



Universiteit
Leiden
The Netherlands

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

Beekhuizen, B.F.

Citation

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

Version: Corrected Publisher's Version

License: [Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden](#)

Downloaded from: <https://hdl.handle.net/1887/35460>

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

Cover Page



Universiteit Leiden



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

Author: Beekhuizen, Barend

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

Issue Date: 2015-09-22

CHAPTER 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

¹Reinterpreting, 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 been 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' consist 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*’]

²In 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

³Interestingly, 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 misleading 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-

⁴Of 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

⁵There 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* interchangeably, 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 WORD-ing*; (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 hitherto 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.e., 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

⁶From 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 acquisition 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.

⁷Except 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

⁸The 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 (2014b) 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 separate 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.e., 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

⁹This 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-

¹⁰Inversely, 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 the 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 from 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

¹¹The 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	<i>You open it</i>
agent-locatives	agent, verb, object, location	<i>You put it on the table</i>
mover-locatives	mover, verb, location	<i>I sit in the chair</i>
patient-locatives	object, verb, location	<i>The ball goes in the box</i>

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 puppet 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 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 construction 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 Categorical 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 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).

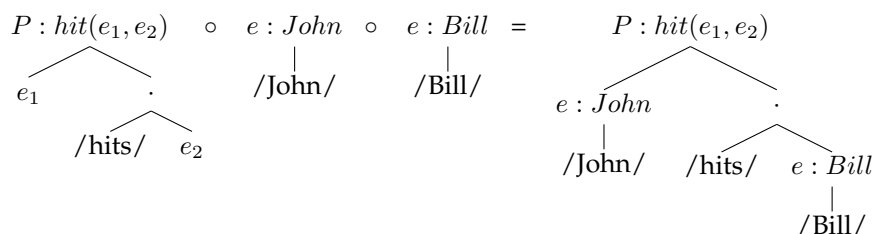
This perspective constitutes a promising starting point for developing a model that takes both form and meaning into account. The reliance of U-DOP on units of heterogeneous size and the fact that the model allows for redundant representations is congenial with usage-based constructivist tenets. The 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) [[E₁] [hits] [E₂]] | HIT(E₁,E₂)
 (21) [E:JOHN / John]
 (22) [E:BILL / Bill]
 (23) [[E₁] → [E₁:JOHN / John] [hits] [E₂] → [E₂: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 successfully.

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(E₁, E₂), which was expressed by utterances such as *Bill sees Abe* or *Mary sees Jack*, there was also an intransitive semantic predicate TURN.50(E₁) that was expressed with the expression *X sees Abe*, with any entity filling the E₁ on the

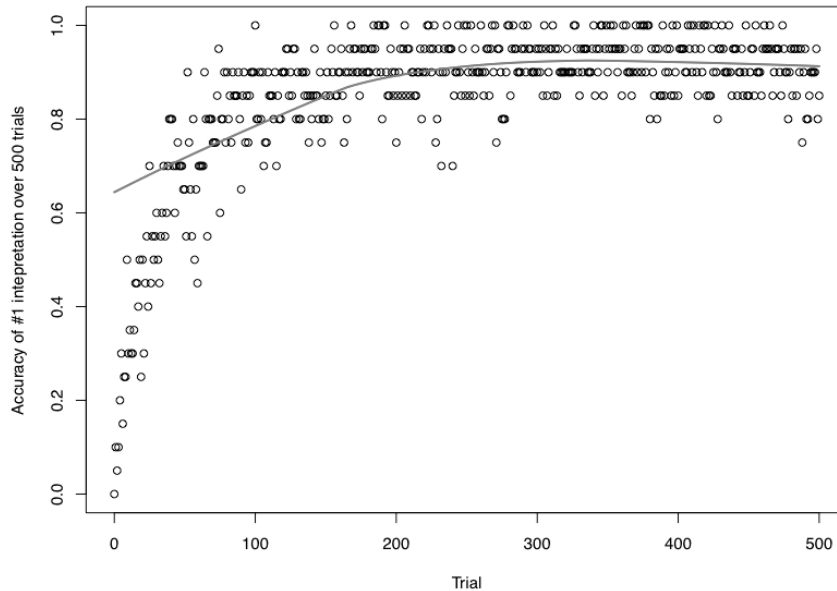


Figure 2.2: The average accuracy of μ -DOP in the first 500 trials.

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

¹²The idiom was modeled after the Dutch expression *Abraham zien*, ‘lit. seeing Abraham, turning fifty’.

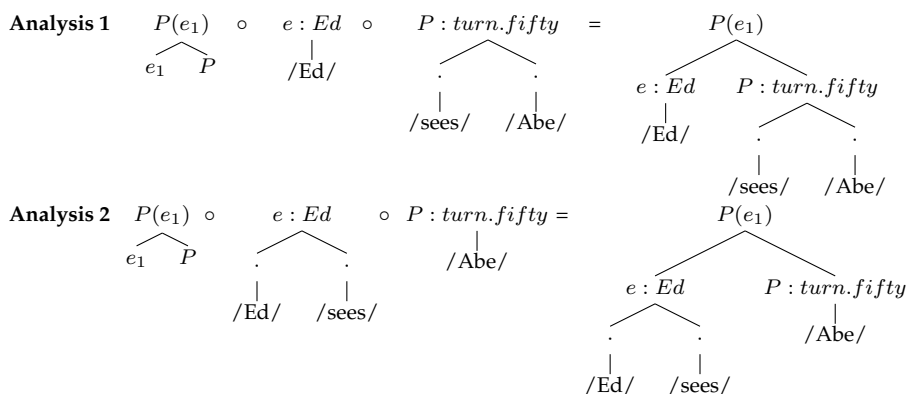


Figure 2.3: Two derivations of *Ed sees Abe* in a situation where TURN.50(ED) is present.

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 (2014a) 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 successfully used to simulate that empirical finding. A '-' means that either the model does not satisfy that desideratum, or has not successfully been shown to simulate that empirical explanandum. A '◇', 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)	(McCaulley & Christiansen 2014a)
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 network 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 acquisition.

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 MO-SAIC 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.

