SupMat_5: Models run on Types

5.1. All data

A Random Forest of 500 trees run on all data, including Set Phrases, gives an overall error rate of 13.53%; most of these errors are made on zero (27% incorrectly classified), next is *the* (10.28% incorrectly classified), and most accurate is *a* (8.54% incorrectly classified).

In terms of variable importance, Type triggers the largest mean decrease in the Gini index (decrease = 522), followed by Number (decrease = 168) and Count (decrease = 110). The mean decreases then drop down to much smaller values for Set Phrase (decrease = 34), Elaboration (decrease = 23) and Corpus (decrease = 15).

Table (5.1): Confusion matrix for a Random Forest on all data

	zero	**a**	**the**
zero	**367**	38	98
a	20	**664**	42
the	49	51	**873**

5.2 Excluding Set Phrases and nouns that have no number

With a 6.36% error rate the prediction accuracy is quite high. Most of these errors are made on zero (11.7% incorrectly classified), next is *the* (6.4% incorrectly classified), and most accurate is *a* (2.7% incorrectly classified).

In terms of variable importance, Type triggers the largest mean decrease in the Gini index (decrease = 567), followed after a drop in values by Number (decrease = 171), and then Count (decrease = 105). Elaboration (decrease = 17) and Corpus (decrease = 12) contribute least. Inspection of a number of trees shows large differences between individual trees, suggesting that the Types (and in particular Types 3 and 4) cannot reliable be distinguished from each other based on the available variables.

Table (5.2): Confusion matrix for a Random Forest on data excluding Set Phrases and nouns without number

	zero	**a**	**the**
zero	**346**	14	32
a	14	**571**	2
the	47	4	**747**

5.3 Learning simulation

In Figure (5.1) we display the results of the computational simulation of learning the English article system with our incremental, error-correction learning algorithm, but this time using referentiality and definiteness as the critical learning cues. While *the* is strongly associated with Type 2 (referential definites), Types 3 (referential indefinites) and 4 (non-referentials) consistently appear together and do not differentiate between any of the articles; neither is Type 1 (generics) a strong cue for any of the articles.