From cradle to control verbs
Extending constructivist models beyond simple syntax

Barend Beekhuizen
Leiden University & University of Amsterdam

5 January 2012

1 Introduction

In recent years, several computational models (Alishahi & Stevenson 2008, Chang 2008, Bannard, Lieven & Tomasello 2009) have begun fulfilling the promise of usage-based construction grammar (Tomasello 2003, Goldberg 2006) as a computationally testable account of the acquisition of grammar.[1] In simulating specific developmental patterns, such as the generalization of argument structure constructions, constructionist models seem to focus on a limited part of the childs development, namely the acquisition of so-called simple syntax (Tomasello 2003, Ambridge & Lieven 2011), i.e. word order in declarative sentences and argument structure. This focus has left the models conceptually ill-equipped to account for the grammatical development on either developmental side of the acquisition of simple syntax.

On the one side, most computational constructivist models do not address the fact that the child will have to learn that language is compositional in the first place, and on the other, to the best of our knowledge, no attempt has been made to scale up constructivist models in order to account for more complex phenomena assumed by nativists to rely on the learners innate linguistic abilities. In order to show that usage-based construction grammar is a valid theory of language acquisition, it is important that a single model can be shown to simulate the full range of a learners grammatical development, from early segmentation of the utterance to the control of such complex grammatical patterns as pronominal binding and control. In this paper, we will discuss in further detail why the aforementioned computational constructivist models lack these desiderata and present, at a very coarse level, a More Complete Usage-Based Learner that aims at incorporating them.

2 Discovering Compositionality

On a usage-based account of language acquisition, the learner comes to the task of acquiring a grammar with no preconceptions about the architecture of the ambient language (Tomasello 2003). A more radical perspective emphasizes that the child is not even aiming at learning a language; it rather tries to communicate, with the acquired grammar being a by-product of this attempt (Hopper 1998). As such, the child is not looking for compositional structure as a goal in itself. In order to begin learning compositional structures, the learner will first have to realize that there are co-varying correspondences and differences over the experienced utterances and the learners hypothesized interpretation of those utterances. Secondly, the learner should understand that assuming compositional constructions capturing these constants and variables leads to more communicative flexibility as well as a more economic cognitive representation. These realizations only emerge in response to the experienced communicative situations, and do so incrementally. Because this process is data-driven, it does not have to occur by necessity, but it is only a rational response to the complexity of the task at hand.[2]

As far as we know, no constructionist model fully incorporates this idea.[3] The closest to doing so is Bannard et al. (2009), in which the segmentation of multi-word strings (but not the elements of meaning) is done through the alignment of strings of words from a corpus, but this model lacks incrementality. Chang (2008), on the other

[1] Cf. (Bod 2009) for a critique on the lack of such endeavors.

[2] Given, for instance, a hypothetical fully non-compositional language, the usage-based learner will never start learning abstract patterns possibly governing that language.

[3] These, and the following comments, should not be taken as criticisms on the models per se or their achievements, but rather as pointers to aspects of the models that make them unfit to be extended in order to simulate segmentation or the development of more complex syntax.

hand, does describe an incremental model, but the learner starts with a vast set of lexical constructions (i.e. words and their meanings) and hence cannot be said to 'discover' compositionality, as it knows already that there are multiple meaningful items in the utterance that will have to be combined in some meaningful way. Similarly, in Alishahi & Stevenson (2008) the quest for abstraction is part of the design of the model: it starts with structured utterances and discovers the optimally compressing clusters. What is needed in order to simulate the emergence of grammar, is a model in which the initial input is unsegmented at some level and in which the option of using partially abstract representations gradually becomes more attractive to the learner than storing full utterances coupled with the hypothesized meaning.

3 Prerequisites for Scaling Up

A second issue is that constructionist models have not addressed the issue of complex syntax extensively. Although experimental research on the acquisition of complex clauses has been done, there are no computational models simulating the development of such phenomena. Similarly, as Ambridge & Lieven (2011, ch. 8) discuss, no constructivist account of cross-clausal dependencies such as control and raising verbs and pronominal binding, has been proposed. Computational modeling on usage-based premises might help to develop such an account and show, for instance, item-based effects and and the presence of semantic and constructional prototypes that match the developing understanding of such structures. In order to deal with co-referentiality within an utterance and semantically embedded structures, a constructivist model would need a representational formalism to do so. Of the models that do incorporate meaning, two do not have the representational machinery to do so (Alishahi & Stevenson 2008, Bannard et al. 2009). Chang‘s model employs Frame Semantics, which is in principle scalable to account for the semantics underlying cross-clausal dependencies.

4 Towards a More Complete Usage-Based Learner

The discussed models seem to lack certain aspects of what it takes to develop a unified model that can in principle cover the whole range of syntactic development from segmentation to complexity beyond argument structure. What would be needed for such a model is the representational richness and incremental Bayesian Model Merging approach (Stolcke 1994) used in Chang's (2008) model combined with a segmentation strategy for breaking down larger structures that display regularities among each other. We can think of strategies akin to those described by Bannard et al. (2009) or van Zaanen (2001), but then applied to form and meaning simultaneously.

We developed and implemented such a model, provisionally titled a *More Complete Usage-Based Learner* or *MCUBL. MCUBL* reorganizes its constructional knowledge (initially consisting of mappings between entire utterances and entire meaning-complexes inferred from the context in which the utterances were produced) by merging similar constructions while storing both the aligned and the unalignable parts of the form and the meaning of the original constructions as novel constructions. Different such reorganizations are then evaluated as candidate grammars against the present and several previous utterances-in-context as well as on their prior probability as a grammar, where smaller, semantically less complex grammars are preferred. The optimal reorganization will then be selected as the base grammar for the next iteration. For a graphical representation, see figure 1. The model is strongly builds on the work of Chang (2008), but has a simpler formalism for representing the grammar and has merging operations that are specifically useful for segmentation purposes.

In the development of this model, most attention to this point has been directed at enumerating and operationalizing the desiderata for a constructionist learner that can in principle start with nothing and end up having a command of the grammar beyond 'simple syntax', something which has been done insufficiently thus far. Only small-scale tests on corpus material concerning the initial segmentation has been done, and only with qualitative evaluations. Because of the use of Frame Semantics as a representational formalism, we believe the model can be used to simulate the emergence of more complex grammatical constructions. Taking stock of such grammatical constructions and developmental patterns in want of a constructivist account as well as setting up the experiments leading to an actual evaluation of the model are presently our main foci. A further point of interest is the nature of the data and (semantic) resources that should be used to appropriately evaluate the outcomes of the model.

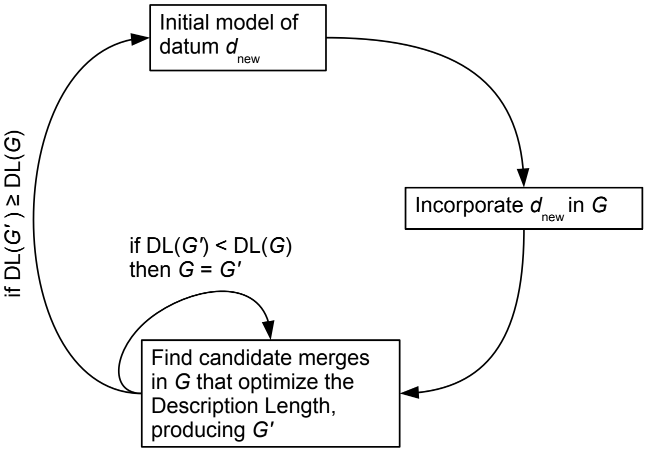
Initial model of datum d_{new}
Incorporate d_{new} in G
Find candidate merges in G that optimize the Description Length, producing G'
if DL(G') < DL(G) then G = G'
if DL(G') ≥ DL(G)

Figure 1: A high-level representation of the More Complete Usage-Based Learner.

References

Alishahi, A. & Stevenson, S. (2008), 'A Computational Model of Early Argument Structure Acquisition', *Cognitive Science: A Multidisciplinary Journal* **32**(5), 789–834.

Ambridge, B. & Lieven, E. V. M. (2011), *Child Language Acquisition. Contrasting Theoretical Approaches*, Cambridge University Press, Cambridge, UK.

Bannard, C., Lieven, E. & Tomasello, M. (2009), 'Modeling children's early grammatical knowledge.', *Proceedings of the National Academy of Sciences of the United States of America* **106**(41), 17284–9.

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

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

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

Hopper, P. (1998), Emergent grammar, *in* M. Tomasello, ed., 'The new psychology of language: cognitive and functional approaches to language structure', Lawrence Erlbaum., Mahwah, NJ, pp. 155–176.

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

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

van Zaanen, M. M. (2001), Bootstrapping Structure into Language: Alignment-Based Learning, Dissertation, University of Leeds.