Morphological schema induction by means of conditional inference trees

Vsevolod Kapatsinski
(1) Department of Linguistics, 1290 University of Oregon, Eugene, OR, 97403, USA

Keywords: Construction Grammar, Cognitive Grammar, schema, CART.

1 Evidence for first-order schemas / constructions

Rule-based approaches to grammar (Albright & Hayes, 2003; Reiss, 2004, inter alia) suggest that knowledge of grammar consists largely of knowledge of rules. A rule describes a change and the context in which that change has been observed to occur and is to be carried out (e.g., Reiss, 2004). In its strongest version (Reiss, 2004), Rule-based Phonology proposes that phonology is based only on rules. Constructionist and Cognitive theories of grammar instead suggest that grammar largely consists of first-order schemas or constructions, which are generalizations about form-meaning pairings.

For instance, the learner of a language that has many singular-plural pairs like *buti*-buti might notice that plurals in the language often end in [Vt]. The speaker of such a language might then derive [Vt]-final plurals from novel singulars, even ones that do not end in [Vt], because s/he thinks that word-final [Vt] is a good expression of plural meaning. These [Vt]-final plurals might be derived from singulars that already end in a consonant, giving rise to palatalization, e.g., bukSG-buksiPL. This result was recently shown experimentally in human miniature artificial language learning (Kapatsinski, 2012, 2013). The result is expected if the learners are extracting first-order schemas but is unexpected if they are extracting rules (changes in context), since tsSG/tiPL supports the change 0>i but actually provides evidence against the changes k>t or t>t (confirmed by computational simulations using the rule-based Minimal Generalization Learner developed by Albright & Hayes, 2003).

Further support for first-order schemas comes from from “wug tests”, in which subjects are often observed to overuse common output patterns, deriving them in ways unattested in the lexicon, as well as the natural language phenomena of hypercharacterization, affix-fusing rule telescoping and haplology (see Booij, 2010, and Bybee, 2001: 126-129). For instance, Booij (2010) mentions the case of UHD, the abbreviation for universitair hoofddocent ‘assistant professor’ in Dutch, being often redundantly marked by the agentive –er to become UHD-er, thus coming to fit the agentive noun schema “…er” characterizing most agentive nouns.

Thus there is much empirical support for human language learners acquiring first-order schemas. However, there has been no computational work explicitly addressing first-order schema learning. Having a model of first-order schema extraction would ideally allow us to predict, given a lexicon, what schemas a human learner of that lexicon would acquire.

Without such a model, Constructionist and Cognitive theories of morphology are seriously underspecified in their predictions. For instance, consider again a learner being exposed to a lexicon containing many examples of Vt-final singulars corresponding to Vt-final plurals. Should the learner think that plurals in this language often end in [V#]? [CV#]? [VCV#]? [Ci#]? [VCi#]? [tfi#]? [Vtfi#]? [Vt/V#]? Note that if the level of schema generality is too high, such examples would not support bukSG-buksiPL over bukSG-bukiPL: both [buti] and [buki] fit the schemas [VCi], [VCV], [V], etc. At the same time, the learners in the same experiment took examples like but/buti and bup/bupi to support buk/buki over buk/but (Kapatsinski 2012, 2013). Thus, for the Cognitive/Constructionist theories of grammar to account for the results, the schemas should be neither too general, nor too specific: [buti], [bupi] and [buki] must fit the same schema, supporting each other, while [buti] must be distinct. Achieving the right level of generality is thus not a trivial problem.
2 Schema induction

First, we depart from the standard approach in Construction/Cognitive Grammar in proposing that schemas, at least to a large extent, become more specific, rather than becoming more general as more of the grammar is acquired. The crucial prediction is early liberalism: upon exposure to a few plural wordforms, one thinks that plurals in the language can be pretty much anything, rather than thinking they can only be the words one has just experienced. Upon hearing [bupi] paired with multiple novel creatures one does not think that the plural form of any word is [bupi], and does not overgeneralize [bupi] to suppletively replace plural forms of other words no matter how often [bupi] is heard. Learners start out thinking that –i and –a can attach to stems to make plurals, but even at the end of training may not grasp that [ka] and [ki] are illegal. When schemas have not yet calcified into strong preferences for the observed sound sequences, the learner is open to experience and is ready to learn.

General-to-specific learning is supported by apparent underspecification of early lexical representations (e.g., Swingley 2007): early on during language acquisition, child and adult learners tolerate single-feature mismatches despite being able to hear the difference. The general pattern of results is that while correctly pronounced words are recognized more easily than slightly mispronounced ones, mispronunciations of low-frequency familiar words are preferred over unfamiliar words. This is expected if learners are gradually strengthening the more specific schema that does not allow for mispronunciations but still retain the more general schema that allows for some featural mismatch. As learners continue hearing a word, the more specific schema strengthens, thus for highly familiar words featural mismatches are not tolerated even by young children. The phonological representation of a word is thus one instance of a schema where the specification process is relatively uncontroversial. We propose that this extends to all first-order schemas. When the schema is weak, one is willing to accept major deviations from the previously encountered examples but the tolerance decreases as the distribution of experienced exemplars grows and its believable extent shrinks.

In the proposed theory, schema specification proceeds by seeking out unexpected bumps in the joint probability space defined by meanings and sounds, i.e., which kinds of sequences are unexpectedly frequent in plural forms? Xu & Tenenbaum (2007) document this kind of inference for semantic categories: suppose you are presented with a picture of a Dalmatian paired with the word fep. At first you are likely to think that fep means “dog”. However, if fep is presented to you three times, each time paired with a picture of a different Dalmatian, you are likely to discard the hypothesis that fep means “dog” as it would be a very suspicious coincidence that a process of randomly sampling dogs would produce three Dalmatians in a row. While Xu & Tenenbaum (2007) use these data to argue for a Bayesian model of the acquisition of lexical semantics, Rogers & McClelland (2004) provide evidence that distributed connectionist networks likewise predict that meanings should grow more specific, and differentiated from each other, over the course of development (they therefore term this general-to-specific learning "progressive differentiation").

We argue that schemas are conditional on meaning, rather than form, i.e., they are fundamentally production-based: schemas are not the most reliable cues to meaning; they are the most common realizations of meanings. This can be seen by considering the effect of adding examples like [but[\text{SG}] but[\text{PL}]] on the productivity of palatalization. These examples actually reduce the reliability of stem-final [t\text{]} as a cue to plurality by introducing cases in which [t\text{]} is found in the singular. Nonetheless, such examples help palatalization. This finding makes sense if schemas are fundamentally based on production, rather than perception experience: they are what is most likely to be produced given that a certain meaning is intended.

We implement general-to-specific schema extraction as conditional inderence tree induction using the ctree() function in the party package (Hothorn et al., 2006) in R (see Daelemans & van den Bosch 2005 for a conceptually similar approach). We believe conditional inference trees to be particularly well suited to this purpose (compared to regression-based techniques) because they are well known to be good at dealing with small-n/large-p problems, where the number of observations is small and the number of predictors is large. This is the actual situation for small lexica, and every speaker’s lexicon is small at early stages of language acquisition.

We coded each of the words presented to learners in Kapatsinski (2012) in terms of the features of the stem vowel, the stem-final consonant, and the identity of the final vowel. We also generated trigrams that are phonotactically legal in the learners’ native language but non-occurring in plural forms in the miniature
artificial language presented to them and entered them into the training set with a frequency of zero. We thus assume learners use implicit negative evidence bootstrapping schema acquisition from prior acquisition of phonotactics (as is commonly assumed in the literature, e.g., Pater & Tessier 2003). The dependent variable to predict was the type frequency of the resulting word-final trigram.

The ctree() recursively partitions the space defined by the predictors into rectangular areas such that at every split entropy reduction is maximized. The predictor producing the best binary split is at the top of the tree, with other predictors entering the tree if they improve predictiveness within the bins defined by the predictors already in the tree. At each step, the predictor that achieves the best split within a branch is entered into that branch.

We follow prior work in Cognitive/Construction Grammar tradition in proposing that schemas are positive, describing observed rather than non-existent words (Bybee, 2001; Nesset, 2008). Thus only the branches with non-zero predicted counts are stored. Generally, a first-order schema is then defined as in (1), and we propose that all schemas in a decision tree are extracted from the data.

(1) A first-order schema is a path through a conditional inference tree in which the predicted variable is occurrence/non-occurrence of an ngram, observations are word types, and the predictors are semantic and phonological features of words containing that ngram. The path must proceed downwards from the root of the tree terminating in a node that is either 1) a leaf with a non-zero type frequency or 2) an ancestor to at least one leaf with a non-zero type frequency.

3 Results

The proposed mechanism of schema induction is able to predict the experimental results in Kapatsinski (2012, 2013). As shown in Figure 1, [buti], [bupi] and [buki] are learned to be instances of the schema [-Pal]i#=PL (as well as the more general schema i#=PL). The reason [+Pal] is singled out by the model is that a large proportion of i#-final plurals are t[i#-final plurals, and [t] is the only [+Pal] segment in the inventory. [buti] is predicted to be distinct, being an instance of [+Pal]i#=PL. Thus, additional examples of but[iSG]-but[iPL] are predicted to help butSG-butiPL, bukSG-bukiPL, and bupSG-bupiPL through the schema [+Pal]i#=PL, while additional examples of bupSG-bupiPL are predicted to help bukSG-bukiSG over bukSG-butiPL through the [-Pal]i# schema. Space prevents me from showing the trees for all languages presented to human learners here. See Kapatsinski (2013) for the full results.

![Figure 1: Schemas extracted for one of the languages in Kapatsinski (2013). 'Yes' means that the ngram defined by the path from the top of the tree to the leaf in question occur in the training lexicon. Final vowel is –i (not shown). Schemas: PL=...i#, PL=...[+Pal]i#, PL=...[-Pal]i#, PL=...V[+Pal]i#](image-url)
4 Limitations and future directions

While the classification and regression trees capture the results in Kapatsinski (2012, 2013), there are a number of issues that remain for future work. While I find the arguments for general-to-specific learning and the primacy of first-order schemas fully convincing, the current implementation is only a very limited proof of concept. Three issues in particular are quite troubling. First, the model at present has no notion of serial order: the features combined into a schema could come from segments that are far apart in the word, or from adjacent segments, or the same segment, with equal ease. This lack of locality bias is clearly unrealistic (Albright & Hayes 2003). Schemas should probably be at least strongly biased to map meanings to perceptually coherent units, excluding (or at least making exceedingly unlikely) the possibility of grouping non-adjacent features coming from different tiers into a single schema. Second, especially in a small lexicon, many features are highly redundant, and the decision of which one of these redundant features will be placed in the tree, and which other ones will be discarded, is very sensitive by small changes in the training data. Because CART trees do not tolerate redundancy, once a feature is placed in the tree, all other redundant features are eliminated. This is clearly incorrect: redundancy is a crucial design feature of language and should not only be tolerated but sought after for the purposes of building a robust perception-production loop. It is, perhaps, then better to read off schemas from the entire collection (random forest) of believable trees. Parallel distributed networks (Rogers & McClelland 2004) may also provide a plausible alternative here. Third, empirical data demand the schemas read off the tree not to necessarily terminate in a leaf: requiring all schemas to be full paths through the tree would result in schemas that are much too specific to be able to deal with novel inputs. This raises the question of how such non-fully-specific schemas should be weighted relative to other schemas. While some proposals are presented in Kapatsinski (2013), much work remains to be done on this issue. Finally, a broader question is how first-order schemas interact with second-order schemas, which are necessary for learning and producing arbitrary paradigmatic mappings (Booij 2010, Nesset 2008), and how both of these interact with perseveratory tendencies militating against stem changes. While the evidence for both first-order and second-order schemas appears convincing, much work remains to be done in integrating schemas into a full-fledged theory of grammar.

References


