Blake Allen, UBC
Justified naivety: limits on constraint conjunction in inflectional morphology

Context and motivation: Recently, sublexical phonology (Becker and Gouskova 2013, Allen and Becker in review) has applied the probabilistic, constraint-based phonological formalism of MaxEnt Harmonic Grammar (Hayes and Wilson 2008) to the domain of inflectional morphology. The scope of this approach has been limited to making predictions given a single known "base" cell in an inflectional paradigm, e.g. modeling the task of predicting the plural form of a novel noun from its singular form. Consequently, the question of how predictions can be made from multiple known base forms of a word has been left unresolved, even though the complexity of many of the world's inflectional systems suggests that such inference must be possible (Stump and Finkel 2013).

This potential complexity poses a severe problem for learnability. Even relying on only a single base cell, the space of constraints to search while learning the phonological correspondences between that cell and the derivative cell is potentially infinite, and it must be reduced to a manageable size through simplyfing assumptions (Hayes and Wilson 2008). In the case of multiple base cells, there is no reason *a priori* to exclude constraints that are conjunctions of constraints on different base cells, e.g. a constraint [*NominativeSingular:* *e# & *GenitiveSingular:* *i#]. The space of these conjoined constraints grows far more quickly than that of non-conjoined constraints. Moreover, it is possible to construct hypothetical inflectional systems that require such *cross-base constraint conjunctions*, meaning that such languages could exist.

Proposal: In this presentation, I provide evidence that no cross-base constraint conjunctions are required by existing inflectional systems. To do so, I show that a computationally implemented model of grammar without these constraints accurately accounts for a variety of inflectional systems selected to provide wide coverage of the morphological typology. I then schematize the type of hypothetical inflectional system that *does* require these constraints. From this mismatch between the typology and the space of possible inflectional systems, I conclude that cross-base constraint conjunctions are absent from the constraint search space. Finally, casting this finding in the language of probability theory, I show that the combinatorial problem posed by cross-base constraint conjunction directly parallels a more general issue in the domain of statistical machine learning, and also that the solution proposed here of disallowing constraint conjunction is effectively equivalent to the well-studied statistical model known as Naive Bayes, opening up to phonologists the extensive literature on this model.

Methods and evidence: I test the adequacy of a model of grammar without cross-base constraint conjunctions on the following datasets: Spanish present tense verbs, Latin "principal parts" and their associated forms, Japanese verbs, and Kwerba nouns. These inflectional systems vary substantially in the predictiveness relations that hold among their various cells (Stump and Finkel 2013), and so I conclude that a model of grammar able to account for all of these datasets can be tentatively assumed to account for inflectional morphology more generally, pending investigation of additional datasets. The testing procedure amounts to performing leave-one-out cross-validation on each paradigm, essentially hiding each form of each word's paradigm one at a time, and having the model predict it from the other forms.

The model lacking cross-base constraint conjunctions that I have tested on these datasets is a simple generalization of the sublexical grammar architecture (Becker and Gouskova 2013, Allen and Becker in review). In order to ban cross-base constraint conjunctions when the probability of an output candidate is predicted from multiple bases, this probability must be derived from only constraints that each refer to a single base form, e.g. two constraints [*NominativeSingular:* *e#] and [*GenitiveSingular:* *i#], but not a single conjoined constraint [*NominativeSingular:* *e# & *GenitiveSingular:* *i#]. I impelement this restriction as follows: for each available base, allow it to predict the probability of the output candidate $p(c|base)$ in the standard way, using only constraints that refer to that base, and then multiply these probabilities across all available bases to reach the final predicted probability for the output candidate. Note that normalization is omitted in this simplified description but not in the model itself.

This model has the advantage of being falsifiable. I will describe examples of hypothetical inflectional systems that this model of grammar predicts to be non-existent, unproductive, or diachronically unstable. Despite the author's efforts to search for such inflectional systems in natural languages, none have yet been found. A generalized description of the property shared by these inflectional systems is forthcoming, but for now the search will be continued by running the leave-one-out cross-validation procedure described above on additional datasets.

Assuming that no such inflectional systems are found, the finding that cross-base constraint conjunctions are outside the constraint search space has a beneficial implication for phonologists: inflectional morphology can be expressed using a statistical model called Naive Bayes. Definitionally, supposing a set of candidates for the unknown form of a word, the probability of one candidate form c given a subset of the other forms of that word $f_1, f_2, ...f_n$ can be written as $p(c|f_1, f_2, ...f_n)$. Applying Bayes's theorem, this probability is proportional to $p(f_1, f_2, ...f_n|c)p(c)$, where $p(c)$ is the prior probability of the candidate c. The term $p(f_1, f_2, ...f_n|c)$ permits the influence of cross-base constraint conjunctions, as the probability of one base form given c can depend on the probabilities of the other base forms. Disallowing cross-base constraint conjuctions, we can decompose this term into $p(f_1|c)p(f_2|c)...p(f_n|c)$, which is equivalent to the generalized sublexical model described above. Notably, this new definition, $p(c|f_1, f_2, ...f_n) \propto p(c)p(f_1|c)p(f_2|c)...p(f_n|c)$, is identical to the form of the Naive Bayes model from statistical machine learning (Maron and Kuhns 1960). This result means that phonologists interested in applying the sublexical approach to inflectional morphology can make use of the extensive literature on Naive Bayes, including papers on its learnability properties and its numerous implementations in various programming languages.

References

Allen, Blake, and Michael Becker. In review, *Phonology.* Learning alternations from surface forms with sublexical phonology.

Becker, Michael, and Maria Gouskova. 2013. Source-oriented generalizations as grammar inference in Russian vowel deletion. Ms. lingbuzz/001622.

Hayes, Bruce, and Colin Wilson. 2008. A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry* 39:379–440.

Maron, Melvin Earl, and John L Kuhns. 1960. On relevance, probabilistic indexing and information retrieval. *Journal of the ACM* 7:216–244.

Stump, Gregory, and Raphael A Finkel. 2013. *Morphological typology: From word to paradigm.* Cambridge: Cambridge University Press.