Automating Construction Work
Data-Oriented Parsing and Constructivist Accounts of Language Acquisition

Barend Beekhuizen & Rens Bod

Abstract

The constructionist approach to language has long proven its merits as a theoretical framework guiding linguistic observations. However, relatively little work has been dedicated to providing a precise, formalized definition of constructions and the mechanisms by means of which they are acquired. In giving an overview of recent work in Data-Oriented Parsing (DOP), we show how the theoretical development of construction grammar and usage-based approaches to language acquisition can benefit from the converging evidence and novel insights that computational models such as DOP can provide us with.

In this chapter, we introduce DOP and compare its properties to usage-based and constructionist ideas about the nature of grammar and its acquisition. We discuss the unsupervised incarnation of DOP, U-DOP, and show how it can be used to address nativist hypotheses about the learnability of grammatical patterns. Finally, we propose an extension of the formalism that is able to learn a meaning-driven grammar from unstructured input data.

1 Introduction

Our world is filled with a vast array of objects and their relations and properties. Human infants face the magnificent task of processing experiences with the outside world in such a way that they can later on respond in an adequate manner when similar, but non-identical experiences present themselves. We can call this processing "learning" and an important question studied throughout the cognitive sciences is how humans do it. One domain for which this question is especially important, is that of linguistic systems of communication, as the complexity and open-endedness found therein has led many to believe that some architectural aspects of the cognitive representations of the phenomenon are not learnable from positive linguistic input alone. This assumption has led to the conclusion that these representations are innately present in the language learner and that there are cognitive mechanisms innately tuned or geared towards acquiring a language (such as a 'principles and parameters' approach, cf. Wunderlich (2007)). With the linguistic nativist conviction comes the assumption that the representations used are of a fairly abstract nature – after all, the learner would have to be able to acquire any of the thousands of languages being used around the world. Nativists, especially within the Minimalist framework (Chomsky 1993), further support this assumption by pointing to the economy of representation as a driving factor for having a system that is as compact as possible. Importantly, the innate knowledge is part of a mental module pertaining only to language. That is, the representations the learner starts with are domain-specific.

Another school of thought, the empiricist one, states the child does not come equipped with inborn, domain-specific knowledge concerning the architecture or properties of a communication system to be acquired. The acquisition of the complexities of linguistic structure are explained (as far as they are not theory-internal concerns that depend on one's preconception of the cognitive representation (cf. Tomasello (2003, p. 7)) from experience, through domain-general structure-finding mechanisms such as categorization, schematization and social understanding. Importantly, these mechanisms and representational biases have to exist in the learner's mind prior to the acquisition of a language system. Hence, usage-based theorists cannot be argued to believe in a blank-slate learner. The crucial difference from a *linguistic* nativist position is that the mechanisms and biases are not specific to language, but are shared with other cognitive domains because they either are functions of how the brain in general works (e.g., working memory, entrenchment processes, abstraction) or are part of known evolved cognitive modules (e.g., the figure-ground distinction from the visual system, notions of object permanence). With the nativist position being the dominant one for the last decades, researchers of the empiricist bent face the task of showing that there are flaws in the empirical observations or subsequent inferential processes leading to linguistic nativist conclusions. At the same time, it is crucial that empiricist theorists develop a substitutive, positive, account of language acquisition through experience and domain-general skills. Important work showing flaws in nativist reasoning and providing a novel account has been done. Construction Grammar, in many of its flavors (Langacker 1989, Goldberg 1995, Croft & Cruse 2004), as well as non-constructivist work in language acquisition (Peters 1983) shows how the nativists' assumed divisions between the core and periphery of the grammar, meaning and the grammar, and linguistic competence and performance cannot be maintained, and at the same time presents an empiricist account of how the architecture and content of linguistic representation emerges as an interaction between a multitude of factors. The work of Tomasello and colleagues (Tomasello 2003) has shown how understanding other people's (communicative) intentions is crucial for and supportive of acquiring a language, demonstrating how a thitherto overlooked aspect of human cognition solves some of the nativist arguments against acquiring a grammar from experience, as well as presenting a coherent explanation of linguistic development.

In this paper, we would like to add something to the developing usage-based constructivist narrative. This contribution is in part a methodological enrichment and in part an account of the possible domain-general cognitive mechanisms behind the acquisition of the grammatical structures. We believe that computational modeling is an important means for providing us with important insights in the theoretical perspective. First of all, it forces us to translate our fuzzy and imprecise natural-language definitions into extremely precise computational ones. Although this often means a loss in accuracy of description (we will have to give up on the description of some aspects of natural language for our model to be understandable), it provides a gain in the testability of certain claims. Using a well-defined model, then, we can assess claims pertaining to the architecture and content of the representations, the processing mechanisms and the timescales on which these operate.

The computational model we present in this paper is Data-Oriented Parsing (DOP; (Scha 1990, Bod 1998, Bod, Scha & Sima'an 2003)), and its instantiations Unsupervised Data-Oriented Parsing (U-DOP) and Meaningful Unsupervised Data-Oriented Parsing (μ-DOP). The data-oriented family of models addresses the question how processing complex, structured exemplars, such as linguistic experiences may lead to a cognitive system by means of which a language user can assign structure to novel exemplars. As such, it is not a theory about the content of representations, but rather a discovery procedure (for

learners and linguists alike) for cognitively useful structured representations. In the following sections, we explain the basic ideas behind the models in greater detail, link it to constructivist assumptions and show how the diverse models can be applied to questions about the acquisition of grammar.

2 Data-Oriented Parsing

A core question in the usage-based approach to language acquisition is that of grammatical productivity. How does a learner, be it an artificial one or one of flesh and blood, know, after having seen a number of exemplars, what patterns it should use to produce and interpret novel utterances? Although many informal discussions of the process have been given (Tomasello 2003, Goldberg 2006), often with reference to Gentner's more formalized work on analogy (Gentner 1983), few models of discovering the productive grammatical units of a language have been developed so far. Similarly, no existing description of construction grammar's parsing principles offers us an account of recombining these productive fragments into analyses of novel utterances. Such an account is desirable, as it can help validate learnability claims and adds to the possibilities for evaluating the theory against the data. It should be noted that construction grammar and usage-based theories are not alone in their lack of precise definitions; it seems that any current linguistic theory has given up on the construction of a precise, testable model of language use and language acquisition.

Formalizations of learning mechanisms for acquiring a grammar such as Embodied Construction Grammar (Chang 2008) and Fluid Construction Grammar (Trijp, Steels, Beuls & Wellens 2009) have been developed over the last decade. Other systems that have claimed relevance to usage-based theorizing are Memory-Based Learning (Daelemans & Van den Bosch 2005) and the memory-access and parsing framework developed by (Jurafsky 1996). All of these add to our understanding, and insight from these different approaches complements DOP's contribution, namely a precise account of how Gestalt-like linguistic units can be discovered in the data. The proposed mechanisms of Data-Oriented Parsing obviously cannot capture the wealth of linguistic phenomena described in full detail, but aim to give us insight in how complex representations can be acquired from the input data, and as such can help understand the domain-general learning processes in want of further specification.

2.1 Data-Oriented Parsing as a constructional learner

Suppose a learner has processed several exemplars, or structured representations of utterances. When processing a novel utterances, the learner can draw on this inventory by recombining its parts in order to come up with an analysis of the novel utterance.

Now, a novel utterance, say the one in example (1), can be analyzed using parts of the processed utterances in figure 1. Let us define a legal part, or *subtree* of a tree representation to be a connected subgraph in which all sisterhood relations of the original tree hold (see Bod & Kaplan (1998) for a more precise definition). Maintaining the sisterhood means that given the first tree in figure 1, we can have a fragment of S going to NP and VP, but not S going just to NP (without its VP sister). But most importantly, it is not just the small parts that can be re-used, larger fragments can be used in analyzing novel utterances as well. The main claim of Data-Oriented Parsing is that all fragments, irrespective of size, can be used in analyzing novel utterances. Given this starting point, we have many ways of analyzing sentence (1). Some of these are given in figure 2. We analyze novel

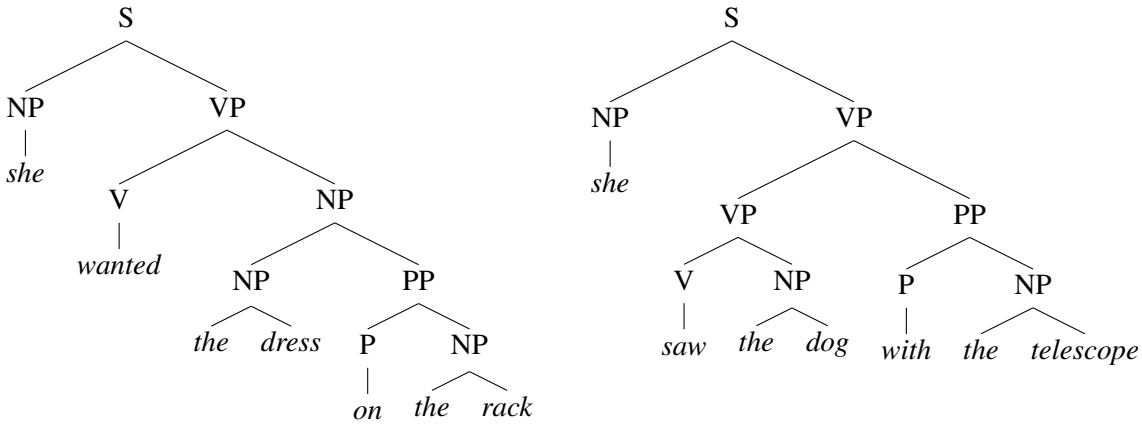

Figure 1: A corpus of two processed utterances

utterances by using a legal part of an utterance the learner has seen and 'subsitituting' its leftmost non-terminal symbol (that is: grammatical symbols, such as 'S', 'Det' and 'PP') with another part the learner has already seen. This procedure is repeated until there are no non-terminal symbols left, that is: all words of the utterance are present in the analysis.[1] The symbol of this substitution operation is '$\circ$'.

(1) *She saw the dress with the telescope*

What we see in figure 2 is that there are multiple analyses possible for the novel utterance, and there are multiple ways to arrive at a single analysis. We call each of these ways of arriving at an analysis a *derivation* and a resulting analysis (which can emerge through different derivations) a *parse*. The first and third derivation thus give us the same parse tree, but get there in different ways. The first tree uses only the smallest fragments possible, while the third tree re-uses larger fragments of the earlier processed experiences.

Now, given this structural ambiguity (traditionally: the PP-attachment problem), the learner has to choose which of the possible analyses to consider as the right one. This is where the frequencies of the fragments come in. First, consider a derivation to be a complex event consisting of a number of smaller events, viz. the subtrees. Suppose that each of these subtrees has a certain probability. This would mean that the probability of the event of them occurring together would be the joint probability of all of the individual events of selecting that subtree. The joint probability of a derivation can thus be given by the product of the probabilities of the individual subtrees $t_1 \dots t_n$ that make up the derivation d:

$$P(d) = P(t_1 \circ t_2 \circ \dots \circ t_n) = \prod_{i=1}^{n} P(t_i)$$

The probability of a subtree, then, can be estimated by the number of times it occurs in the corpus of processed utterances, divided by the number of times a fragment is found with the same syntactic category at the root of that subtree. This is to say that the event of drawing a specific subtree from a bag of subtrees with the same syntactic label in the root

[1] This procedure may strike some readers as very top-down. Although it is presented as such, the same principles can be applied in a bottom-up parsing algorithm equally well.

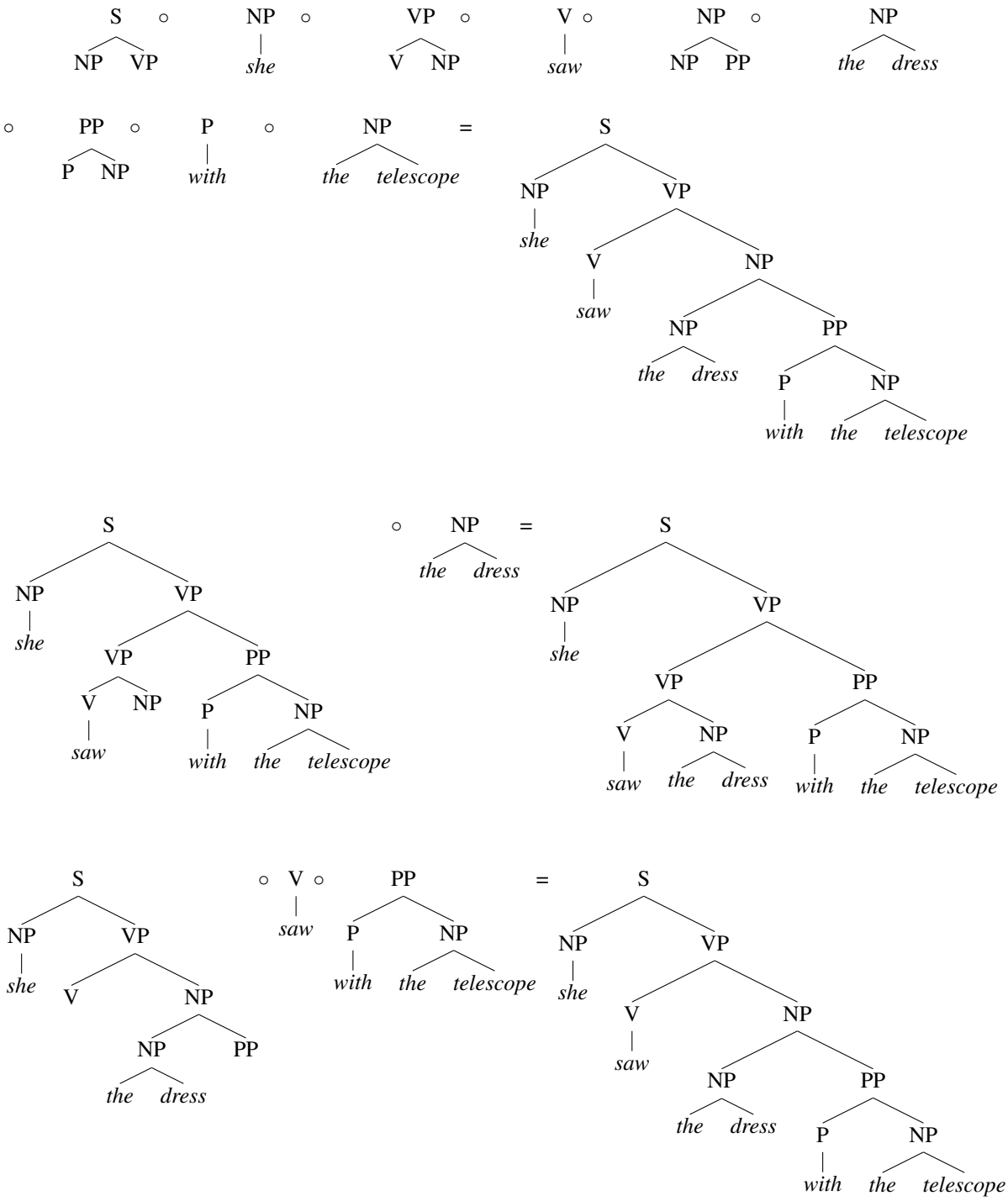

Figure 2: Three derivations of *She saw the dress with the telescope*

node is the number of times that subtree occurs divided by the number of subtrees in the bag. With this estimation procedure, the subtree *competes* with all other subtrees that can occur at the same place in a derivation, viz. at an open position in another fragment that

has the syntactic category of that subtree. Stated more formally:

$$P(t) = \frac{|t|}{\sum_{t':root(t')=root(t)} |t'|}$$

The estimation of these probabilities thus involves finding all possible subtree types in all trees in the corpus and establishing their frequency. For a simple tree such as the one in figure 3, we can extract all possible subtrees and arrive at the set shown in figure 4.

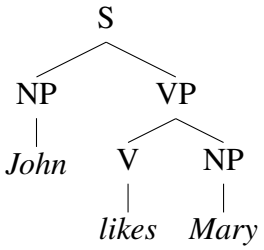

Figure 3: A simple parse tree

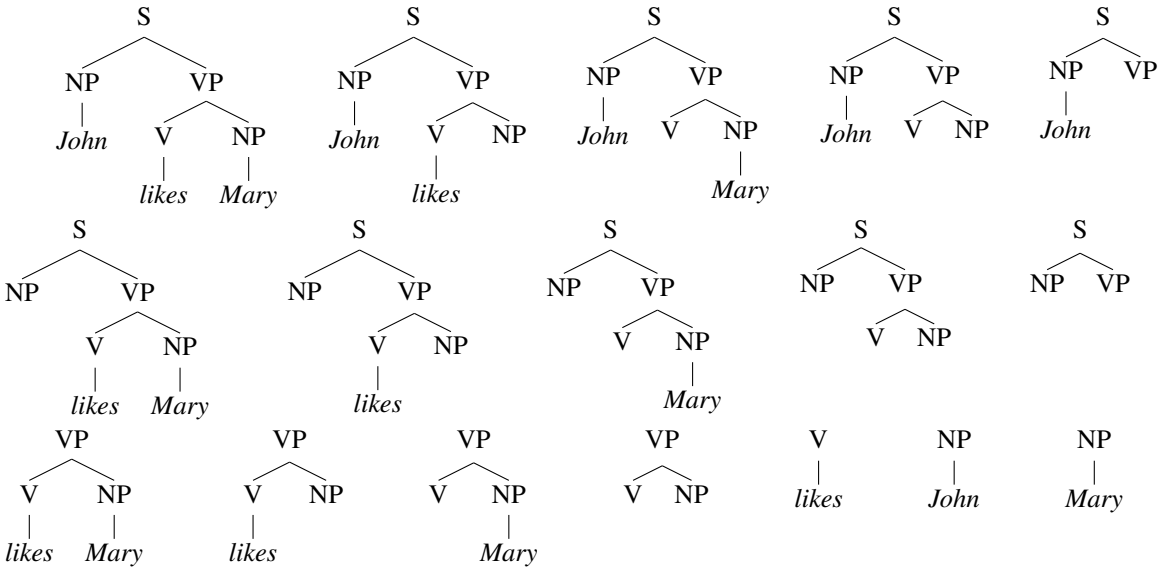

Figure 4: All subtrees of the tree in figure 3

Now, we have multiple derivations leading to the same parse of an utterance, as we have seen in figure 2. We determine the probability of a parse to be the sum of the probabilities of all derivations yielding that parse tree. Again, this is grounded in relatively simple probability theory: the probability of a parse is the probability of either derivation one leading to that parse, or derivation two, or derivation three, and so forth, and so it is the sum of the probabilities of the individual events. For any parse tree t we can thus calculate the probability as follows:

$$P(t) = \sum_{d \text{ is a derivation of } t} P(d)$$

We then select the parse with the highest probability mass to be the most likely analysis of the novel utterance (cf. Zuidema (2006) for a discussion of this and other estimation and evaluation methods).

It is important to note that, using the same mechanisms, DOP can also *generate* utterances from the grammar. This process boils down to probabilistically selecting a subtree fragment rooted in the starting label, and expanding its open substitution sited by other subtrees. When coupled with meaning, this approach can give us a Data-Oriented generator as well.

2.2 DOP as a usage-based, constructionist model

DOP shares most of its core principles with the usage-based constructionist approach. In this section, we discuss on what theoretical positions DOP and the usage-based constructionist approach converge.[2]

An important effect of this modeling procedure is that the most likely parses will be the ones that have derivations consisting of a few larger subtrees. Why is this? The probability of a derivation will generally be higher if it consists of fewer subtrees because there are fewer subtree probabilities to multiply. Hence, the model has a bias towards interpreting utterances by using as few and hence as large fragments as possible. In that sense, the model tries to maximize analogy with the previously processed utterances and by doing so, the model adheres to the usage-based principles that grammatical productivity comes about through experience and a domain-general ability to make schemas (Tomasello 2003, Gentner 1983).

The reliance on experience is another aspect on which usage-based constructionist approaches and DOP converge. First of all, the hypothesis space of possible grammatical constructions emerges through the experience with language, as well as the conception that we have to understand language, at least to some extent, hierarchically (see Frank, Bod & Christiansen (2012) for arguments why hierarchical processing is not a procedure applied all the time). Hierarchicality, then, does not have to be the a priori template for a learner to understand language. The learner may start with a number of possible data structures, some of which are hierarchical and some are not, and find out, in response to processing the data that a hierarchical template to store, process and produce language may be an optimal cognitive strategy (Perfors, Tenenbaum & Wonnacott 2010). For this paper, we assume that this property of language has been discovered.

Secondly, experience means that routinization and Gestalt-like effects take place (Bybee 2006). It is well known that frequency affects language use, at the very least by governing choice among acceptable alternatives (Schuchardt 1885, Mehler & Carey 1968, Jurafsky 2003). DOP incorporates this insight by allowing for larger fragments to be stored and used as Gestalts in linguistic processing. Moreover, the trade-off between computation (composing two fragments by the substitution operator) and storage (using one larger fragments) is driven by frequency as well: the more likely the parts are relative to the whole, the more likely a computed analysis (as opposed to a retrieved one) is. But most importantly: it does not have to be either/or. Because all derivations, whether they are directly retrieved as one chunk, or composed of minimal bits, are used in calculating the probability of the analysis, DOP avoids the rule-list fallacy (Langacker 1989, ch. 1): language users maintain both, sometimes perhaps redundantly.

Furthermore, DOP starts from the same maximalist conception of language as constructionist approaches do. This conception entails a couple of things. First, the basic building blocks are heterogeneous in size. This means that they can be small, like words

[2]For a more thorough discussion of the principles of experience, heterogeneity and redundancy, see Beekhuizen, Bod & Zuidema (2013)

or depth-one rules, or larger. And it means that they can be abstract, having no lexical material, or highly concrete. An important insight following from this principle and the previous one is that rules and exemplars are not ontologically different entities, but are created out of the same matter, viz. processed experience. Every subtree in DOP then, is a schema from the processed experience that can be recombined with parts of other experiences to understand something novel. These ideas resonate core properties of a constructivist, usage-based understanding of grammatical knowledge (Croft & Cruse 2004, ch. 10-11).

Finally, the inventory of the basic building blocks may be redundant, as hinted at earlier. DOP gives the artificial learner fragments which it can, in principle, build up out of other subtrees it has. The idiom *What time is it?* can of course be built up out of its components, but there is reason to believe that language users keep a representation of the whole in mind as well (Bybee 2006).[3] Although this is not a position shared by all constructionist linguists (Construction Grammar tries to minimize redundancy for instance (Fillmore & Kay 1996)), the usage-based theorists seem to embrace this idea. Accepting redundancy as a core property of the linguistic system follows rather naturally from the rejection that linguistic structure has to be either stored as a rule or as a list (i.e., the rule-list fallacy, cf. Langacker (1989))

In fact, the DOP framework has been used to address issues in language acquisition that relate to the issues of heterogeneity and redundancy. Given a hypothesis space of all possible subtrees, we can find out what set of subtrees was most likely used in deriving an utterance. Without going into the details, Borensztajn, Zuidema & Bod (2008) did so for a syntactically annotated corpus of young children's utterances. What they showed was that, in line with the usage-based perspective, the most likely subtrees behind the children's utterances become more abstract with age. More examples of applying the DOP principle to language acquisition can be seen in the next sections.

2.3 Analogy, acquisition and the unlearnable

If we think about DOP as a model of language acquisition, the model effectively says that children acquire grammar by constructing analogies with previous utterances, guided by statistical generalization. This starting point, of using analogy to construct a grammar, has not gone unchallenged in linguistic theorizing. Examples of linearly similar but structurally different sentences, as discussed by Pinker (1979) and Chomsky (1986), show how proportional analogy, the simplest format of analogical reasoning, might lead a learner to wrong conclusions (this particular example being Pinker's):

(2) John likes fish : John likes chicken :: John might fish : John might chicken

(3) Swimming in the sea is dangerous : The sea is dangerous :: Swimming in the rivers is dangerous : The rivers is dangerous

As Pinker and Chomsky correctly point out, analogies like these do not hold because there is no notion of structural dependency, nor a concept of syntactic category that would be required to make them work. However, this is not a problem of analogical reasoning per se, but rather of the structure and content of the input analogical reasoning applies to. Analogical reasoning, which can be described as trying to solve a problem (categorizing an object, parsing an utterance) by comparing our knowledge about the object to a knowledge base of similar objects. If we grant people the ability to infer grammatical categories and

[3] That is, regardless of whether this whole is opaque to the user or whether its constituency is transparent.

hierarchical representations for sentences, analogical reasoning over such a knowledge based would not come up with the erroneous predictions of the grammaticality of *John might chicken*. If we let the the learner make the analysis without anything like grammatical categories or a notion of hierarchical structure, we do arrive at this prediction. Hence, it is not the mechanism that yields ungrammatical results, it is the nature of the content.

An extension of the original DOP model presented in the previous sections, Unsupervised Data-Oriented Parsing, or U-DOP, has been developed as an attempt to address this issue. Unsupervised techniques, developed in machine learning, allow a learner to build up some representation (of structure or categories, for instance), without having 'correct' representations for a set of training items (which would be supervised learning). Instantiated in U-DOP, these techniques grant the learner the domain-general starting point of understanding data as hierarchically structured, that is, as containing different levels of analysis, in which a concept on one level is *triggered by* (communicatively, cf. Verhagen (2009)) or *consists of* (cognitively) small, less inclusive parts, but do not give the learner the correct analyses of the structure to train on. These assumptions are not language-specific, as we can apply the template of meronymy (the *consists-of* relation) to our understanding of body parts, artefacts, grouping relations of identical individuals, only in language we combine it with symbolic understanding (the *triggered-by* relation).

Using the idea that language is hierarchical and a more strict notion of analogical reasoning, U-DOP can be shown to predict the ungrammaticality of *The rivers is dangerous* (Bod 2009). It also predicts that a child can learn that, if it wants to form a polar interrogative of a sentence like (4), it's not the first (as in example (5)) but the second *is* (as in example (6) that is produced at the front of the utterance.

(4) The man who is sick is singing.

(5) *Is the man who sick is singing?

(6) Is the man who is sick singing?

How does the model acquire these dependencies correctly? Unlike DOP, U-DOP assumes that the learner does not know how to interpret its initial input. Instead, the learner will store all possible analyses of the input, and uses that as a basis for extracting subtrees and estimating their probabilities. Furthermore, we leave the problem of syntactic categories out of scope for now, focussing solely on the hierarchical structure. Effectively, all nodes in the tree representation, except for the lexical leafs, are of the same category, say 'X', and can thus be substituted for one another with the combination operation. So, suppose the U-DOP learner has heard the two utterances *the dog walks* and *watch the dog*. Each of these has two possible analyses:

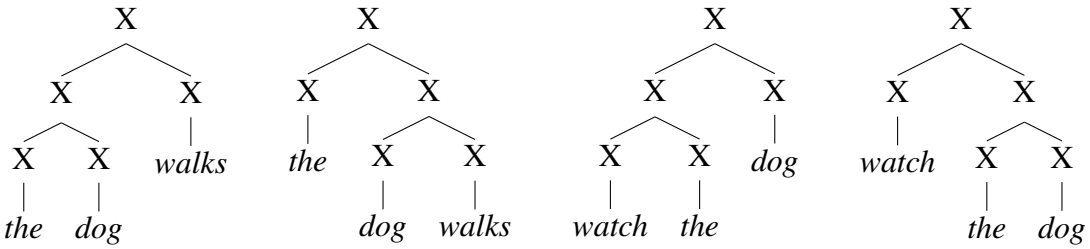

Figure 5: The analysis of the two utterances *the dog walks* and *watch the dog*

From this collection of all possible analyses, we can extract all possible subtrees, just like we did with DOP. What will become immediately clear, is that a subtree like [[*the*]$_X$

[*dog*]$_X$]$_X$ forms a reliable constituent, being found in two out of four parse trees. A less reliable subtree is [[*dog*]$_X$ [*walks*]$_X$]$_X$, which is only found in one parse tree. Using these subtrees, then, we can infer the hierarchical structure of an unseen utterance.

In order to do this, we use a stricter notion of analogy than in DOP. U-DOP starts from the insight that the more similar a novel analysis is to earlier analyses, the better an analysis it is. The model will therefore choose in the first place that parse tree that has the shortest derivation. We consider the length of the derivation to be the number of subtrees used in that derivation. Often, there are multiple parse trees that have an equally long derivation. In that case, the learner selects the most probable parse among the ones that have the shortest derivations. The probability of the parse is calculated as in DOP. This idea of selecting the most probable parse from among the shortest-derivations (MPSD) is in essence a probability driven model of analogy.

When we train this model on child-directed speech, we can simulate the acquisition of hierarchical structure and grammatical dependency. If we train U-DOP on the Adam corpus (Brown 1973), which consists of two hours of child directed speech per fortnight over the course of approximately two years, which constitutes only a fraction of a child's input, the model correctly assigns more probability to a sentence like 6 than to one with the wrong auxiliary at the sentence-initial position 5 (Bod & Smets 2012)

Why is this of interest? The issue of auxiliary-fronting with subjects that have relative clauses has been a parade case for nativist approaches to grammar. Crain (1991) and others have tried to argue that the fact that children make no errors like the one in (5) when learning these patterns shows that they directly home in on the correct hypothesis, i.e. that there is a main clause and a subordinate clause, and that the auxiliary in main clause ought to be fronted. U-DOP uses no concepts of 'clause' to explain the phenomenon, but grounds it in the experience of a learner and its attempt to stick as close as possible to that experience. A typical nativist argument against the use of experience is that this specific construction is rarely, if ever, observed in the primary data, yet children seem be sensitive to the difference in grammaticality between examples (6) and (5). U-DOP tackles this problem by saying that we can combine subtrees from different processed utterances. A learner may have never seen a case of auxiliary fronting with a subject containing a relative clause, it will probably have processed some auxiliary fronting *without* subjects with a relative clause as well as some relative clauses in other grammatical constellations.

Subtrees from these parses make the learner able to produce a more probable short derivation for sentences like example (6), but not for cases like (5). A pattern like *the X who is X* might be used, along with *is X singing*. However, the model will have never or very scarcely seen patterns like *who sick* or *is X is singing*. Because of this, the model with find more probable short derivations for the good pattern, and longer, or less likely derivations for the erroneous one. The observed behavior of the model is in line with the pattern of errors observed in Ambridge, Rowland & Pine (2008)

Moreover, this manner of analyzing the learnability of complex grammatical phenomena can be extended to other cases. Whereas most empiricist computational linguists use a specific model to address a specific phenomenon in order to refute nativist explanations (e.g., auxiliary fronting (Clark & Eyraud 2006) or anaphoric *one* (Foraker, Regier, Khetarpal, Perfors & Tenenbaum 2009)),[4] Bod & Smets (2012) show that a single, unified model, viz. U-DOP, can learn (virtually) all existing cases of hierarchical dependencies that are thought of as unlearnable. Work like this complements the analysis done by Pullum & Scholz (2002) and show how general learning and structuring principles may lead

[4]See Clark & Lappin (2011) for an overview of different refutations using different models

to the behavior or judgements we can observe. As such, they provide us with a cognitively leaner, simpler and hence a priori more likely model of the acquisition of grammatical structure.

3 Meaning

One crucial aspect of constructional approaches has been ostensibly lacking from the discussion so far: meaning. Constructional theories hold that the grammatical building blocks are pairings of some signifying form with a signified meaning. Although much work in DOP has been done on grammatical form per se, the model is not incompatible with this approach to grammar. In fact, the model has no restriction on the representation it processes, as long as they are well-formed according to some formal criterium. This follows from the claim that DOP is a domain-general learner; as such it has to be able to detect structure regardless of the topology or content of the structure.

DOP has a long history in trying to accommodate meaningful representations. (Bonnema, Bod & Scha 1997) can be seen as a first attempt. In this model, the syntactic representations on the tree's nodes were enriched with lambda-calculus logical formulae. Later developments were the integration of DOP and Lexical-Functional Grammar (LFG-DOP, (Bod & Kaplan 1998) and Head-Driven Phrase Structure Grammar (HPSG-DOP, (Arnold & Linardaki 2007)). Building on the insights of these models, we propose an unsupervised variant of Data-Oriented Parsing that incorporates meaning. Because this is the first exploration of an unsupervised learning mechanism to meaning-enriched structures, we chose not to use the rich representations of LFG or HPSG, but rather take a very simple and limited formalism to illustrate the priniciple. We will show how it functions, how a learner may derive productive patterns with it, and what its limitations are.

3.1 A U-DOP approach to learning meaningful grammars

What would acquiring a grammar involve, if we use the constructionist starting point of form-meaning pairings as the basic building blocks of a language? First of all, the problem of learning the mapping has become bigger, as not only word-meaning mappings have to be learned, but also mappings between all other kinds of constructions and meanings. Secondly, the learner would have to have some mechanism for arriving at a set of schemas, productive and less so, that allow it to talk about novel events and understand novel utterances. Thirdly, the model would have to be incremental: we cannot expect a real language learner to wait until it has seen some number of utterances. As we will see, incrementality in fact does not make the problem bigger, but rather can function as a bootstrap for the learner.

The basic idea behind meaningful U-DOP, or μ-DOP, is that a learner analyzes a sentence using meaningful, heterogeneously sized and possibly redundant hierarchical representations. It starts analyzing utterances with the fragments it knows already, and then maps the unanalyzed parts of the utterance with parts of the semantic situation that are unexpressed in the analysis, thereby filling the gaps in its knowledge. Next, it decomposes these analyses and adds them to its knowledge base.

In this experiment, we describe a learner that can deal with very simple meaning representations. For understanding more complex semantic operations, obviously more complex representations are needed. The goal, however, is not to develop an account of meaning, but rather to show how DOP can acquire symbolic structures (i.e., form-meaning pairings)

of different size, complexity and abstractness. By doing so, we show how the concept of building up meaningful constructions works with an unsupervised Data-Oriented model and how the acquired representations are in line with the constructivist understanding of language. As such, we argue that DOP is a substrate-neutral learner (as long as the substrate can be structured in some graphical way), that can be thought of as domain-general.

How can we envisage the representation of an analysis in μ-DOP on which the data-oriented learning algorithm applies? It has to contain hierarchical structures, meaning representations and represent words. As a simple formalism for representing meaning, we use predicate logic, restricted to representations of predicates and entities. The semantic representations take the place of the syntactic categories of DOP as the contents of the nodes in the analyses. A node may contain a value of a variable (predicate or entity), denoted as $P : watch$ for the predicate $watch$ or $e : John$ for the entity $John$, or denote an operation on its constituent parts. $watch(e_1, Mary)$, for instance, means that there is some entity that fills the slot position e_1 that can be found somewhere else in the subtree. Fragments like these can be combined by logical composition. Suppose we have the fragments in figure 6. We can compose these and the interpretation will be $hit(John, Bill)$, as the slots with which the two small, word-like subtrees are combined, dictate which entity is the first argument (e_1) and which the second (e_2). Not all (legal) fragments can be composed: the subtrees in figure 8 cannot be composed, as there is no slot in the subtree to the left to fit $P : hit$ to: both open positions are of the category e.

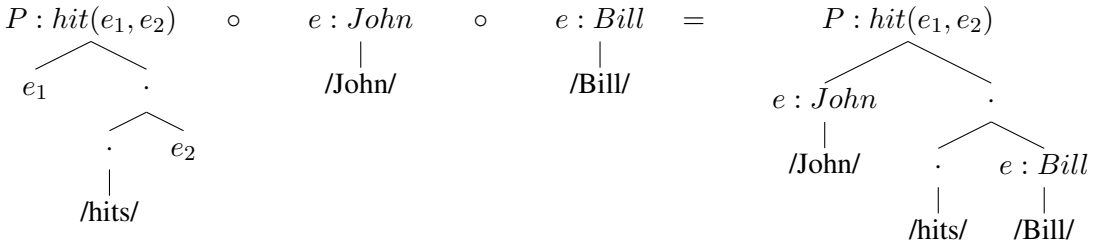

Figure 6: Legal composition of three subtrees. The empty nodes (visualized with · as the node label) are necessary to preserve binarity, but do not have any analytical significance. The character before the colon in the top-node of a subtree represents the semantic category of the subtree.

Let us look at the learning procedure taking us to an inventory of such subtrees step by step. First, the learner analyzes a sentence using its inventory of grammatical patterns. Initially, this inventory is empty, but as the learner processes more utterances, it will become bigger. Suppose for the sake of the argument that the learner knows that the word /John/ means that there is an entity called John, or $e : John$. This can be represented as a simple subtree, connecting a node $e : John$ with a node /John/. Suppose the learner also knows that the word /Mary/ refers to an entity called Mary.

The learner hears an utterance like *John watches Mary*, uttered in the context of John watching Mary and Mary walking. What the learner will try to do, then, is analyze this utterance using its known constructions, i.e. the word-meaning pairing for /John/ and for /Mary/. If the learner does not know how to analyze the utterance, it can use a meaningless binary subtree (with a very small probability) to combine any two fragments. Analyzing *John watches Mary* we arrive at two analyses, as follows.

In both parse trees, we arrive at an interpretation that involves at least an entity called John and one called Mary. When trying to understand what situation the utterance refers to, the learner can now exclude that of Mary walking, as it has seen a known fragment

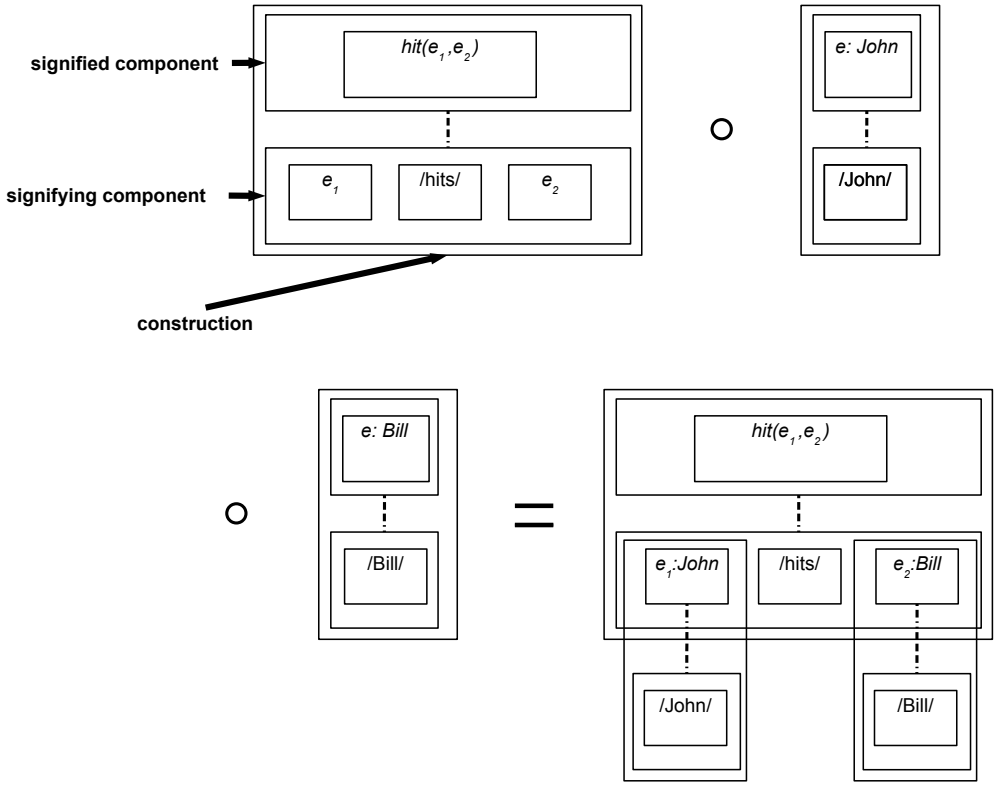

Figure 7: The composition of figure 6, represented in box diagrammes. Note that the two poles of the construction correspond to Verhagen's (2009) conception of the construction as a symbolic assembly

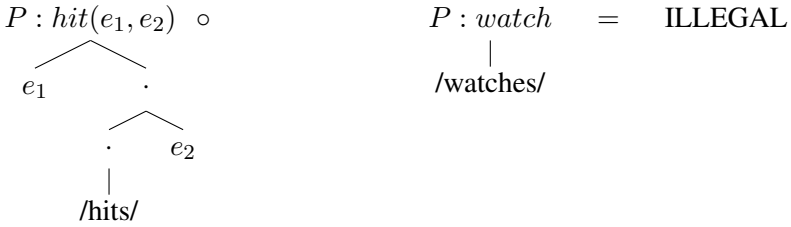

Figure 8: Illegal composition of two subtrees

Figure 9: The starting knowledge of the model in the example

that refers to John, who is not a participant in the situation of Mary walking. The learner will then try to complete the analysis by adding parts of the semantic representation to the analyses. Because the learner does not know to what nodes of the analyses these

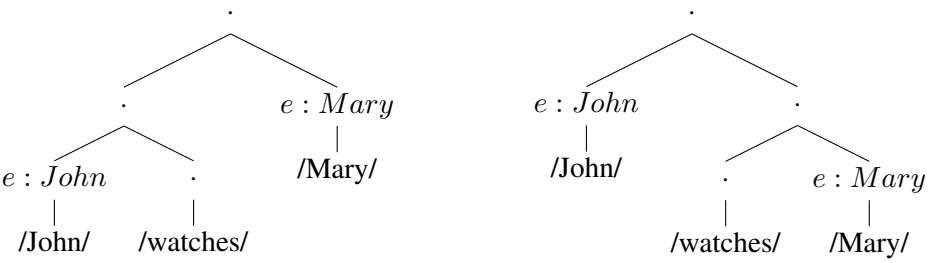

Figure 10: Two analyses of *John watches Mary*

representations belong, it will try all possible combinations and store them in its memory. The distribution of unseen meanings is guided by the constraint that the same fragment of meaning does not occur twice in the utterance. This constraint can be seen as a form of a mutual exclusivity constraint or as more simple Gricean pragmatic reasoning (i.e. balancing the maxims of quality and quantity).

To what observed situations can the partial analyses in figure 10 apply? Recall that two situations take place, viz. John watching Mary and Mary walking. The second can be excluded, as John, who was found to be referred to in both analyses, is no part of it. This leaves us the first one, and the model will try to complete its analyses using parts of that meaning. If there were more situations compatible with the inferred meaning in an analysis, all of them would be used in the add-unseen-meaning step to complete that analysis.

In our example, we have two analyses, both of which can be mapped to the situation $watch(John, Mary)$. In both cases, the learner has found the meaning $e : John$ and $e : Mary$ and hence it is missing the part $watch(e_1, e_2)$. This partial semantic representation can be further decomposed into $P(e_1, e_2)$ and $P : watch$ (in general: the models tries all decompositions of the unobserved parts of the meaning) and these partial representations can be distributed over the nodes in the tree that were unanalyzed according to the constraint discussed before. This is effectively a U-DOP approach to semantic representations: we take all parts of all missing semantic representations, all nodes in the tree analyses of an utterance and map them to each other, letting the statistics decide which ones would be relevant for analyzing novel utterances. For the first analysis of *John watches Mary* and the representations we found missing, we can add them to the partial analysis in the ways given in figure 11.

Using these completed representations, we can update our inventory of subtrees that can be reused in analyzing new utterances. We define DOP's decomposition operation here to take all connected subgraphs of the parse tree that have both a root node with a semantic representation, and only leaf nodes with either semantic or phonological representations in them. That is, the two subtrees in figure 13 are excluded on these grounds (not a meaningful root, not all leaf nodes are meaningful), whereas all legal subtrees of the third completed analysis in figure 11 are given in figure 12. When decomposing a tree, leaf nodes retain only the type and its index, so that an indexed slot comes into existence. The non-trivial part of the decomposition operation is that a tree can be decomposed only at nodes where a meaning representation is found; nodes lacking these are taken to be 'internal' (signified with a dot '·') and are not used in the interpretation process in any way, but are retained only to maintain (the positive computational properties of) binarity.

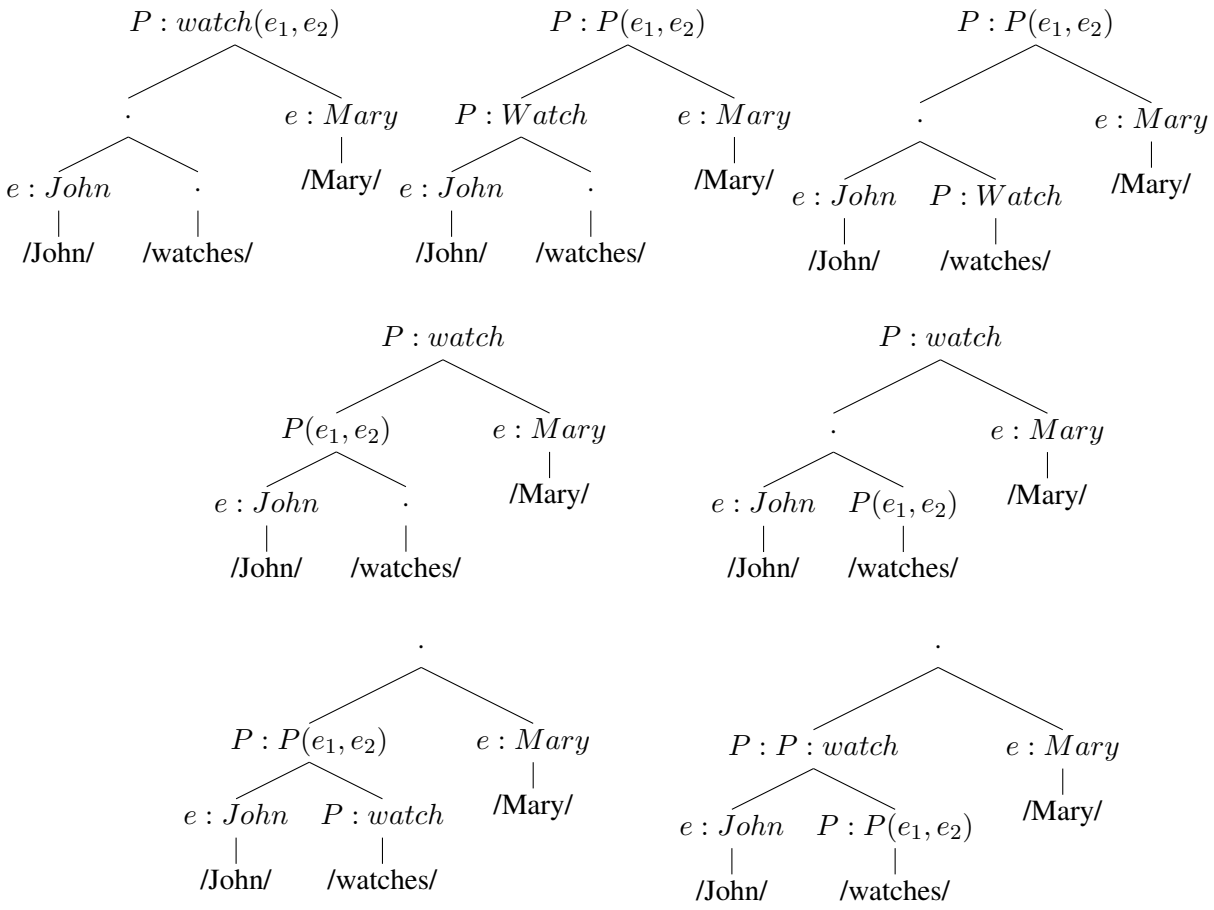

Figure 11: All additions of unseen meaning to the first analysis in figure 10

Among these decompositions, we can find various structures that remind us of constructions as we know them. There are words, fully open patterns, ones with only lexical items as leaf nodes and everything in between. All of these have the same ingredients: a graph structuring the part-whole relations with phonological or semantic structure as the content of the nodes. Our procedure, of trying to parse an utterance, adding the unseen parts of the meaning to unanalyzed parts of the utterance and decomposing that thus provides a learner with a way of discovering the patterns that are useful for understanding novel language material.

However, in a more realistic setting, there may be many analyses that are incorrect and thus harm the model if their parts are added to the inventory of fragments. Do we add all of these to the knowledge base for interpreting the next utterance too? In fact, we don't have to, as DOP provides us with a good mechanism to separate the wheat from the chaff, namely the probabilities of the analyses.

As in DOP, the most likely analysis of an utterance is the analysis that has the highest probability mass summed over all of its derivations. The probability mass of a derivation, then, is given by the joint probability of all the subtrees used. Since the model has to add unseen rules, we have to reserve some small probability for these. We do so by smoothing the probability of the seen rules, so that some probability mass becomes available for the

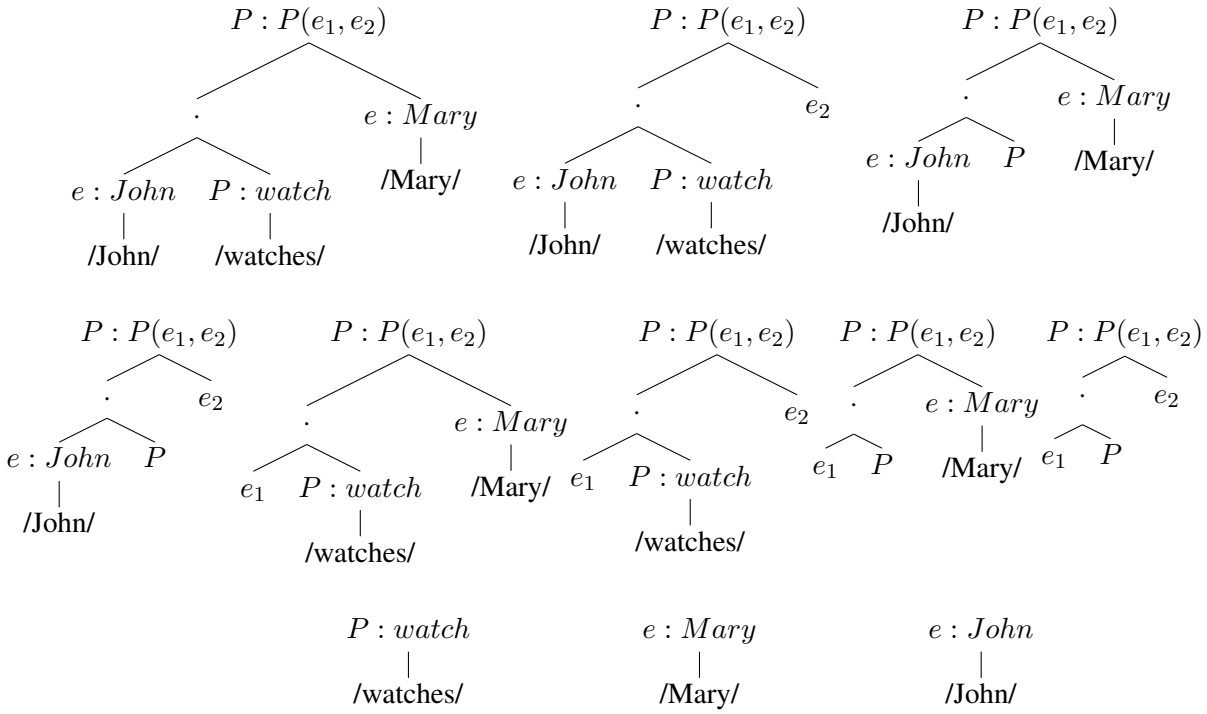

Figure 12: Legal subtrees of the third analysis in figure 11

Figure 13: Illegal subtrees of the third analysis in figure 11

unseen components.[5]

We weight all parse trees by their probability before decomposing them, so that the subtrees from the more likely analyses will be more likely to be reused in subsequent analyses. The subtrees in figure 12 thus are not added to the inventory of patterns with a frequency of one, but with a frequency of one *times the probability of the derivation*. After all subtrees are added, we can recalculate the probability of every subtree in the same way as we did with DOP, viz. by dividing its (weighted) frequency by the summed (weighted) frequencies of all subtrees sharing the exact content of the root node, i.e. the semantic representation of it.

Expanding the DOP-idea to another representation, we can see what Data-Oriented Parsing can and cannot do, as well as what the assumptions of the model are. The model is not a category learner: semantic structures and their combinatoriality, and sound structures are assumed as given information to the model. The power of DOP is then to recognize productive, complex patterns in the structured data given these assumptions. To this end, the model tries all possible structures (the unsupervised component) as guided by the context, but uses this information judiciously by weighing it according to the model's prior knowledge. The main assumptions of the model are in the construction of likely analyses,

[5]More precisely, we used Simple Good-Turing smoothing (Gale & Sampson 1995).

which have been argued to be driven by pragmatic processes (match with the situation and a mutual exclusivity bias), and the deconstruction of complex wholes, where slots are assumed to be necessary meaningful. This latter constraint is motivated by a learner's desire to communicate: if a slot of a subtree specifies some meaning (an entity or a predicate) co-indexed with the subtree's global interpretation pattern in its root node, the learner can use it productively for generating and understanding novel utterances. If no such constraint is present, the pattern cannot be used, and the learner will not bother to extract it.

3.2 An experiment with artificial data

Can the model described above induce a grammar from utterances and situations these utterances are found in? As an initial test, we used artificial data loosely based on natural language to see if the model was able to discover patterns that allowed it to find the situation that the utterance was intended to refer to. The model saw each sentence paired with seven situations (which are represented as predicate-argument structures), of which one was the intended one. There were eight entities, ($Abe, Ben, Carl, Didi, Ed, Fay, Gerold, Hannah$), four single-place predicates ($laugh$, cry, $turn.fifty, die$, and four two place predicates ($see, shave, hit, push$). In total, there were $8 \times 4 + 8 \times 8 \times 4$ possible predicate-argument structures. Obviously, this is a gross oversimplification of the issues a child faces (it ignores the packaging problem (Gentner 1982, Gleitman 1990), does not use extralinguistic understanding of speaker intentions (Tomasello 2001) and overly restricts the space of possible situations), but as a toy example demonstrating the dynamics of the model, it will do.

The intended predicate was expressed by a simple subject-verb-object sentence if it was transitive (e.g. *Carl hits Ben* for $hit(Carl, Ben)$, and a subject-verb sentence if it was intransitive (e.g. *Carl cries* for $cry(Carl)$. If the subject and object were coreferential, the pronoun *himself* was used at the object position (e.g. *Carl hits himself* for $hits(Carl, Carl)$), except in the case of the predicate $shave$, where the object is simply not expressed if it is coreferential with the subject (e.g. *Gerold shaves* for $shaves(Gerold, Gerold)$). Two other exceptions are the predicate $turn.fifty$, which is expressed with the verb-object idiom *see Abe*, so $turn.fifty(Didi)$ is expressed with *Didi sees Abe*, and the predicate die, which is expressed with the verb-object idiom *kick bucket*, so $die(Carl)$ is expressed with *Carl kicks bucket*.[6]

Using this system, we can generate artificial data. From our 288 possible predicate-argument structures, we select seven at random for every entry. One of these seven is expressed according to the rules mentioned above. With this procedure, we generated twenty data sets of 1200 entries each. Because we are sampling, we have to repeat the experiment so that we know the results are not due to chance. Moreover, this gives us an insight in what it might mean for the cognitive representation of different learners acquiring a grammar from (slightly) different input. Furthermore, note that unlike in the example we gave earlier, here the model starts with an empty inventory. It will thus have to learn words, grammatical patterns and their meanings.

We measure the success of the model by evaluating if it can find the correct situation given the utterance. Because μ-DOP is a discovery mechanism rather than a declarative model of how the grammar should look like, we cannot evaluate if the parses are good or not, only whether they lead to interpretations that were intended. We will have a closer, more qualitative look at the parse trees that the model produces.

[6]For the puzzled reader, the first idiom is loosely based on the Dutch idiom *Abraham zien*, 'lit: to see Abraham; to turn fifty'

How well can the model find the intended predicate-argument structure from among the seven ongoing situations at that moment? Figure 14 shows the performance at each trial for all of the twenty simulations, with every dot standing for the average of the model's performance over the twenty samples. So, after having seen zero utterances, the model will be inconclusive about what situation utterance one refers to, and will therefore predict none of them correctly. With only a few subtrees in its inventory, it will be making mistakes, but we can see that after about 250 situations the model reached its asymptotic peak in performance at understanding around 90% of cases correctly.[7]

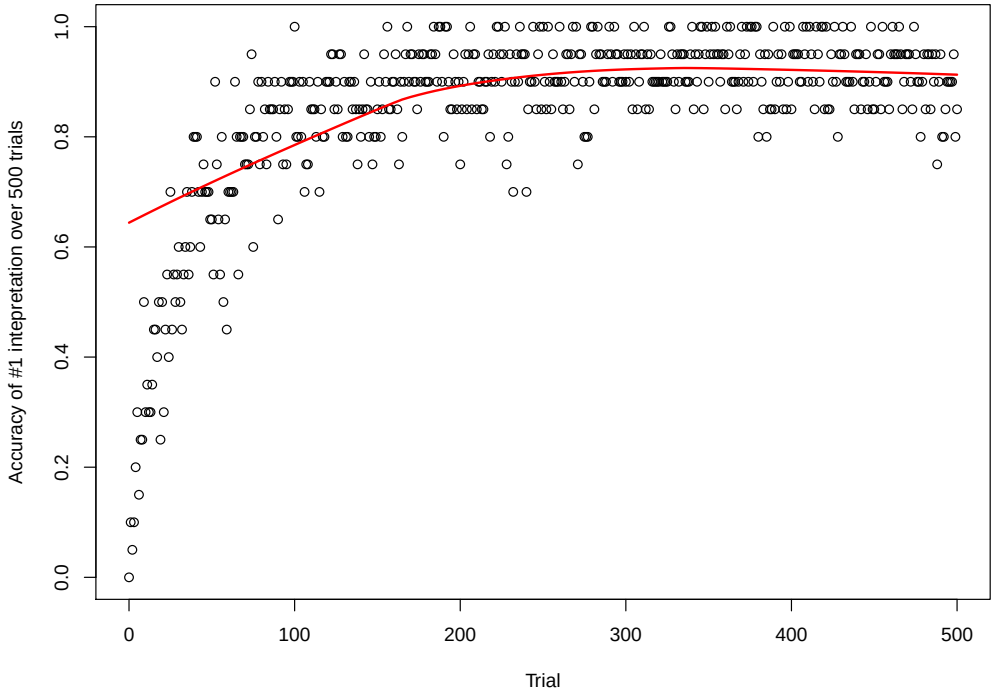

Figure 14: The average accuracy of μ-DOP in the first 500 trials

So exactly what patterns does the learner use to understand the utterances? Let us take a closer look at three cases.

First, the intransitive pattern was after a few instances being used as a fully abstract pattern. So the most likely derivation of a sentence like *A walks* is the one in figure 15. The model probably directly inferred this abstraction, because the semantic representations are very simple. Given these simple representations, the generalization quickly pays off: breaking down the analysis into its smallest component parts allows the learner to use the words and the more abstract pattern. Note that this means that there are no restrictions on the nature of P: there is nothing witholding the parser from substituting P with a transitive predicate like $P : see$

A second case is that of the *sees Abe*-construction (meaning that the subject turns fifty). This case is interesting, because it shows the effect of slightly different inputs on the learn-

[7]It should be noted here that the line in the figure oversmooths the data; it should be interpreted as merely indicating that the model displays a learning curve leading up to an asymptote. The actual curve seems to rise in a much steeper way.

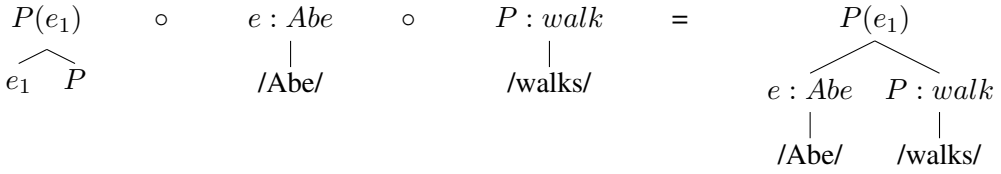

Figure 15: Most likely derivation for *Abe walks*

ers. Recall that we generated twenty different corpora of utterance-situation pairs. Because all of these are different, learners might extract different patterns. It may be that later they converge on the same representation, after having seen more evidence, but they may also retain different representations. As long as this does not hamper communication, that is not a problem. For the *sees Abe*-construction, we see such a development. Basically, there are two paths the learners take. In the first, they almost directly infer that /sees/ + /Abe/ is a chunk that should be treated as a whole and that combines with the intransitive pattern and a subject to form a parse (see Analysis 1 in figure 16). However, six out of fourteen learners started with another, communicatively correct analysis, viz. Analysis 2 in 16. These analyses lead to the same interpretation, and hence are fine for these six learners to use. However, after having seen more utterance-situation pairs, all learners abandon Analysis 2 in favor of Analysis 1. Arguably, they do so because the patterns decomposed from Analysis 1 will be reinforced more over other analysis (the pattern where /Ed/ + /sees/ means $e : Ed$ and the one where /Abe/ means $P : turn.fifty$ can only be used rarely, and hence will obtain lower frequency scores over time, whereas /Ed/ meaning $e : Ed$ are often reused in analyzing all sorts of sentences.

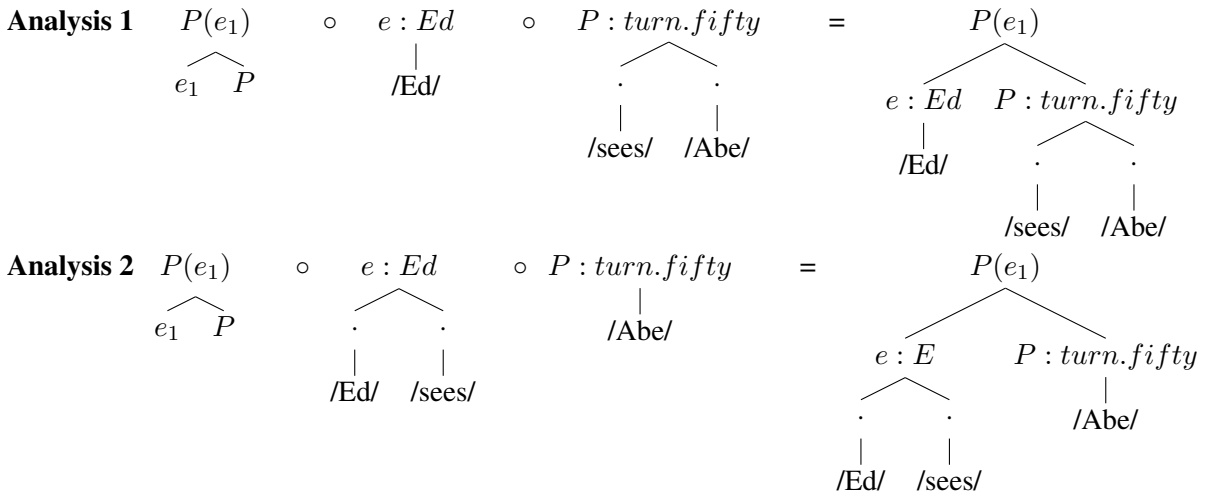

Figure 16: Two analyses of *Ed sees Abe* in the situation where $turn.fifty(Ed)$ is present

Another interesting aspect of the *sees A* construction is that it is ambiguous, in principle, between a literal reading ($see(e_1, Abe)$) and an idiomatic one ($turn.fifty(e_1)$). Thus, when facing a sentence *E sees A* and both the situations $see(Ed, Abe)$ and $turn.fifty(Ed)$ are present in its context, the model cannot make a well-founded choice. In our experiment, the learner selected the former situation as the intended one, presumably because of the higher frequency of $see(e_1, e_2)$ situations (with eight subject and eight objects: 64

instances) than $turn.fifty(e_1)$ situations (with only eight subjects: 8 instances).

A final case is that of the no-reflexive construction. With the predicate $shave$ and a coreferential subject and object, the reflexive is left out, so $shave(Ben, Ben)$ is expressed as *Ben shaves*. How does the learner respond to these sentences? In nineteen out of twenty cases, it will arrive at an analysis such as the one in Analysis 1 (17), where it has a pattern stating that there is a two-place predicate of which the coreferential argument is the first element (e_1) and the predicate the second. This means that the model constructs a generalization that is too broad: this pattern could now in principle also be used to parse and produce (ungrammatical) utterances like *Ed sees* for the situation $see(Ed, Ed)$. However, the model has no grounds of restricting this overgeneralization: there is an incentive to extract the third subtree (the 'word' /shaves/) from a more verb-island like pattern, viz. its use in the transitive argument-structure construction, where it fits the P-slot. We see here the effect of the whole system being interconnected by sharing members or parts of constructions. If it were not for the transitive pattern combining with the word /shaves/ (or [P : $shave$ /shaves/]), the no-reflexive sentences would have probably been analyzed using a more restrictive pattern like '[[e_1] [P : $shave$ [/shaves/]]] meaning $shave(e_1, e_1)$'.

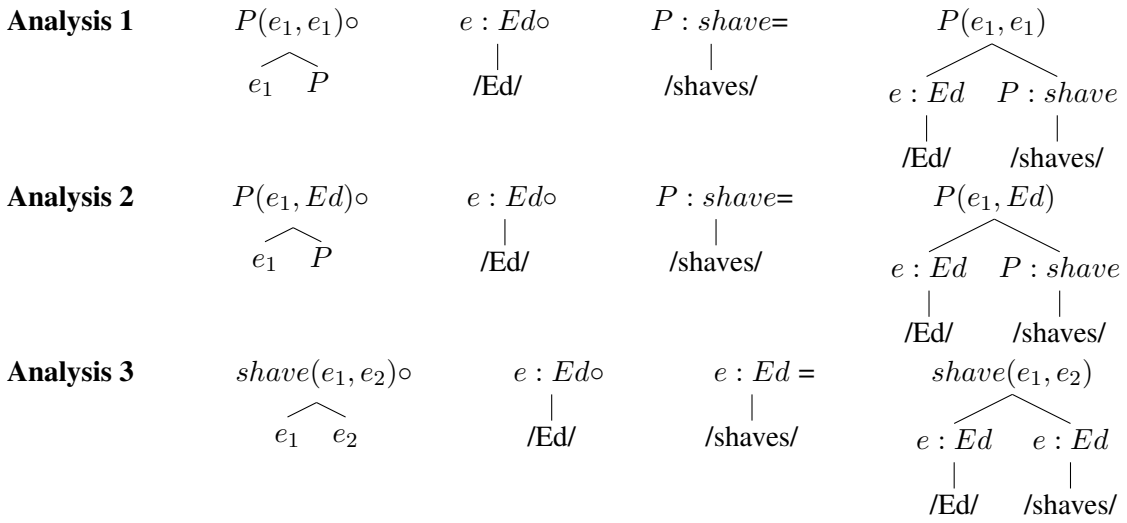

Figure 17: Three analyses of the no-reflexive sentence type

But apart from Analysis 1, the model also comes up with other analyses. In six simulations, the learner started with a pattern like in Analysis 3, that is, one in which the sound /shaves/ is paired with a range of entities and there is a pattern that states that if you juxtapose two entities, the interpretation is that the first entity shaves the second. After having seen more utterances, five out of six learners give up on this pattern in favor of Analysis 1, whereas one keeps using Analysis 2 and Analysis 3 for this sentence type. In one case, the learner starts with Analysis 2 and shifts to Analysis 1 after a while.

What these cases show, is that the model can acquire both compositional and idiomatic structures from the data, using one and the same mechanism. It makes overgeneralizations, which is realistic, and the learners arrive at different representations on the basis of slightly different inputs, but converge mostly after having seen more evidence.

3.3 Approaching the learner: whither μ-DOP?

Any formalization will leave certain questions unanswered: indeed, the formalization of a language learner from a constructivist perspective is an AI-complete problem, as not

only linguistic structure, but also conceptual structure and the memory system will have to be modeled. Model criticism on these levels is nevertheless very welcome and helps modelers develop more realistic models. Extending the representational formalism would be required to be able to understand alternations of the argument structure patterns (e.g., passivization and clefting), for instance, but this point does not bear on the essential learning strategy proposed by DOP, but rather on the nature of the structure in the input. We will discuss one further criticism in this paragraph that does relate to DOP's structure discovering mechanism, acknowledging the essential limitations present to the current line of reasoning.

One central problem with the μ-DOP approach is the fact that it allows for all possible fragments from the beginning onwards. This means that after having processed its first experience, the learner already has a (weak) generalization of, say, the subject preceding the verb. We have seen how the intransitive pattern is directly generalized to its most abstract form. This direct generalization is obviously unrealistic in the light of early language being rather holophrase-based. We can expect a learner to go about in a more conservative manner, only making a schema after it has seen multiple instances of an utterances. Much effort has been spent within the framework of Bayesian learning to understand how a conservative learner can create patterns that are both restrictive and open-ended enough to avoid both undergeneralization and overgeneralization (Stolcke 1994, O'Donnell, Snedeker, Tenenbaum & Goodman 2011), and it is in similar mechanisms that our understanding of generalization has to be looked at (see also Beekhuizen et al. (2013)).

4 Conclusion

In this paper, we have tried to show the potential of formal models for the constructivist approach to grammatical structure. We presented a domain-general, meronymy-based learning mechanism, viz. Data-Oriented Parsing, and show how it can be used as an unsupervised learning mechanism (U-DOP) to address learnability issues using constructivist principles such as heterogeneity of representation size and redundancy. We further gave an example of how we can incorporate meaning in U-DOP and showed how such a model can in principle learn from noisy data, although it remains to be seen how such a model behaves given naturalistic data. Although the specifics of the presented models may not be viable (perhaps there are problems with the learning mechanism, perhaps there are cognitively unrealistic assumptions), they show at least how we can formalize our understanding of linguistic learners in such a way that we can shed light on known phenomena and perhaps discover new ones.

References

Ambridge, B., Rowland, C. & Pine, J. (2008), 'Is structure dependence an innate constraint? New experimental evidence from chilrens complex-question production.', *Cognitive Science* **32**, 222–255.

Arnold, D. & Linardaki, E. (2007), HPSG-DOP: Towards Exemplar-based HPSG, *in* 'ESSLLI'.

Beekhuizen, B. F., Bod, R. & Zuidema, J. (2013), 'Refining the all-fragments assumption: The search for parsimony in redundancy', *Language and Speech* **56**(3).

Bod, R. (1998), *Beyond Grammar: An Experience-Based Theory of Language*, CSLI, Stanford, CA.

Bod, R. (2009), 'From Exemplar to Grammar: A Probabilistic Analogy-Based Model of Language Learning', *Cognitive Science* **33**(5), 752–793.

Bod, R. & Kaplan, R. (1998), A Probabilistic Corpus-Driven Model for Lexical-Functional Analysis, *in* 'Proceedings of ACL/COLING', pp. 145–151.

Bod, R., Scha, R. & Sima'an, K., eds (2003), *Data-Oriented Parsing*, University of Chicago Press, Chicago, IL.

Bod, R. & Smets, M. (2012), Empiricist Solutions to Nativist Puzzles by means of Unsupervised TSG, *in* 'Proceedings EACL'.

Bonnema, R., Bod, R. & Scha, R. (1997), A DOP Model for Semantic Interpretation, *in* 'Proceedings ACL-EACL'.

Borensztajn, G., Zuidema, W. & Bod, R. (2008), Childrens grammars grow more abstract with age Evidence from an automatic procedure for identifying the productive units of language, *in* 'Proceedings of the Annual Conference of the Cognitive Science Society', Vol. 1, pp. 175–188.

Brown, R. (1973), *A First Language*, Harvard University Press, Cambridge, MA.

Bybee, J. (2006), 'From Usage to Grammar: The Mind's Response to Repetition', *Language* **82**(4), 711–733.

Chang, N. C.-L. (2008), Constructing Grammar: A computational model of the emergence of early constructions, Dissertation, University of California, Berkeley.

Chomsky, N. (1986), *Knowledge of Language: Its Nature, Origin, and Use.*, Praeger, Westport, CT.

Chomsky, N. (1993), A minimalist program for linguistic theory, *in* K. L. Hale & S. J. Keyser, eds, 'The view from Building 20: Essays in linguistics in honor of Sylvain Bromberger', MIT Press, Cambridge, MA, pp. 1–52.

Clark, A. & Eyraud, R. (2006), Learning auxiliary fronting with grammatical inference, *in* 'Proceedings of CoNLL', pp. 125–132.

Clark, A. & Lappin, S. (2011), *Linguistic Nativism and the Poverty of the Stimulus*, Wiley-Blackwell, London.

Crain, S. (1991), 'Language acquisition in the absence of experience.', *Behavioral and Brain Sciences* **14**, 597–612.

Croft, W. & Cruse, D. A. (2004), *Cognitive Linguistics*, Cambridge University Press, Cambridge, UK.

Daelemans, W. & Van den Bosch, A. (2005), *Memory-Based Language Processing*, Cambridge University Press, Cambridge, UK.

Fillmore, C. J. & Kay, P. (1996), 'Construction Grammar'.

Foraker, S., Regier, T., Khetarpal, N., Perfors, A. & Tenenbaum, J. (2009), 'Indirect Evidence and the Poverty of the Stimulus: The Case of Anaphoric One', *Cognitive Science* **33**(2), 287–300.

Frank, S. L., Bod, R. & Christiansen, M. H. (2012), 'How hierarchical is language use?', *Proceedings. Biological sciences / The Royal Society* **279**(November 22), 4522–4531.

Gale, W. A. & Sampson, G. (1995), 'Good-Turing without tears', *Journal of Quantitative Linguistics* **2**(3).

Gentner, D. (1982), Why Nouns are Learned Before Verbs: Linguistic Relativity versus Natural Partitioning, *in* Stan A. Kuczaj II, ed., 'Language Development. Volume 2: Language, Thought, and Culture', Lawrence Erlbaum Associates, Hillsale, New Jersey, chapter 11, pp. 301–334.

Gentner, D. (1983), 'Structure-mapping: A theoretical framework for analogy', *Cognitive Science* **7**(2), 155–170.

Gleitman, L. (1990), 'Sources of Verb Meanings', *Language Acquisition* **1**(1), 3–55.

Goldberg, A. E. (1995), *Constructions. A construction grammar approach to argument structure.*, Chicago University Press, Chicago, IL.

Goldberg, A. E. (2006), *Constructions at Work. The Nature of Generalization in Language*, Oxford University Press, Oxford.

Jurafsky, D. (1996), 'A Probabilistic Model of Lexical and Syntactic Access and Disambiguation', *Cognitive Science* **20**(2), 137–194.

Jurafsky, D. (2003), Probabilistic Modeling in Psycholinguistics: Linguistic Comprehension and Production, *in* R. Bod, J. Hay & S. Jannedy, eds, 'Probabilistic Linguistics', MIT Press, Cambridge, MA.

Langacker, R. W. (1989), *Foundations of Cognitive Grammar, Volume I*, Stanford University Press.

Mehler, J. & Carey, P. (1968), 'The Interaction of Veracity and Syntax in the Processing of Sentences', *Perception and Psychophysics* **3**, 109–111.

O'Donnell, T. J., Snedeker, J., Tenenbaum, J. B. & Goodman, N. D. (2011), Productivity and Reuse in Language, *in* 'Proceedings CogSci'.

Perfors, A., Tenenbaum, J. B. & Wonnacott, E. (2010), 'Variability , negative evidence , and the acquisition of verb argument constructions', *Journal of Child Language* **37**, 607–642.

Peters, A. M. (1983), *The Units of Language Acquisition*, Cambridge University Press, Cambridge, UK.

Pinker, S. (1979), 'Formal models of language learning.', *Cognition* **7**(3), 217–83.

Pullum, G. K. & Scholz, B. C. (2002), 'Empirical assessment of stimulus poverty arguments', *The Linguistic Review* **19**(1-2), 9–50.

Scha, R. (1990), Taaltheorie en taaltechnologie; competence en performance, *in* R. de Kort & G. Leerdam, eds, 'Computertoepassingen in de Neerlandistiek', LVVN, Almere, pp. 7–22.

Schuchardt, H. (1885), *Ueber die Lautgesetze: Gegen die Junggrammatiker*, Robert Oppenheim, Berlin.

Stolcke, A. (1994), Bayesian Learning of Probabilistic Language Models, Dissertation, University of California, Berkeley.

Tomasello, M. (2001), Perceiving intentions and learning words in the second year of life, *in* M. Bowerman & S. C. Levinson, eds, 'Language Acquisition and Conceptual Development', Cambridge University Press, Cambridge, UK, chapter 5, pp. 132–158.

Tomasello, M. (2003), *Constructing a language: A Usage-Based Theory of Language Acquisition*, Harvard University Press, Cambridge, MA.

Trijp, R. V., Steels, L., Beuls, K. & Wellens, P. (2009), Fluid Construction Grammar : The New Kid on the Block, *in* 'ACL 2012'.

Verhagen, A. (2009), 'The conception of constructions as complex signs. Emergence of structure and reduction to usage', *Constructions and Frames* **1**, 119–152.

Wunderlich, D. (2007), Why assume UG?, *in* M. Penke & A. Rosenbach, eds, 'What counts as evidence in linguistics?', Benjamins, Amsterdam, pp. 147–174.

Zuidema, W. (2006), Theoretical Evaluation of Estimation Methods for Data-Oriented Parsing, *in* 'Proceedings EACL', pp. 1–4.