freudenthal et al. 2010)	(McCauley & Christiansen 2014*a*)	SPL
D1 (explicitness)	+	+	+	+	+	+	+
D2 (comprehensiveness)	◊	−	◊	◊	−	−	+
D3 (simultaneity)	◊	−	+	+	−	−	+
D4 (representational realism)							
D4-1 (qualitative grounding)	+	+	−	+	+	+	+
D4-2 (quantitative grounding)	+	+	+	+	+	+	+
D4-3 (immanence)	+	+	−	+	+	−	+
D5 (processing realism)							
D5-1 (heterogeneous structure building)	−	−	−	−	+	−	+
D5-2 (linear processing)	−	−	−	−	+	+	+
D6 (ontogenetic realism)							
D6-1 (cumulative complexity)	◊	−	−	−	+	+	+
D6-2 (learning-by-processing)	−	+	+	+	+	+	+
D6-3 (parts-to-whole and v.v.)	+	−	−	−	+	−	+
D6-4 (developmental continuity)	+	+	+	+	+	+	
D7 (explanatory insight)	+	+/−	+	+/−	+/−	+/−	+
D3-1 (unification)	−	+	−	−	+	−	+
E1 (decreasing argument omission)	◊	−	−	−	+	−	+
E2 (prevalence of subject omission)	◊	−	−	−	+	−	−
E3 (co-varying complexity)	−	−	−	−	−	−	−
E4 (overgeneralization and retreat)	◊	+	−	−	−	−	+
E5 (mechanisms overgeneralization)	−	−	−	−	−	−	+

Table 8.1: A comparison of SPL to the various learners discussed in section 2.5.

whether or not SPL satisfies each particular desideratum or explanandum.

To the best of my knowledge, SPL constitutes the first usage-based computational model that is able to parse and generate utterances while starting with no representational content (D2 and D3). Furthermore, I believe it most closely instantiates the full set of ideas put forward within the usage based perspective: the representations are both qualitatively and quantitatively grounded in the linguistic usage events through their reinforcement in analyzing the usage event. Any learned abstractions are furthermore immanent: they merely restate commonalities across more concrete constructions. In making the analyses, SPL reasonably satisfies the constraints on the realism of processing. Although this was not the focus of this dissertation, it satisfies the baseline conditions that processing is incremental over the utterance and does not involve the search for an optimal analysis over the full utterance.

Obviously, SPL is not a complete model: no model ever is, which is why we call it a model. Several design features of SPL function as 'stubs' in the model to make it work.[1] These stubs are well grounded in our knowledge of pragmatic reasoning, linguistic processing, and learning theory, but I do see room for improvement over the current formulations: a more gradient application of them, over the discrete 'constraints' that have been formulated for the model, is definitely a locus of such improvement.

On the empirical side, more evaluation of the model to experimental data is needed. My reason to focus on 'naturalistic' comprehension and production is that the natural situation of linguistic interaction forms a baseline: if we cannot explain that, the fact that we do understand behavior in artificial settings is to my mind a worthless one. After all, this is the context in which languages are culturally evolved and where the cognitive mechanisms involved in linguistic behavior are geared towards (whether developmentally or biologically). However, once we understand the naturalistic case (to some extent), going back and forth between the evaluation on naturalistic behavior and experimentally elicited behavior vastly enriches our knowledge of the cognitive mechanisms. I hope to contribute to this evaluation in future work.

Crucially, however, SPL satisfies the developmental desiderata: it obeys to the cognitive law of cumulative complexity by gradually building up more complex representations (both in length and abstraction) from simpler ones. All learning in the model can be seen as the traces left by the processing of the usage event: there are no reorganization operations on the constructicon as a whole, nor does the model 'allow' constructions 'in' or not on the basis of how useful they are: many representations are extracted from the usage events, but only a few get reinforced in subsequent usage events. The issue of parts-to-whole and whole-to-parts learning is interesting. SPL does parts-to-whole learning by means of the syntagmatization operator, but does not break down larger units into its components (e.g., when chunks are acquired). In chapter 6, I argued that this kind of offline blame assignment may be at odds with

[1] I owe this way of regarding aspects of the model to Suzanne Stevenson (p.c.).

the idea that learning is a by-product of processing. It requires the learner to re-analyze her previous experiences in terms of a novel conception. Perhaps this is not impossible, but I believe this aspect of the starting-big conception is in want of some more elaboration. Interestingly, SPL does display kinds of whole-to-parts learning, for instance through the bootstrapping operator, whereby a novel word is learned on the basis of a larger linguistic gestalt. Finally, the learning operators are available to the model throughout time, although, as I discussed earlier, their frequencies vary.

I believe SPL provides a good example of narrowing the gap between a theoretical conception and a computational model (D7). Most aspects of the model are readily interpretable as aspects of the usage-based perspective, as I have argued in chapter 3. SPL furthermore provides some unifying explanations: effects of type frequency, token frequency, overgeneralization and the retreat from overgeneralizatio all emerge simply from the reinforcement procedure of the model by means of which the representational potential changes over time.

Looking at the explananda, finally, we see that SPL meets explananda E1, E4, and E5. I did not discuss explanandum E3 anywhere in this dissertation and have not attempted to model it myself, but I do believe it to be a crucial empirical observation that future studies should address. Explanandum E2 is not met by the model: as I argued in chapter 7 it requires either an implemented notion of discourse salience or a right-edge bias. Perhaps adding either of those to SPL may help satisfy this explanandum.

8.5 Suggestions for the usage-based conception

All of the observations discussed in sections 8.2 and 8.3 are effects found within the computational model. As such, we may easily dismiss them as artefacts of the model. I believe, however, that in many cases this is not the best thing to do. SPL instantiates a rather close implementation of a usage-based conception of language acquisition, and as such constitutes a way of studying the various aspects of a usage-based account in interaction, something not possible in the lab or from the armchair. The interpretation can lead to two kinds of conclusions: either SPL is right about some aspect of the theory, or SPL is wrong, but then a better implementation of a particular cognitive mechanism has to be proposed in order to replace the proposal made in SPL.

A first aspect of the model I would like to draw attention to is the notion of a competence-performance distinction it embodies. Looking at the behavior of the model, it seems that it only has acquired highly concrete constructions. However, these are the constructions it uses most frequently (which is why they are stored at that level of concreteness in the first place), and for rarer events, the model quickly arrives at a high level of abstraction in its representational potential. With the back-and-forths between adherents of the early-abstraction and lexical conservatism perspectives, it is hard to find empirical

data that are not contradicted by other data. The point I want to make with SPL, however, is that the usage-based perspective is not at odds with an early-abstraction view. Given the close implementation of an immanent abstraction procedure, SPL quickly arrives at abstraction. Perhaps children do so as well.

Secondly, SPL supports the view that, despite their linguistic behavior being roughly the same, different language users may have different representations from one another. We have seen this in chapter 6 for several phenomena: the number and abstraction of the constructions varies across simulations, and whereas some simulations operate on the basis of pronoun frames like '*you* X *it*', others have independent pronouns. Nevertheless, in all simulations, the model arrives at a very similar performance on the comprehension and production tasks.

In the discussion of the independence of items, a factor was found that, to my knowledge, has not been studied well within the usage-based framework. Earlier in this conclusion, I coined it 'promiscuity': the ease with which a word is used in the slots of various constructions. This may be a factor, besides type frequency of a constructional slot and the token frequency of a word in that slot and it would be interesting to study its effects on processing, both through corpus studies and experimental work.

8.6 Suggestions for cognitive modeling

A first central contribution of this dissertation is the empirical grounding of the situational context in empirical findings on the actual situational contexts in which children experience linguistic usage events. Although not scalable by itself to function as input to a computational model, the method did provide us with valuable insights in the situational contexts in which children acquire language. When studying the acquisition of meaningful units, I believe, one cannot simply make up reasonable estimations of the uncertainty and noise present in the situation, rather, an empirical grounding of these estimations is required.

Nonetheless, the way most computational models approach conceptualization is still far from perfect. Meaning is hard.[2] A future direction I would like to suggest is the combination of continuous representations with naturalistic settings. The use of resources like WordNet is simply not suited to capture the subtleties of constructional meaning, and, more importantly, displays a cultural bias. The induction of universal semantic maps and subsequent acquisition of categories within this map forms an interesting way forward (cf. Beekhuizen, Fazly & Stevenson 2014).

In computational modeling, the shadow of scalability is always looming. The case is not different for SPL, I believe. I presented the performance of

[2]Or, as Hugo Brandt Corstius, famously, and intranslatably said "Wat je ook doet, de semantiek gooit roet" (lit. 'whatever you do, semantics throws soot', 'whatever you do, semantics is a spoilsport').

the model given an empirically grounded toy setting: the model processes certain words, but not others, given a limited representation of the meaning. Nonetheless, the distribution of the words, as per Alishahi & Stevenson's (2010) input generation procedure, as well as the parameters of the situational noise and uncertainty, are grounded in empirical work. What I have attempted is to trade-off the often conflicting notions of the faithfulness to a theoretical perspective, achieving realism in the simplifying assumptions, maximizing a model's empirical coverage, and maximizing the number of things a model can do (comprehensiveness). I have mainly focussed on the faithfullness and comprehensiveness, perhaps slightly at the expense of attempting to find more empirical coverage. There is a time for everything and only through a methodologically heteredoxical approach to computational modeling can we use its full potential.

Another issue of scalability that I believe needs to be addressed in usage-based frameworks, is that of 'hard' constructions. Within the generative paradigm, several constructions have been proposed to be unlearnable from the input data alone. Two approaches are typically pursued within the usage-based framework to counter these claims, namely the reconceptualization of the construction (e.g., Verhagen (2005) for long-distance *Wh*-questions, or van Hoek (1997) for pronominal binding), and corpus-driven work, possibly involving computational models that show that one can arrive at, at least, representations leading to the correct outputs (Smets 2010, Bod & Smets 2012). However, computational models doing so typicaly do not take meaning into account. It would be interesting to see if such 'hard' constructions can be acquired in a framework such as SPL.

Finally, I believe that cognitive modeling should be used more as a tool for theory formation than only as a hypothesis testing device. Not that that latter should be done *less*, but I believe that we, as a community of linguists and cognitive scientists have not yet understood the full potential of the method, which goes well beyond the mere empirical evaluation of a theory. At all levels of the scientific process, modeling provides a tool for shaping our endeavors: as a discovery procedure, a helping hand (but also a constraint) in formulating and scrutinizing theories, a means of giving existence proofs, and a means of both making as well as evaluating predictions. I hope to have presented a case where many of these possible applications of computational modeling come together, and furthermore hope that more research along similar lines will be done.

Bibliography

Abbot-Smith, K., Lieven, E. & Tomasello, M. (2008), 'Graded representations in the acquisition of English and German transitive constructions', *Cognitive Development* **23**(1), 48–66.

Abbot-Smith, K. & Tomasello, M. (2006), 'Exemplar-learning and schematization in a usage-based account of syntactic acquisition', *The Linguistic Review* **23**(3), 275–290.

Akhtar, N. (1999), 'Acquiring basic word order: evidence for data-driven learning of syntactic structure', *Journal of Child Language* **26**(2), 339–56.

Alishahi, A. & Stevenson, S. (2008), 'A computational model of early argument structure acquisition', *Cognitive Science* **32**(5), 789–834.

Alishahi, A. & Stevenson, S. (2010), 'A computational model of learning semantic roles from child-directed language', *Language and Cognitive Processes* **25**(1), 50–93.

Ambridge, B. (2013), 'How do children restrict their linguistic generalizations? An (un-)grammaticality judgment study', *Cognitive Science* **37**(3), 508–43.

Ambridge, B. & Lieven, E. V. M. (2011), *Child Language Acquisition. Contrasting Theoretical Approaches*, Cambridge University Press, Cambridge, UK.

Ambridge, B., Pine, J. M. & Rowland, C. F. (2012), 'Semantics versus statistics in the retreat from locative overgeneralization errors', *Cognition* **123**(2), 260–79.

Ambridge, B., Pine, J. M., Rowland, C. F., Freudenthal, D. & Chang, F. (2014), 'Avoiding dative overgeneralisation errors: semantics, statistics or both?', *Language, Cognition and Neuroscience* **29**(2), 218–243.

Arnon, I. (2010), *Starting Big. The Role of Multi-Word Phrases in Language Learning and Use*, Doctoral dissertation, Stanford University.

Bailey, D. R. (1997), *When Push Comes to Shove : A Computational Model of the Role of Motor Control in the Acquisition of Action Verbs*, Doctoral dissertation, University of California, Berkeley.

Baillargeon, R. & Wang, S.-H. (2002), 'Event categorization in infancy', *Trends in Cognitive Sciences* **6**(2), 85–93.

Baker, M. (2001), *The Atoms of Language: The Mind's Hidden Rules of Grammar*, Basic Books, New York, NY.

Baldwin, D. A. (1993), 'Early referential understanding: Infants' ability to recognize referential acts for what they are', *Developmental Psychology* **29**(5), 832–843.

Bannard, C., Lieven, E. & Tomasello, M. (2009), 'Modeling children's early grammatical knowledge', *Proceedings of the National Academy of Sciences of the United States of America* **106**(41), 17284–9.

Bannard, C. & Matthews, D. (2008), 'Stored word sequences in language learning: the effect of familiarity on children's repetition of four-word combinations', *Psychological science* **19**(3), 241–8.

Beckwith, R., Tinkler, E. & Bloom, L. (1989), 'The acquisition of non-basic sentences', in *Proceedings of the Boston University Conference on Language Development*.

Beekhuizen, B. (2010), *On abstraction in construction grammar. An exercise in methodology*, M.A. thesis, Leiden University.

Beekhuizen, B. (2011), *Annotation guidelines for the Block Game Project*, Technical report.

Beekhuizen, B. & Bod, R. (2014), 'Automating Construction Work. Data-Oriented Parsing and Constructivist Accounts of Language Acquisition', in R. Boogaart, T. Colleman & G. Rutten, eds, *Extending the Scope of Construction Grammar*, Mouton, Berlin, pp. 47–74.

Beekhuizen, B., Bod, R. & Verhagen, A. (2014), 'The linking problem is a special case of a general problem none of us has solved', *Linguistics* **90**(3), e91–e96.

Beekhuizen, B., Bod, R. & Verhagen, A. (to appear), 'Acquiring relational meaning from the situational context . What linguists can learn from analyzing videotaped interaction', in J. Evers-Vermeul & E. Tribushinina, eds, *Usage-based approaches to language acquisition and language teaching*.

Beekhuizen, B., Bod, R. & Zuidema, J. (2013), 'Refining the all-fragments assumption: The search for parsimony in redundancy', *Language and Speech* **56**(3), 257–264.

Beekhuizen, B., Fazly, A., Nematzadeh, A. & Stevenson, S. (2013), 'Word learning in the wild: What natural data can tell us', in *Proceedings of the 35th Annual Meeting of the Cognitive Science Society*.

Beekhuizen, B., Fazly, A. & Stevenson, S. (2014), 'Learning meaning without primitives: Typology predicts developmental patterns', in *Proceedings of the 36th annual meeting of the Cognitive Science Society*.

Beekhuizen, B., Zuidema, J. & Bod, R. (2013), 'Three design principles of language: The search for parsimony in redundancy', *Language and Speech* **56**(2).

Berk, S. & Lillo-Martin, D. (2012), 'The two-word stage: motivated by linguistic or cognitive constraints?', *Cognitive Psychology* **65**(1), 118–40.

Bidgood, A., Ambridge, B., Pine, J. M. & Rowland, C. F. (2014), 'The retreat from locative overgeneralisation errors: A novel verb grammaticality judgment study', *PLOS one* **9**(5), e97634.

Bloom, L. (1970), *Language development: Form and function in emerging grammars*, MIT Press, Cambridge, MA.

Bloom, L. (1991), *Language development from two to three*, Cambridge University Press, New York, NY.

Bloom, L., Lightbown, P. & Hood, L. (1975), 'Structure and variation in child language', *Monographs of the Society for Research in Child Development* **40**.

Blythe, R., Smith, K. & Smith, A. D. M. (2010), 'Learning times for large lexicons through cross-situational learning', *Cognitive Science* **34**(4), 620–42.

Bod, R. (1998), *Beyond grammar: An experience-based theory of language*, CSLI, Stanford, CA.

Bod, R. (2009), 'From exemplar to grammar: A probabilistic analogy-based model of language learning', *Cognitive Science* **33**(5), 752–793.

Bod, R., Scha, R. & Sima'an, K., eds (2003), *Data-Oriented Parsing*, University of Chicago Press, Chicago, IL.

Bod, R. & Smets, M. (2012), 'Empiricist Solutions to Nativist Puzzles by means of Unsupervised TSG', in *Proceedings of the European Chapter of the Association for Computational Linguistics*.

Boogaart, R. (2009), 'Semantics and pragmatics in construction grammar: the case of modal verbs', in A. Bergs & G. Diewald, eds, *Contexts and constructions*, Benjamins, Amsterdam, pp. 213–241.

Borensztajn, G. (2011), *The neural basis of structure in language. Bridging the gap between symbolic and connectionist models of language processing*, Doctoral dissertation, University of Amsterdam.

Borensztajn, G., Zuidema, W. & Bod, R. (2009), 'Children's grammars grow more abstract with age. Evidence from an automatic procedure for identifying the productive units of language', *Topics in Cognitive Science* **1**, 175–188.

Boster, C. T. (1997), *Processing and parameter setting in language acquisition*, Doctoral dissertation, University of Connecticut.

Bowerman, M. (1974), 'Learning the structure of causave verbs: A study in the relationship of cognitive, semantic and syntactic development', *Papers and reports on Child Language Development* **8**, 142–178.

Bowerman, M. (1982), 'Evaluating competing linguistic models with language acquisition data: Implications of developmental errors with causative verbs', *Quaderni di Semantica* **3**, 5–66.

Bowerman, M. (1990), 'Mapping thematic roles onto syntactic functions: Are children helped by innate linking rules?', *Linguistics* **28**, 1253–1290.

Bowerman, M. (1993), 'Typological perspectives on language acquisition: Do crosslinguistic patterns predict development?', in E. V. Clark, ed., *Proceedings of the Twenty-fifth Annual Child Language Research Forum*, CSLI Publications, Stanford, CA, pp. 7–15.

Braine, M. D. (1963), 'On learning the grammatical order of words', *Psychological Review* **70**, 323–346.

Braine, M. D. (1976), *Children's First Word Combinations*, University of Chicago Press, Chicago, IL.

Bretherton, I., McNew, S., Snyder, L. & Bates, E. (1982), 'Individual differences at 20 months: analytic and holistic strategies in language acquisition', *Journal of Child Language* **10**(2), 293–320.

Brown, R. (1957), 'Linguistic determinism and the part of speech', *Journal of Abnormal and Social Psychology* **55**, 1–5.

Brown, R. (1973), *A First Language*, Harvard University Press, Cambridge, MA.

Brugman, H. & Russel, A. (2004), 'Annotating multimedia/multi-modal resources with ELAN', in *Proceedings of the Fourth International Conference on Language Resources and Evaluation*.

Bybee, J. (2006), 'From usage to grammar: The mind's response to repetition', *Language* **82**(4), 711–733.

Carletta, J. (1996), 'Assessing agreement on classification tasks: the kappa statistic', *Computational Linguistics* **22**(2), 249–254.

Chang, N. C.-L. (2008), *Constructing grammar: A computational model of the emergence of early constructions*, Doctoral dissertation, University of California, Berkeley.

Choi, S. (2006), 'Preverbal spatial cognition and language-specific input: Categories of containment and support', in K. Hirsh-Pasek & R. M. Golinkoff, eds, *Action Meets Word. How Children Learn Verbs*, Oxford University Press, Oxford, UK, chapter 7, pp. 191–207.

Chomsky, N. (1957), *Syntactic Structures*, Mouton, The Hague.

Chomsky, N. (1962), 'Explanatory models in linguistics', in E. Nagel, P. Suppes & A. Tarski, eds, *Logic, Methodology and Philosophy of Science*, Stanford University Press, Stanford, CA, chapter 9, pp. 528–550.

Chomsky, N. (1965), *Aspects of the Theory of Syntax*, MIT Press, Cambridge, MA.

Chomsky, N. (1975), *Reflections on Language*, Pantheon, New York, NY.

Chomsky, N. (1986), *Knowledge of Language: Its Nature, Origin, and Use*, Praeger, Westport, CT.

Chomsky, N. (1993), 'A minimalist program for linguistic theory', in K. L. Hale & S. J. Keyser, eds, *The View from Building 20: Essays in Linguistics in Honor of Sylvain Bromberger*, MIT Press, Cambridge, MA, pp. 1–52.

Clark, E. V. (2003), *First Language Acquisition*, Cambridge University Press, Cambridge, UK.

Clark, H. H. (1996), *Using Language*, Cambridge University Press, Cambridge, UK.

Croft, W. (2001), *Radical Construction Grammar: Syntactic Theory in Typological Perspective*, Oxford University Press, Oxford, UK.

Croft, W. & Cruse, D. A. (2004), *Cognitive Linguistics*, Cambridge University Press, Cambridge, UK.

Cross, T. (1977), 'Mothers' speech adjustments: the contribution of selected child listener variables', in C. Snow & C. Ferguson, eds, *Talking to Children: Language Input and Acquisition*, Cambridge University Press, Cambridge, UK, pp. 151–188.

Daelemans, W. & Van den Bosch, A. (2005), *Memory-Based Language Processing*, Cambridge University Press, Cambridge, UK.

Dąbrowska, E. (2012), 'Different speakers, different grammars: Individual differences in native language attainment', *Linguistic Approaches to Bilingualism* **2**(3), 219–253.

Dąbrowska, E. (2014), 'Recycling utterances: A speaker's guide to sentence processing', *Cognitive Linguistics* **25**(4), 617–653.

de la Higuera, C. (2010), *Grammatical Inference. Learning Automata and Grammars*, Cambridge University Press, Cambridge, UK.

de Saussure, F. (1916), *Course de Linguistique General*, Payot, Paris.

de Villiers, P. A. & de Villiers, J. G. (2000), 'Linguistic determinism and the understanding of false beliefs', in P. Mitchell & K. J. Riggs, eds, *Children's reasoning and the mind*, Psychology Press, Hove, UK, pp. 191–228.

Deprez, V. & Pierce, A. (1993), 'A cross-linguistic study of negation and functional projections in early grammar', *Linguistic Inquiry* **24**, 25–67.

Dodson, K. & Tomasello, M. (1998), 'Acquiring the transitive construction in English: the role of animacy and pronouns', *Journal of Child Language* **25**(3), 605–22.

Dowty, D. (1991), 'Thematic proto-roles and argument selection', *Language* **67**(3), 547–619.

Du Bois, J. W. (1987), 'The discourse basis of ergativity', *Language* **63**, 805–855.

Fazly, A., Alishahi, A. & Stevenson, S. (2010), 'A probabilistic computational model of cross-situational word learning', *Cognitive Science* **34**(6), 1017–1063.

Feldman, J. A. (2006), *From Molecule to Metaphor: A Neural Theory of Language*, MIT Press, Cambridge, MA.

Ferreira, F., Bailey, K. G. & Ferraro, V. (2002), 'Good-Enough Representations in Language Comprehension', *Current Directions in Psychological Science* **11**(1), 11–15.

Ferreira, F. & Patson, N. D. (2007), 'The 'Good Enough' Approach to Language Comprehension', *Language and Linguistic Compass* **1**(1-2), 71–83.

Fillmore, C. J., Kay, P. & O'Connor, M. C. (1988), 'Regularity and Idiomaticity in Grammatical Constructions: The Case of Let Alone', *Language* **63**(3), 501–538.

Fisher, C. (2002), 'The role of abstract syntactic knowledge in language acquisition: a reply to Tomasello (2000).', *Cognition* **82**(3), 259–78.

Fleischman, M. & Roy, D. K. (2005), 'Why verbs are harder to learn than nouns. Initial insights from a computational model of intention recognition in situated word learning', in *Proceedings of the 27th Annual Meeting of the Cognitive Science Society*.

Frank, M. C., Goodman, N. D. & Tenenbaum, J. B. (2008), 'A Bayesian framework for cross-situational word-learning', *Advances in Neural Information Processing Systems* **20**, 1–8.

Frank, M. C., Goodman, N. D. & Tenenbaum, J. B. (2009), 'Using speakers' referential intentions to model early cross-situational word learning', *Psychological Science* **20**(5), 578–585.

Frank, S. L. & Bod, R. (2011), 'Insensitivity of the human sentence-processing system to hierarchical structure', *Psychological Science* **22**, 829–834.

Frank, S. L., Bod, R. & Christiansen, M. H. (2012), 'How hierarchical is language use?', *Proceedings of the Royal Society B: Biological Sciences* **279**(22), 4522–4531.

Freudenthal, D., Pine, J. & Gobet, F. (2010), 'Explaining quantitative variation in the rate of Optional Infinitive errors across languages: a comparison of MOSAIC and the Variational Learning Model', *Journal of Child Language* **37**(3), 643–69.

Geeraerts, D. (2010), 'Schmidt redux: How systematic is the linguistic system if variation is rampant?', in K. Boye & E. Engeberg-Pedersen, eds, *Language Usage and Language Structure*, De Gruyter Mouton, Berlin/New York, pp. 237–262.

Gentner, D. (1978), 'On Relational Meaning: The Acquisition of verb meaning', *Child Development* **49**, 988–998.

Gentner, D. & Boroditsky, L. (2001), 'Individuation, relativity, and early word learning', in M. Bowerman & S. C. Levinson, eds, *Language Acquisition and Conceptual Development*, Cambridge University Press, Cambridge, UK, chapter 8, pp. 215–256.

Gentner, D. & Bowerman, M. (2009), 'Why some spatial semantic categories are harder to learn than others. The Typological Prevalence Hypothesis', in J. Guo, E. Lieven, N. Budwig, S. Ervin-Tripp, K. Nakamura & S. Ozcaliskan, eds, *Crosslinguistic approaches to the psychology of language. Research in the tradition of Dan Isaac Slobin*, Psychology Press, New York, NY, chapter 34, pp. 465–480.

Gigerenzer, G. (1991), 'From tools to theories: A heuristic of discovery in cognitive psychology', *Psychological Review* **98**(2), 254–267.

Gigerenzer, G. & Brighton, H. (2009), 'Homo Heuristicus: Why biased minds make better inferences', *Topics in Cognitive Science* **1**(1), 107–143.

Gleitman, L. (1990), 'Sources of Verb Meanings', *Language Acquisition* **1**(1), 3–55.

Gleitman, L. R., Cassidy, K., Nappa, R., Papafragou, A. & Trueswell, J. C. (2005), 'Hard words', *Language Learning and Development* **1**(1), 23–64.

Goldberg, A. E. (1995), *Constructions. A Construction Grammar Approach to Argument Structure*, Chicago University Press, Chicago, IL.

Goldberg, A. E. (2006), *Constructions at Work. The Nature of Generalization in Language*, Oxford University Press, Oxford.

Golinkoff, R. M., Hirsh-Pasek, K., Bailey, L. M. & Wenger, N. R. (1992), 'Children and adults use lexical principles to learn new nouns', *Developmental Psychology* **28**(1), 99–108.

Goodall, J. (1986), *The Chimpanzees of Gombe. Patterns of Behavior*, Harvard University Press, Cambridge, MA.

Graf, E., Theakston, A., Lieven, E. & Tomasello, M. (2015), 'Subject and object omission in children's early transitive constructions: A discourse-pragmatic approach', *Applied Psycholinguistics* **36**(3), 1–27.

Hespos, S. J. & Baillargeon, R. (2001), 'Reasoning about containment events in very young infants', *Cognition* **78**(3), 207–45.

Hollich, G. J., Hirsh-Pasek, K. & Golinkoff, R. M. (2000), 'Breaking the language barrier: An emergentist coalition model for the origins of word learning', *Monographs of the Society for Research in Child Development* **65**(3), 1–135.

Hyams, N. (1986), 'Core and peripheral grammar and the acquisition of inflection', in *Proceedings of the Boston University Conference on Language Development*, Boston, MA.

Hyams, N. (2011), 'Missing subjects in early child language', in J. de Villiers & T. Roeper, eds, *Handbook of generative approaches to language acquisition*, Springer Verlag, Dordrecht, the Netherlands, pp. 13–52.

Jackendoff, R. (1990), *Semantic Structures*, MIT Press, Cambridge, MA.

Jackendoff, R. (2002), *Foundations of Language: Brain, Meaning, Grammar, Evolution*, Oxford University Press, Oxford, UK.

Jackendoff, R. & Wittenberg, E. (2014), 'What you can say without syntax: A hierarchy of grammatical complexity', in F. J. Newmeyer & L. B. Preston, eds, *Measuring Linguistic Complexity*, Oxford University Press, Oxford, UK, chapter 4.

Jones, G., Gobet, F. & Pine, J. M. (2000), 'A process model of children's early verb use', in *Proceedings of the 22nd Annual Meeting of the Cognitive Science Society*, pp. 723–728.

Jurafsky, D. (1996), 'A Probabilistic Model of Lexical and Syntactic Access and Disambiguation', *Cognitive Science* **20**(2), 137–194.

Jurafsky, D. (2003), 'Probabilistic modeling in psycholinguistics: Linguistic comprehension and production', in R. Bod, J. Hay & S. Jannedy, eds, *Probabilistic Linguistics*, MIT Press, Cambridge, MA, pp. 39–96.

Jurafsky, D. & Martin, J. (2009), *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*, Prentice Hall.

Kaminski, J., Call, J. & Fischer, J. (2004), 'Word learning in a domestic dog: evidence for "fast mapping"', *Science* **304**(5677), 1682–3.

Kay, P. (2002), 'An Informal Sketch of a Formal Architecture for Construction Grammar', *Grammar* **5**, 1–19.

Kirby, S. (1999), *Function, Selection and Innateness. The Emergence of Language Universals*, Oxford University Press, Oxford, UK.

Kwiatkowski, T. (2011), *Probabilistic Grammar Induction from Sentences and Structured Meanings*, Doctoral dissertation, University of Edinburgh.

Lakoff, G. (1987), *Women, Fire and Dangerous Things: What Categories Reveal About the Mind*, Chicago University Press, Chicago, IL.

Lakoff, G. (1990), 'The invariance hypothesis: Is abstract reason based on image-schemas?', *Cognitive Linguistics* **1**, 39–74.

Landau, B., Smith, L. B. & Jones, S. (1992), 'Syntactic context and the shape bias in children's lexical learning and adults'', *Journal of Memory and Language* **31**, 807–825.

Langacker, R. W. (1987), *Foundations of Cognitive Grammar. Volume I: Theoretical Prerequisites*, Stanford University Press, Stanford, CA.

Langacker, R. W. (1988), 'A usage-based model', in B. Rudzka-Ostyn, ed., *Topics in Cognitive Linguistics*, Benjamins, Amsterdam, The Netherlands, pp. 127–161.

Langacker, R. W. (2000), 'A dynamic usage-based model', in M. Barlow & S. Kemmer, eds, *Usage Based Models of Language*, CSLI Publications, Stanford, CA, pp. 1–64.

Langacker, R. W. (2005),'Construction grammars: Cognitive, Radical and less so', in F. J. R. de Mondoza Ibañez & M. S. Peña Cervel, eds, *Cognitive Linguistics. Internal Dynamics and Interdisciplinary Interaction*, Mouton, Berlin, pp. 101–159.

Langacker, R. W. (2009), 'A dynamic view of usage and language acquisition', *Cognitive Linguistics* **20**(3), 627–640.

Lebeaux, D. & Pinker, S. (1981), 'The acquisition of the passive', in *Proceedings of the Boston University Conference on Language Development*.

Levinson, S. C., Meira, S., & The Language and Cognition Group (2003), ''Natural Concepts' in the Spatial Topological Domain – Adpositional Meanings in Crosslinguistic Perspective: An Exercise in Semantic Typology', *Language* **79**(3), 485–516.

Lewis, D. (1969), *Convention. A Philosophical Study*, Harvard University Press, Cambridge, MA.

Lieven, E., Behrens, H., Speares, J. & Tomasello, M. (2003), 'Early syntactic creativity: a usage-based approach', *Journal of Child Language* **30**(2), 333–370.

Lieven, E., Pine, J. & Baldwin, T. (1997), 'Lexically-based learning and early grammatical development', *Journal of Child Language* **24**(1), 187–219.

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

Macnamara, J. (1972), 'Cognitive basis of language learning in infants', *Psychological Review* **79**(1), 1–13.

MacWhinney, B. (1985), 'Hungarian language acquisition as an exemplification of a general model of grammatical development', in D. I. Slobin, ed., *The Crosslinguistic Study of Language Acquisition. Vol. 2: Theoretical Issues*, Erlbaum, Hillsdale, New Jersey, pp. 1069–1155.

Majid, A., Boster, J. S. & Bowerman, M. (2008), 'The cross-linguistic categorization of everyday events: a study of cutting and breaking.', *Cognition* **109**(2), 235–50.

Marcotte, J.-P. (2005), *Causative alternation errors in child language acquisition*, Doctoral dissertation, Stanford University.

Matusevych, Y., Alishahi, A. & Vogt, P. (2013), 'Automatic generation of naturalistic child-adult interaction data', in *Proceedings of the 35th Annual Meeting of the Cognitive Science Society*.

McCauley, S. M. & Christiansen, M. H. (2014*a*), 'Acquiring Formulaic Language: A Computational Model', *The Mental Lexicon* **9**(3), 419–436.

McCauley, S. M. & Christiansen, M. H. (2014*b*), 'Prospects for usage-based computational models of grammatical development: argument structure and semantic roles', *Wiley Interdisciplinary Reviews: Cognitive Science* **5**(4), 489–499.

McClelland, J. & Kawamoto, A. (1986), 'Mechanisms of sentence processing: Assigning roles to constituents of sentences', in J. L. McClelland, D. Rumelhart & T. P. research Group, eds, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume II*, MIT Press, Cambridge, MA, chapter 16.

McClure, K., Pine, J. M. & Lieven, E. V. M. (2006), 'Investigating the abstractness of children's early knowledge of argument structure', *Journal of Child Language* **33**, 693–720.

Moerk, E. (1972), 'Principles of interaction in language learning', *Merill-Palmer Quarterly* **18**, 229–257.

Naigles, L. R., Hoff, E. & Vear, D. (2009), 'Flexibility in early verb use: Evidence from a multiple-n diary study', *Monographs of the Society for Research in Child Development* **74**(2).

Needham, A. & Baillargeon, R. (1993), 'Intuitions about support in 4.5-month-old infants', *Cognition* **47**, 121–148.

O'Grady, W. (1997), *Syntactic Development*, University of Chicago Press, Chicago, IL.

Pérez-Leroux, A. T., Pirvulescu, M. & Roberge, Y. (2007), 'Null objects in child language: Syntax and the lexicon', *Lingua* **118**(3), 370–398.

Peters, A. M. (1983), *The Units of Language Acquisition*, Cambridge University Press, Cambridge, UK.

Pinker, S. (1984), *Language Learnability and Language Development*, Harvard University Press, Cambridge, MA.

Pinker, S. (1989), *Learnability and Cognition: The Acquisition of Argument Structure*, MIT Press, Cambridge, MA.

Quine, W. (1960), *Word and Object*, MIT Press, Cambridge, MA.

Radford, A. (1990), *Syntactic Theory and the Structure of English: The Nature of Early Child Grammar*, Blackwell, Oxford, UK.

Regier, T. (1992), *The Acquisition of Lexical Semantics for Spatial Terms : A Connectionist Model of Perceptual Categorization*, PhD thesis, University of California, Berkeley.

Rissanen, J. (1978), 'Modeling by shortest data description', *Automatica* **14**(5), 465–471.

Roy, D. K. & Pentland, A. P. (2002), 'Learning words from sights and sounds: a computational model', *Cognitive Science* **26**, 113–146.

Sabbagh, M. A. & Baldwin, D. A. (2005), 'Understanding the role of communicative intentions in word learning', in N. Eilan, C. Hoerl, T. McCormack & J. Roessler, eds, *Joint Attention: Communication and Other Minds*, Oxford University Press, Oxford, UK.

Sadock, J. (1982), 'The Bennish optative: The spontaneous ergative construction in child speech', in *Chicago Linguistic Society Parasession on Nondeclaratives*, pp. 186–193.

Savage-Rumbaugh, S., Murphy, J., Sevcik, R., Brakke, K., Williams, S. & Rumbaugh, D. (1993), 'Language comprehension in ape and child', *Monographs of the Society for Research in Child Development* **58**(3-4).

Scha, R. (1990), 'Taaltheorie en Taaltechnologie; Competence en Performance', in R. de Kort & G. Leerdam, eds, *Computertoepassingen in de Neerlandistiek*, LVVN, Almere, pp. 7–22.

Schlesinger, I. M. (1971), 'Production of utterances and language acquisition', in D. I. Slobin, ed., *The Ontogenesis of Grammar. A Theoretical Symposium*, Academic Press, New York, NY, pp. 63–101.

Siskind, J. M. (1996), 'A computational study of cross-situational techniques for learning word-to-meaning mappings', *Cognition* **61**(1-2), 39–91.

Skousen, R. (1989), *Analogical Modeling of Language*, Kluwer Academic Publishers, Dordrecht.

Smets, M. (2010), *A U-DOP approach to modeling language acquisition*, M.Sc. thesis, University of Amsterdam.

Smith, K., Smith, A. D. M. & Blythe, R. A. (2011), 'Cross-Situational Learning: An Experimental Study of Word-Learning Mechanisms', *Cognitive Science* **35**(3), 480–498.

Snow, C. E. (1977), 'Mothers' speech research: From input to interaction', in C. E. Snow & C. A. Ferguson, eds, *Talking to Children. Language Input and Acquisition*, Cambridge University Press, Cambridge, UK, chapter 1, pp. 31–50.

Steedman, M. (2000), *The Syntactic Process*, MIT Press, Cambridge, MA.

Stevens, J. S. (2011), 'Learning Object Names in Real Time with Little Data', in *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*.

Stolcke, A. (1994), *Bayesian Learning of Probabilistic Language Models*, Doctoral dissertation, University of California, Berkeley.

Theakston, A. L., Maslen, R., Lieven, E. V. M. & Tomasello, M. (2012), 'The acquisition of the active transitive construction in English: A detailed case study', *Cognitive Linguistics* **23**(1), 91–128.

Tomasello, M. (1992), *First Verbs: A Study of Early Grammatical Development*, Cambridge University Press, Cambridge, UK.

Tomasello, M. (1995), 'Pragmatic contexts for early verb learning', in M. Tomasello & W. E. Merriman, eds, *Beyond Names for Things. Young Children's Acquisition of Verbs*, Lawrence Erlbaum Associates, Hillsale, New Jersey, chapter 5, pp. 115–146.

Tomasello, M. (1999), *The Cultural Origins of Human Cognition*, Harvard University Press, Cambridge, MA.

Tomasello, M. (2003), *Constructing a Language: A Usage-Based Theory of Language Acquisition*, Harvard University Press, Cambridge, MA.

Tomasello, M. (2008), *Origins of Human Communication*, MIT Press, Cambridge, MA.

Tomasello, M. & Farrar, M. J. (1986), 'Joint attention and early language', *Child Development* **57**(6), 1454–63.

Valian, V. (2009), 'Abstract linguistic representations and innateness. The development of determiners', in W. Lewis, S.Karimi, H. Harley & S. Farrar, eds, *Time and Again: Theoretical Perspectives on Formal Linguistics in Honor of D. Terence Langendoen*, John Benjamins, Amsterdam, The Netherlands, pp. 189–206.

van Hoek, K. (1997), *Anaphora and Conceptual Structure*, Chicago University Press, Chicago, IL.

van Trijp, R. (2008), *Analogy and Multi-Level Selection in the Formation of a Case Grammar*, Doctoral dissertation, Universiteit Antwerpen.

Verhagen, A. (2005), *Constructions of Intersubjectivity*, Oxford University Press, Oxford, UK.

Verhagen, A. (2006), 'Dreigt polysemie uit de hand te lopen? Een oefening in semantische categorisering', *Voortgang: Jaarboek van de Neerlandistiek* **24**, 159–168.

Verhagen, A. (2009), 'The conception of constructions as complex signs. Emergence of structure and reduction to usage', *Constructions and Frames* **1**, 119–152.

Wexler, K. & Culicover, P. (1980), *Formal principles of language acquisition*, MIT Press, Cambridge, MA.

Xu, F. & Tenenbaum, J. B. (2000), Word learning as Bayesian inference., in *Proceedings of the 22nd Annual Meeting of the Cognitive Science Society*.

Yang, C. (2002), *Knowledge and learning in natural language*, Oxford University Press, Oxford, UK.

Yang, C. (2011), 'A Statistical Test for Grammar', in *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, number June, Portland, Oregon, pp. 30–38.

Yu, C. & Smith, L. B. (2007), 'Rapid word learning under uncertainty via cross-situational statistics', *Psychological Science* **18**(5), 414–20.

Zipf, G. K. (1935), *Psycho-biology of Languages*, MIT Press, Cambridge, MA.

Summary

Understanding how children acquire the language of their community within a limited amount of time is a central question in linguistics. The usage-based constructivist approach to language acquisition holds that children do so by using domain-general learning mechanisms such as social cognition and pattern recognizing mechanisms. Computational cognitive modeling (simulating a child's behavior by formalizing and implementing important aspects of these hypotheses as software) is becoming an increasingly important method in the field of language acquisition. This dissertation addresses four central issues in the field of language acquisition and computational cognitive modeling:

- Achieving greater comprehensiveness of computational cognitive models: the model should be able to produce, as well as interpret utterances, and not just a part of the process.
- Achieving greater naturalism in the computational modeling of the acquisition of meaning: the interpretability of utterances should be as realistic as possible.
- A reappreciation of the starting-small hypothesis within the usage-based framework: children do not only break down larger wholes into their component parts, they also learn to arrive at larger linguistic structures by combining smaller ones.
- A reassessment of proposed learning mechanisms (cognitive) and algorithms (computational): many learning mechanisms are still framed in deductivist or rationalist terms, two perspectives on cognition which do not connect naturally to the usage-based approach.

Besides these particular theoretical issues, I set out a list of general theoretical desiderata and empirical explananda the model has to satisfy in chapter 2. Previous models have made important contributions by focussing on parts of

this list and my main aim in developing yet another model was to bring these insights together.

If we want to build a comprehensive model, that is: one that can interpret as well as produce utterances, we need to have a hypothesis of how children arrive at an understanding of the communicative intention without the help of language. Computational models that deal with meaning typically have a set of situations to which the utterance potentially refers. In chapter 4, I studied the realism of this assumption. I found that the levels for noise (the absence of the meaning of an element of the utterance from experience) and uncertainty (the overwhelming presence of possible meanings that are not referred to in experience) typically used in computational modeling studies are low compared to the ones we find in actual caregiver-child interaction. I studied the latter by looking at a corpus of videotaped caregiver-child interaction and annotated the corpus for all conceptual elements reasonably thought to be present in the situation around the speech situation. Another insight following from this study was that chains of events are highly dependent on each other: if the mother engages in an action with a ball, it is very likely that she will engage in another action with the ball afterwards, or perhaps in the same action with another object. Given the tediousness of hand-coding the data, this method did not prove scalable to the demands of a computational model. The study of these properties of interaction 'in the wild', however, did lead to an adaptation of Alishahi & Stevenson's (2010) input generation procedure. In this adapted procedure, we generate pairs of an utterance and the situational context in which the utterance occurs, with the latter consisting of a set of situations, one of which is the target situation, unless the target situation is absent. Notably, the similarity of the situations within the situational context, and between subsequent situational contexts to each other is given by the similarity we found in the caregiver-child interaction. Furthermore, the setting of the parameters for noise and uncertainty was derived from the video data as well.

In chapter 3 I formalize the model: the Syntagmatic-Paradigmatic Learner (SPL). The model starts off with no linguistic-representational content, and learns to comprehend as well as produce utterances. SPL processes utterances in a context of situations, and in doing so, gradually builds up a constructicon, an inventory of both lexical and grammatical constructions. The 'learning mechanisms' involved in the learning process are best thought of as mere traces of processing operations, rather than actual hypothesis-testing operations (which is the metaphor, grounded in deductivist thought, that is often used to describe the acquistion of linguistic representations). SPL uses the representational format of the construction, a pairing of signifying elements (both phonological and conceptual) and a signified conceptualization.

For every processed input item, the model arrives at an optimal analysis, and does so without engaging in utterance-wide optimization. That is: SPL processes the utterance linearly and while keeping track of only the most likely analysis up to that point. The best analysis constitutes the input for SPL's

learning mechanisms. Through a set of learning mechanisms, SPL gradually builds up an inventory of constructions allowing it to comprehend and produce utterances. The learning mechanisms constitute the central innovation of the model in the aim to stay close to the usage-based approach as set out by Langacker (1988). I believe this aim has been fulfilled in the design of the model in several ways. Crucially, all of the learning mechanisms, with perhaps the exception of cross-situational learning, are online mechanisms. That is: they do not constitute post-hoc operations on the constructicon (the inventory of constructions), but rather reflect the traces left by the processing of the input item. These traces are found at several levels.

First, a trace of the most concrete representations of the utterances the processes is left in the representational system of SPL through the use of most-concrete constructions. This operation has the effect that highly concrete representations, if they are reinforced often enough, can become stronger over time. We can interpret this as the formation of category prototypes: the well-reinforced, highly-concrete representations are readily available to the model in analyzing and generating utterances.

Second, the mechanism of reinforcing the most-concrete used constructions, i.e. the most-concrete constructions, allows the model to accrue reinforcement mass for those constructions that are used frequently. The effect of this operation is that abstract constructions may obtain reinforcement if they are used to analyze utterances. Because the model only reinforces the most-concrete used construction, the reinforcement operation rewards patterns that are actually used. The usefulness of a construction is therefore determined by its frequency of use. Notably, this design feature implements Bybee's (2006) notion of type frequency. An abstract construction will typically only be reinforced once for each unique usage event for which it is used in an analysis. If the same usage event is encountered again, it is very likely that the more concrete construction blocks the use of the more abstract one. Routinization through high token frequency follows from the same learning operation: if a construction is used frequently, it is more readily available for subsequent analyses. If this construction happens to be a highly concrete one (i.e., one with many constituents lexically specified) the model will acquire such a construction as a routine.

Third, the model builds up increasingly long constructions through the use of the syntagmatization operation. Syntagmatization is the trace left by the processing of multiple, smaller, constructions for which the model has found no analysis in which they are connected to each other with a grammatical construction. These smaller constructions then form the constituents of a novel, wider, construction. Syntagmatization is the primary means through which SPL builds up grammatical constructions.

Finally, the paradigmatization operation allows the model its potential to generalize to unseen usage events. By taking the joint structure of any two constructions that have been reinforced, the paradigmatization 'extracts' abstractions from more concrete constructions. These abstractions, however, are

only extracted in the implementational sense: as no selection over them takes place, they can be considered immanent in the more concrete constructions from which they are abstracted, by simply restating their overlap. However, through the reinforcement of the most-concrete used construction, they can be reinforced themselves, in a way akin to Langacker's (2009) description of how abstractions may obtain unit status without the more concrete patterns doing so. This way, selection of 'good' or 'useful' abstractions takes place, but without any selection mechanism performing a global evaluation of the usefulness of a novel abstraction.

The model gets off the ground by the cross-situational learning mechanism, which compares recent usage events and extracts any reliable overlap as initial lexical constructions. Another way of obtaining lexical constructions is through the bootstrap operation. Bootstrapping is a property of the utterance analysis mechanism that fills a non-phonologically-specified slot of a construction with a substring of the utterance, by assuming that substring is an actual word filling that slot.

Both cross-situational learning and bootstrapping allow for the extraction of chunks: lexical constructions that are larger than a single word in the 'adult' language. These chunks, unlike what many within the usage-based framework assume, are not broken down by the paradigmatization operation. This would require the model to engage in a post-hoc re-analysis of the chunks, which was an operation I wanted to avoid, as it makes learning more than a mere by-product of processing.

I argued in chapter 3 that the developed model reasonably succeeds in satisfying the desiderata set out in chapter 2. To the best of my knowledge, it constitutes the first usage-based computational model that is able to analyze and produce utterances while starting its development with no representational content. Furthermore, I believe it most closely instantiates the full set of ideas put forward within the usage based perspective: the representations are both qualitatively and quantitatively grounded in the linguistic usage events: their reinforcement depends on their frequency of use in analyzing linguistic usage events. Any learned abstractions are furthermore immanent: they merely restate commonalities across more concrete constructions rather than extracting novel cognitive representations from the more concrete constructions. In analyzing utterances, SPL reasonably satisfies the constraints on the realism of processing. Although this was not the focus of this dissertation, it satisfies the baseline conditions that processing is incremental over the utterance and does not involve the search for an optimal analysis over the full utterance.

I evaluated SPL's behavior both in a comprehension (chapter 5) and a production (chapter 7) experiment. In the comprehension experiment, I looked at the performance of the model in identifying the correct situation out of all possible situations the utterance could refer to, as well as the coverage of the utterance and the situation with the best analysis. On all three measures, SPL gradually becomes a more competent language user over time. Similarly, for production, SPL was tested by having it generate utterances on the basis of

a situation and its constructicon at that point in time. The generated utterances become longer over time, and increasingly capture the linguistic material found in the utterance that would have been produced by the input generation procedure. Interestingly, the model displayed high scores of precision, or correctness, from the outset: whatever it produced was mostly correct. This is in line with the finding that children mainly make errors of omission (leaving out elements present in adults' speech), but few errors of comission (producing linguistic elements an adult would not produce).

Next, I looked at the robustness of the model. Recall that we set the parameters for the similarity of the situations in the situational context, as well as the noise and uncertainty of the situational context on the basis of the empirical study of caregiver-child interaction. We may, however, ask how the model performs given different values for these parameters. I found that if the situations are similar to each other, the model is relatively robust to higher levels of noise and uncertainty (on the measures discussed above). Generating each situation independently of the previous one creates a situational context in which the situations are more dissimilar from each other, and in that condition, noise and uncertainty do affect the model's performance negatively. This suggest that the coherence of the situational contexts in which children have their early linguistic experiences plays an important role in bootstrapping a linguistic system: even if the child misidentifies the precise situation, the erroneously identified situation likely contains many elements that are correct.

It is, however, at a more detailed level that the interesting behavioral patterns can be seen, and especially from the failure of the model to behave as we expect, we learn important things about how the mechanisms work. In the two experimental chapters, I studied several behavioral patterns of the model in qualitative detail, to try to understand why the model behaves in certain ways.

In the production experiments, we observed that the number of expressed arguments grew over time as an effect of an increasing number of syntagmatized and subsequently paradigmatized constructions being acquired. I was not able to simulate the prevalence of subject omissions, but argued that this is likely due to a lack of pragmatics and of a right-edge processing bias. What I did find was that the omission of early arguments was not only a matter of a small vocabulary: for many aspects of the situation the model had to express, it had a lexical construction available, but it simply did not have a grammatical construction ready to fit the lexical construction in.

A central question in language acquisition is why children sometimes overgeneralize argument-structure (and other) constructions and how they retreat from this overgeneralization. The overgeneralization of argument structure constructions and the subsequent retreat were modeled in chapter 7. The answer of SPL to these two questions is that it quickly builds up an inventory of abstract, generalizable, grammatical constructions (which it, however, hardly uses in comprehension) that it combines with verbs that cannot occur in these constructions (e.g., *you fall ball*). The presence of an alternative con-

struction pre-empts this kind of combinations after a phase of overgeneralization. I argued that pre-emption works in two ways. First, the more entrenched this alternative construction is, the quicker the model retreats from overgeneralization. Second, we find an entrenchment effect of the 'correct' construction: when the model experiences more cases of *ball fall* with a causative meaning (someone dropping a ball), the constructions underlying such utterances are reinforced more, and because of this, highly general constructions allowing for the overgeneralization become less entrenched. I argued that, rather than describing this as entrenchment per se, we could better regard this effect as 'latent pre-emption', that is: as a pre-emption effect that is not seen in the behavior (the model does not produce *ball fall*, as it is less expressive than *you drop ball*), but that does block the use of a novel, erroneous, combination of an abstract construction and a verb.

One interesting property of computational models is that we can study their representations independently of the model's behavior. I did so in chapter 6. A first finding reported there is that, even though all learning mechanisms are available over time, their use varies over time. For the acquisition of lexical constructions we found that cross-situational learning, the naïve method by means of which the model extracts similarities across linguistic usage events, is only used for the first few hundreds of input items. Afterwards, the model has built up an inventory of semi-open and open grammatical constructions that it can use to bootstrap the meaning of words it has not seen. The paradigmatization operation, secondly, displays interesting 'bursts' of activity over time, meaning that the model does not arrive at abstractions gradually, but encounters exemplars that 'unlock' new subspaces of the design space of linguistic representations.

The abstractions learned by SPL display the interesting property that they are not directly obvious from the behavior of the model in comprehension and production. If we would not have looked under the hood of the model, we might have arrived at the erroneous conclusion that its representational system is very concrete. This is a false line of reasoning: given the usage-based tenet that language users prefer the use of more concrete constructions over more abstract ones (as implemented in the probability model of SPL), we expect the highly concrete constructions to show up most of the time. However, representationally, the model has great potential for making generalizations. In fact, generalizations are found rather early, and the model spends the later iterations mainly by adding more relatively concrete constructions to the abstract ones that pre-empt the latter. This is not strange, given the overgeneralization behavior we observe in both children and SPL: once abstraction is available, the model will use it for expressivity's sake, unless it has something more concrete that is equally expressive.

An interesting feature of the abstractions found in the model is that they clearly reflect the type frequencies of the items occurring in them: the transitive construction is strongly reinforced as a non-verb-specific construction, because many verbs occur in it, whereas the caused-motion construction is

only seen with two verbs, and hence reinforced in verb-island-like constructions rather than as constructions that abstract over verbs.

Reversing the perspective, we furthermore saw how certain words are more readily learned as independent lexical constructions whereas others are primarily learned as the constituents of grammatical constructions. Notably, words referring to entities ('nouns'), are typically learned as independent entities. For the other kinds of words, there was more variation, both between the words and between simulations. Pronouns are used in a lot of different contexts, hence boosting the likelihood of their independent acquisition, but they are also used frequently *within* particular constructions. What we find for pronouns, as well as for prepositions and verbs displaying similar distributions, is that they are acquired independently in some simulations, but as 'bound' elements of constructions in others. I identified three possible factors that determined a word's independence. First, the more different elements occur in a slot, the more likely it is that the abstraction over them will be used in comprehension and production, and the more likely it is that the filler word will be acquired independently. Second, the frequency of the word in the slot: the higher this value is, the more likely it is that it will not be acquired independently, as it will be reinforced as part of a grammatical construction often. Finally, the word's 'promiscuity' matters: if a word occurs across the slots of many grammatical constructions, it is more likely that it will be acquired independently.

On several aspects of the representations, we found high degrees of 'individual' variation between the simulations: the abstraction of the representations as well as the relative independence of various words varied between simulations. This is interesting, as the various simulations display grossly the same behavior – they perform equally well on the global tasks in comprehension and production.

Samenvatting

Hoe kinderen binnen zo'n korte tijd de taal van hun gemeenschap verwerven, is een van de centrale vragen in de taalkunde. De gebruiksgebaseerde, constructivistische benadering van taalverwerving stelt dat kinderen dit doen door gebruik te maken van domein-algemene leermechanismes zoals sociale cognitie en patroonherkenning. Het computationeel modelleren hiervan (dus: het nabootsen van het gedrag van een kind door belangrijke aspecten van de gebruiksgebaseerde hypothese the formaliseren en implementeren als computerprogramma's) wordt een steeds belangrijkere methode in het veld van de kindertaalverwerving. Dit proefschrift snijdt vier centrale zaken in het computationeel modelleren van de gebruiksgebaseerde benadering aan:

- Het bereiken van een grotere omvattendheid van de computationele cognitieve modellen: het model moet het hele process, van geen tot veel taalkennis, zowel in productie als begrip, kunnen uitvoeren.
- Het bereiken van een naturalistischere manier van het modelleren van betekenisverwerving: de informatie die een kind tot haar beschikking heeft, moet zo getrouw mogelijk aan het model gegeven worden.
- Een herwaardering van de 'begin-klein' hypothese binnen de gebruiksgebaseerde theorie: kinderen breken niet alleen grotere gehelen op in hun delen (zoals de 'begin-groot' hypothese stelt), maar leren ook grotere gehelen te vormen op basis van de kleine delen.
- Een nader onderzoek naar de aard en conceptualisatie van de leermechanismes en de algoritmes die deze mechanismes instantiëren in het modelleren ervan: veel leermechanismes worden nog steeds beschreven in deductivistische of rationalistische termen, terwijl deze twee perspectieven geen natuurlijke denkwijzen binnen de gebruiksgebaseerde benadering zijn.

Naast deze specifieke zaken, bespreek ik in hoofdstuk 2 een verzameling algemene theoretische desiderata en empirische explananda waaraan een computationeel model zou moeten voldoen. Eerdere modellen hebben belangrijke voortgang op deze desiderata en explananda gemaakt, en een belangrijk doel van deze dissertatie is het bijeenbrengen van deze inzichten.

Als we een omvattend model willen ontwikkelen (d.w.z. een model dat zowel taalbegrip als -productie nabootst), hebben we een antwoord nodig op de vraag hoe kinderen de communicatieve intentie van de spreker begrijpen zonder de talige elementen die deze uitdrukken te begrijpen. Computationele modellen die zich met betekenisverwerving bezig houden, nemen meestal aan dat het kind beschikt over een verzameling conceptualisaties van de situatie waar de taaluiting over zou kunnen gaan. In hoofdstuk 4, heb ik de aannemelijkheid van deze aanname onderzocht. Ik bevond dat de niveau's van 'ruis' (de afwezigheid van in de zin aanwezige betekeniselementen uit de situationele context) en 'onzekerheid' (de mate van aanwezigheid van in de situationele context aanwezige mogelijke betekeniselementen die niet in de zin worden uitgedrukt) in de meeste computationele modellen laag worden ingeschat vergeleken met de waardes die we in eigenlijke ouder-kindinteractie aantreffen. Deze conclusie bereikte ik door middel van een studie van een grote verzameling ouder-kindinteracties waarin alle conceptuele elementen in de aandacht van het kind en de ouder en alle taaluitingen precies beschreven zijn. Een verder inzicht uit deze studie was dat opeenvolgende gebeurtenissen in hoge mate afhankelijk van elkaar zijn: als de ouder op het ene moment een handeling op een bal uitvoert, dan is het zeer waarschijnlijk dat deze nog een handeling op die bal zal uitvoeren erna (i.p.v. op een ander object).

Aangezien het met de hand beschrijven van de data een tijdrovend proces is, is deze methode niet op te schalen naar de hoeveelheid data die een computationeel model nodig heeft. De studie naar de eigenschappen van ouder-kindinteractie 'in het wild' gaf ons evenwel wel een mogelijkheid om de methode van Alishahi & Stevenson (2010) om kunstmatig input voor het model te genereren, aan te passen. Deze aangepaste procedure genereert input items, paren van een taaluiting en een situationele context op basis van een realistische inschatting van de waarden van 'ruis' en 'onzekerheid'. Daarnaast vormt de situationele context een keten: welke situatie de volgende zal zijn, hangt af van de huidige situatie.

In hoofdstuk 3 formaliseer ik het model: de Syntagmatisch-Paradigmatische Leerder (SPL). Het model begint zonder enige talige kennisinhouden, en leert geleidelijk zinnen te begrijpen en te produceren. In het proberen zinnen in situationele contexten te verwerken, bouwt SPL een constructicon op, een inventaris van zowel lexicale als grammaticale constructies. De leermechanismes van SPL kunnen het best gezien worden als de sporen van het verwerken van taal (i.p.v. als hypothese-testende operaties, zoals taalverwerving vaak, op deductivistische wijze, voorgesteld wordt). De constructies die SPL leert zijn paren van signifiërende elementen (fonologische en conceptuele structuren) en gesignifieerde conceptuele

structuur.

Het model verwerkt elk input-item met het constructicon dat het op dat moment heeft. Deze verwerking geschiedt zonder een optimalisatie over de hele zin: SPL verwerkt de zin lineair en onthoudt alleen de meest waarschijnlijke analyse na elk woord. De analyse die hier uitkomt vormt de input voor de leerprocedure. Door middel van een verzameling leermechanismes vormt SPL nieuwe constructies en versterkt het bestaande. De leermechanismes vormen de centrale theoretische vernieuwing van het model door nauwgezet de gebruiksgebaseerde benadering van Langacker (1988) te operationaliseren en implementeren. Het model bereikt dit doel doordat alle leermechanismes, mogelijk met uitzondering van het cross-situationele leren (zie beneden), online leermechanismes zijn. Dat wil zeggen: het zijn geen post-hoc, of achteraf plaatsvindende, operaties op het constructicon die het constructicon reorganiseren. De leermechanismes kunnen gezien worden als de sporen die het verwerken van de zinnen in hun situationele contexten achterlaten in de geest van de taalgebruiker. Deze sporen kunnen op verschillende niveau's gevonden worden.

Ten eerste laten de concreetste representaties van de uitingen die verwerkt worden een spoor achter op het representationele systeem. Deze operatie heeft het effect dat zeer concrete constructies, als ze maar vaak genoeg gebruikt worden, met de tijd representationeel sterker kunnen worden. We kunnen dit effect interpreteren als de formatie van prototypes van categorieën: de vaak versterkte, zeer concrete representaties zijn hogelijk toegankelijk voor het model in het analyseren en produceren van uitingen.

Ten tweede worden de concreetste *gebruikte* constructies versterkt. Dit zijn de constructies die het model heeft gebruikt in de analyse, maar die mogelijk abstracter zijn dan de gebruikssituatie zelf. De versterking van deze representatie staat het model toe om vaak gebruikte abstractere constructies in representationele kracht te laten groeien. Het potentieel van een constructie hangt daarmee dus af van de frequentie van gebruik ervan. Dit kenmerk van SPL instantieert de notie van het effect van typefrequentie, zoals Bybee (2006) die bespreekt. Een abstracte constructie zal normaliter alleen één maal versterkt worden voor elk uniek gebruiksgeval waar het voor gebruikt wordt. Wanneer het model datzelfde gebruiksgeval opnieuw tegenkomt, zal het model immers een concretere constructie hebben (door het eerste leermechanisme) die dat gebruiksgeval beter dekt. Dit mechanisme instantieert niet alleen het effect van typefrequentie, maar ook van tokenfrequentie: als een constructie frequent gebruikt wordt, zal zijn representationele kracht toenemen en een cognitieve routine gaan vormen

Ten derde bouwt het model toenemend lange constructies op door middel van de syntagmatizatie-operatie. Syntagmatizatie is het spoor dat achtergelaten wordt in de representaties wanneer er meerdere, kortere, constructies naast elkaar verwerkt worden. Deze kortere constructies vormen dan tezamen de constituenten van een langere constructie. Syntagmatizatie is een bottom-up leerprocedure waarmee SPL grotere representationele eenheden opbouwt.

Ten slotte vormt het model abstracties met de paradigmatizatie-operatie. Deze operatie staat het model toe taalkennis op ongeziene uitingen toe te passen. Paradigmatizatie neemt de gemeenschappelijke elementen van twee constructies en 'onttrekt' deze om er een nieuwe constructie van te vormen. Deze abstracties zijn evenwel alleen onttrokken in de implementatie. Er vindt immers geen selectie over de abstracties plaats en derhalve kunnen ze als 'immanent' in de concretere constructies worden gezien. Door de versterking van de concreetste gebruikte constructie (het tweede leermechanisme) kunnen ze zelf ook representationele kracht opdoen (cf. de beschrijving van Langacker (2009) van hoe abstracties de status van een 'unit' kunnen bereiken). Op deze manier vindt de selectie van 'goede' of 'bruikbare' abstracties plaats zonder dat het model een globale evaluatie van het constructicon uitvoert.

De eerste representaties van het model komen tot stand door cross-situationeel leren. Dit mechanisme vergelijkt recente gebruiksgevallen met elkaar en onttrekt alle overlappen in uiting en situatie tussen deze als initiële lexicale constructies. Een tweede manier om nieuwe constructies te ontdekken is door het gebruik van de 'bootstrapping'-operator. Bootstrapping is het vermogen van het model om een niet-fonologisch gevulde constituent van een constructie van toepassing te laten zijn op een woord, zonder dat het dat woord kent.

Beide leermechanismes zorgen ervoor dat het model 'chunks' kan extraheren: intern niet geanalyseerde lexicale constructies die groter zijn dan een enkel woord in de taal van een volwassene. Deze chunks worden evenwel niet in hun delen opgebroken door de paradigmatizatie-operatie. Dit zou immers veronderstellen dat het model ze achteraf, en dus niet on-line herinterpreteert, en dat is een operatie die ik wilde voorkomen om het leren daadwerkelijk een neveneffect van het verwerken te laten zijn.

In hoofdstuk 3 betoog ik dat dit model behoorlijk goed de desiderata instantieert. Voor zover ik weet, is het het eerste gebruiksgebaseerde computationele model dat zowel zinnen kan begrijpen als produceren zonder dat het met enige representationele kennis van de taal begint. SPL instantieert verder belangrijke aspecten van de gebruiksgebaseerde theorie: de representaties zijn zowel kwalitatief als kwantitatief gegrond in de talige gebruiksgevallen, hun representationele kracht hangt af van de frequentie van gebruik, de geleerde abstracties zijn immanent, en SPL verwerkt zinnen op een redelijk realistische wijze (lineair en zonder alle mogelijke analyses bij te houden.

SPL's gedrag wordt vervolgens geëvalueerd in hoofdstukken 5 (begrip) en 7 (productie). In het begripsexperiment onderzoek ik hoe goed het model de juiste interpretatie aan een zin kan geven door het te laten kiezen uit een aantal mogelijke interpretaties. Daarnaast bekijk ik hoe goed SPL de zin en de geïnterpreteerde situatie dekt met de beschikbare taalkennis. Op alle drie de vlakken zien we dat SPL een steeds competentere taalgebruiker wordt. In de productie-experimenten zien we dat, gegeven een situatie die uitgedrukt moet worden, SPL steeds langere en adequatere zinnen vormt. De fouten die het model maakt zijn voornamelijk fouten van weglating (het niet produceren

van woorden die een volwassene wel zou produceren) en niet van toevoeging (het wel produceren van woorden die een volwassene niet zou produceren. Dit is in overeenstemming met wat kinderen doen.

Daarnaast observeer ik in beide hoofdstukken dat het model robuust is. De parameters voor ruis en onzekerheid zijn voor de hierboven besproken experimenten ingesteld op grond van het onderzoek in hoofdstuk 4, maar we kunnen ons afvragen hoe het model presteert als we deze waardes hoger leggen. SPL blijkt relatief goed hogere niveaus van ruis en onzekerheid aan te kunnen, als de situaties maar een keten vormen. Zodra de 'coherentie' van de situatie wegvalt, gaat de prestatie beduidend achteruit. Deze bevinding suggereert dat de coherentie van de situatie een belangrijke rol speelt. De reden hiervoor zou kunnen zijn dat zelfs bij de misidentificatie van de situatie, er nog steeds relatief veel elementen in die verkeerd geïdentificeerde situatie zijn die wel correct zijn.

Naast deze algemene tendensen kijk ik in de hoofdstukken 5 en 7 ook op een gedetailleerder niveau naar het gedrag van het model. In de productie-experimenten, bijvoorbeeld, observeerde ik dat het aantal uitgedrukte argumenten groeide als een effect van het toenemend aantal grammaticale constructies, en niet per se als gevolg van een toenemend aantal woorden. In veel gevallen had het model de juiste woorden wel, maar produceerde het deze toch niet, omdat het geen grammaticale constructies had om deze woorden mee te combineren. Het bekende effect dat vooral grammaticale onderwerpen worden weggelaten kon ik niet nabootsen, maar dit is, m.i., te verklaren doordat het model geen pragmatische kennis meeneemt en doordat het geen focus op de rechterkant van de zin heeft, twee fenomenen waarvan we weten dat ze van invloed zijn op het weglaten van onderwerpen.

Een centrale vraag in de kindertaalverwerving is waarom kinderen soms argumentstructuurconstructies gebruiken waar volwassenen deze niet zouden gebruiken en hoe ze dit gedrag 'afleren'. Deze twee fenomenen zijn gemodelleerd in hoofdstuk 7. Het antwoord dat SPL biedt is dat het vrij snel een inventaris van abstracte, en dus generaliseerbare constructies opbouwt en die vrij vroeg combineert met werkwoorden waar deze constructies niet mee gecombineerd kunnen worden (bv. *you fall ball* in een situatie waarin iemand een bal laat vallen). De aanwezigheid van alternatieve constructies (bv. *you drop ball*) voorkomt, even later, dat deze combinaties nog gebruikt worden. Ik betoog in dat hoofdstuk dat het blokkeren van de 'foute' constructie op twee manieren gebeurt. Ten eerste is de mate van representationele kracht van de alternatieve constructie van belang: hoe sterker deze is, hoe sneller het model de overgeneralisatie 'afleert'. Ten tweede vinden we het effect dat hoe vaker het model zinnen als *ball fall* tegenkomt, hoe waarschijnlijker de blokkade van *you fall ball* wordt. Dit laatste effect is een geval van latente blokade: we zien het niet terug in het gedrag, aangezien het model niet *ball fall* zal produceren (omdat dit minder expressief is dan, bv. *you fall ball* of *you drop ball*, waar telkens de agens genoemd is).

Een interessante eigenschap van computermodellen is dat we, in tegenstelling tot bij kinderen, hun interne representaties kunnen bestuderen los van het gedrag dat het model vertoont. Dit doe ik in hoofdstuk 6. Een eerste bevinding hier is dat, hoewel alle leermechanismes te allen tijd beschikbaar zijn voor het model, ze met een variabele frequentie gebruikt worden. Voor de verwerving van lexicale constructies zien we dat de naïeve methode van het cross-situationeel leren slechts in de allereerste fases gebruikt wordt, waarna het model een inventaris van half-open en open grammaticale constructies opgebouwd heeft waarmee het de betekenis van onbekende woorden kan 'bootstrappen'. De paradigmatizatie-operatie, op het grammaticale vlak, vertoont verder interessante uitbarstingen van activiteit over de ontwikkelingstijd, wat betekent dat het model niet gradueel tot nieuwe abstracties komt, maar dat het gebruiksgevallen tegenkomt die nieuwe deelruimtes van de mogelijkhedenruimte van grammaticale constructies ontsluit.

De abstracties die SPL op deze manier leert, vertonen verder de interessante eigenschap dat ze niet direct te observeren hoeven te zijn in het gedrag van het model in begrip en productie. Als we niet 'onder de motorkap' hadden gekeken, zouden we tot de onjuiste conclusie kunnen komen dat het representationele systeem hogelijk concreet is. Deze redenering klopt evenwel niet: uitgaand van de gebruiksgebaseerde stelling dat taalgebruikers in het gebruik concretere constructies de voorkeur geven boven abstractere, valt het te verwachten dat in het gebruik voornamelijk die concrete constructies zullen opduiken. Tegelijkertijd heeft het model echter een sterker potentieel om te generaliseren. Sterker nog: dit potentieel komt zeer vroeg op en het model leert daarna voornamelijk concretere constructies erbij die deze abstracties verder blokkeren. Dit is te verwachten: zodra abstracties beschikbaar zijn, zal het model deze gebruiken om expressief te zijn, tenzij het een concretere constructie beschikbaar heeft die minstens even expressief is.

Een verdere interessante eigenschap van de abstracties die we in het model aantreffen, is dat ze direct de typefrequentie van de meer concrete patronen die er in voorkomen, weerspiegelen: de transitiefconstructie, bij voorbeeld, is representationeel sterk met een open werkwoordspositie omdat er veel verschillende werkwoorden in voorkomen, terwijl de veroorzaakte-verplaatsingsconstructies met slechts twee werkwoorden gezien wordt door het model, en daarom als twee werkwoordsspecifieke constructies wordt opgeslagen.

Als we het perspectief omdraaien (kijkend vanuit de woorden in constructies i.p.v. vanuit de constructies), zien we dat sommige woorden als onafhankelijke lexicale constructies geleerd worden, terwijl andere woorden voornamelijk als constituent van een grotere constructie worden geleerd. Woorden die naar entiteiten verwijzen (zelfstandig naamwoorden) worden typisch geleerd als onafhankelijke lexicale constructies. Voor andere woordsoorten nemen we meer variatie waar, zowel tussen woorden als tussen verschillende simulatierondes. Persoonlijk voornaamwoorden worden in veel en diverse contexten gebruikt, wat zou moeten leiden tot een verwerving als on-

afhankelijke lexicale constructies, maar ze worden ook zeer frequent *binnen* die constructies gebruikt, wat zou moeten leiden tot een opslag als deel van de grotere constructie. Voor deze voornaamwoorden, maar ook voor voorzetsels en werkwoorden, vinden we dat ze in sommige simulatierondes onafhankelijk verworven worden, terwijl ze in andere als deel van een groter patroon geleerd worden. Er zijn drie factoren te identificeren voor de neiging van een woord om onafhankelijk verworven te worden. Ten eerste: hoe meer verschillende elementen er op die plek voorkomen, hoe hoger de waarschijnlijkheid dat het woord onafhankelijk geleerd wordt. Ten tweede: hoe hoger de frequentie van dat woord op specifieke posities van grammaticale constructies, hoe lager de waarschijnlijkheid dat het onafhankelijk geleerd wordt. Ten derde: hoe meer verschillende constructies er zijn waarin dat woord voorkomt, hoe hoger de waarschijnlijkheid dat het onafhankelijk geleerd wordt.

Ook op andere aspecten van de representaties vinden we individuele variatie tussen de simulatierondes. De gemiddelde abstractie van de representatie varieert ook tussen verschillende simulatierondes. Dit is interessant, aangezien de diverse simulatierondes wel hetzelfde gedrag in productie en begrip vertonen – ze presteren even goed op de verschillende taken.

Curriculum Vitæ

Barend Beekhuizen was born in The Hague on the 7th of January 1987, where he attended Gymnasium Haganum (1998-2004). After high school, Beekhuizen studied Dutch Language and Culture at Leiden University, receiving his B.A. in 2007, and subsequently the research master Linguistics at Leiden University (M.Phil., *cum laude*, in 2010). Supported by an NWO *Promoties in de Geesteswetenschappen* grant, he was a PhD candidate at Leiden University Centre for Linguistics, Leiden University and the Institute for Logic, Language, and Computation, University of Amsterdam from 2010 till 2015. Besides the PhD candidacy, Beekhuizen was appointed as a lecturer at the department of Dutch Language and Culture, Leiden University from 2013 till 2015. In the fall of 2015 he will take up a postdoctoral position at the department of Computer Science, University of Toronto.