

# Using Minimum Description Length to make Grammatical Generalizations

Mike Dowman

Graduate School of Arts and Sciences

University of Tokyo

It has often been proposed that language acquisition presents a learnability paradox, because there is not sufficient information in the language data to which children are exposed for them to be able to unambiguously determine the underlying linguistic system. Children have to make appropriate generalizations from language data, in order to produce novel grammatical sentences, but must also avoid generalizations that result in sequences of words that are not grammatical. While previous studies have shown that minimum description length can result in appropriate generalizations when learning verb subcategorizations from simplified artificial data sets, little research has addressed this same problem using naturally occurring language data.

The work reported here learned verb classes from the Switchboard part of the Penn Treebank. This consists of conversational speech that is syntactically annotated for word class and phrasal constituents. The top level constituents in each verb phrase containing a past tense verb were taken to correspond to one possible subcategorization for the verb. Verbs were grouped into classes, each class containing the full range of subcategorizations observed with any of the verbs in the class. A minimum description length evaluation metric was used to assess how well each set of verb classes matched the data. Grouping together verbs which shared some subcategorizations would result in a simpler grammar, but, as it would also predict that all of the verbs in the class could occur with any of the subcategorizations it contained, the new grammar might not fit the observed data so well.

Starting with all verbs in a single class, an annealing search gradually found divisions of verbs into classes that resulted in better overall evaluations. A grammar with six verb classes was learned using this technique. Four of these classes appeared to be linguistically coherent, as they contained verbs with similar syntactic properties and one class contained only the verb *do*, which is syntactically distinct from any other verb. The final class covered verbs appearing in a wide range of structures, and might best be described as a miscellaneous class. This work therefore demonstrates that minimum description length can be used to learn verb subcategories from naturally occurring data.