Description of data sets

MTLC project with Rob Reynolds and Andrea Sims; 4.20.2017

I created 12 data sets as potential input data sets for our model. All of the data sets share the following properties:

* They have 7 classes and 7 morphosyntactic property sets. In meeting last week Rob and I decided that 7x7 is likely the best balance between computationally reasonable and enough size to get significant variation in strength of implicative structure.
* They are symmetrical in their implicative structure. This means that removing any single class will affect the system in the same way as removing any other class. I assume and hope this also means they will turn out to be homogeneous networks when we decide how we want to measure that.
* They only exhibit syncretism across classes, e.g., exponent x realizes MSPS A in class I and class II. There is never any syncretism within a class, e.g., exponent y realizes MSPS A and B in class I. I’m pretty sure that within class syncretism would not affect the complexity measures, but since I am not sure if it would affect how systems develop over time in the model (could it?), I decided to keep it as simple as possible by avoiding any within class syncretism.

The data sets differ in the following ways:

* Each system exhibits a different degree of uncertainty absent implicative structure (entropy).
* Each system exhibits a different degree of complexity when implicative structure is taken into account (conditional entropy).
* Each system differs in the strength of implicative structure (mutual information; difference between entropy and conditional entropy).

I created these systems by hand. The differences in implicative structure can be seen in two ways. The plot below shows the entropy and conditional measures of the system. The structure that leads to these differences is most visible in the excel sheet with all the systems, because I have highlighted cells that are unique to each class. I \*think\* these may be all of the possible ways to structure systems so they (1) have all the similarities listed above and (2) all the classes in each system are identical in how many unique cells they have and how much overlap the non-unique cells have with other classes.

The data sets are ordered by how many unique cells are in each class. In some cases, there systems with the same number of unique cells per class because there are multiple possible patterns of overlap in the non-unique cells. For example, when there is only one unique cell in each class (data6, data7, data8, data9), there are multiple possible ways for the remaining cells to overlap. In data9, MSPS\_A has a unique exponent in the first class ‘aa’; the remaining classes share a single exponent ‘b’. Contrast this with data7 where MSPS\_A has a unique exponent in the last class ‘aa’, but half of the remaining classes have exponent ‘a’ and half have exponent ‘h’ for MSPS\_A. The other two sets, data6 and data8 represent the other possible patterns of overlap when one cell is unique in each class. And so on and so forth.

I am not sure whether we should use all of these systems or a subset of them for the model. I figured I would come up with all the options I could, look at the distribution of their complexity measures, and then we can decide as a group which will be best to use.

