SupMat_4: Models including all data and Set Phrases only

4.1. All data

A Random Forest of 500 trees run on all data, including Set Phrases, gives an overall error rate of 8.77%; most of these errors are made on zero (20.9% incorrectly classified), next is *the* (6.5% incorrectly classified), and most accurate is *a* (3.4% incorrectly classified). A single tree shows a comparable performance and is presented in Figure (4.1).

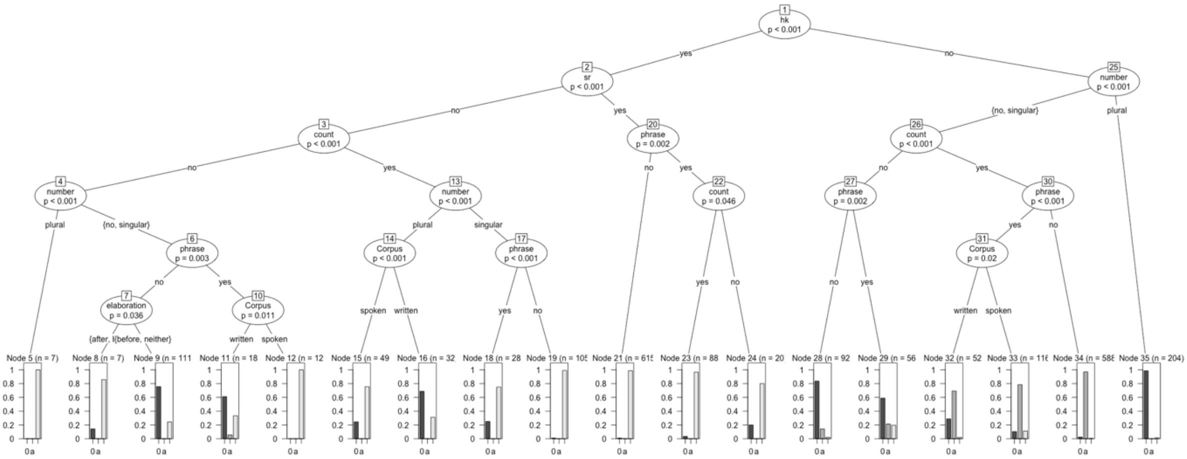

Figure (4.1): Classification tree for all articles data

In terms of variable importance, HK triggers the largest mean decrease in the Gini index (decrease = 521), followed by Number (decrease = 161) and SR (decrease = 132). Count follows at a respectable distance (decrease = 105) before the mean decreases suddenly drop down to much smaller values for Elaboration (decrease = 19), Corpus (decrease = 15) and Set Phrase (decrease = 12).

Table (4.1): Confusion matrix for one Tree on all data

	zero	**a**	**the**
zero	**428**	26	59
a	41	**697**	18
the	34	2	**895**

Table (4.2): Confusion matrix for a Random Forest on all data

	zero	**a**	**the**
zero	**398**	42	63
a	22	**700**	3
the	45	18	**909**

4.2 Set phrases only

With an 20.53% error rate the prediction accuracy is down significantly compared to the model without set phrases. HK remains the strongest predictor, with HK+ predominantly leading to *the*, but encountering competition from zero in a small number of cases. However, HK- does not have a clear preference: only in a small minority of cases that concern countable plural nouns is zero the clear winner. In case of singulars, *a* competes with zero, and at times even with *the*.

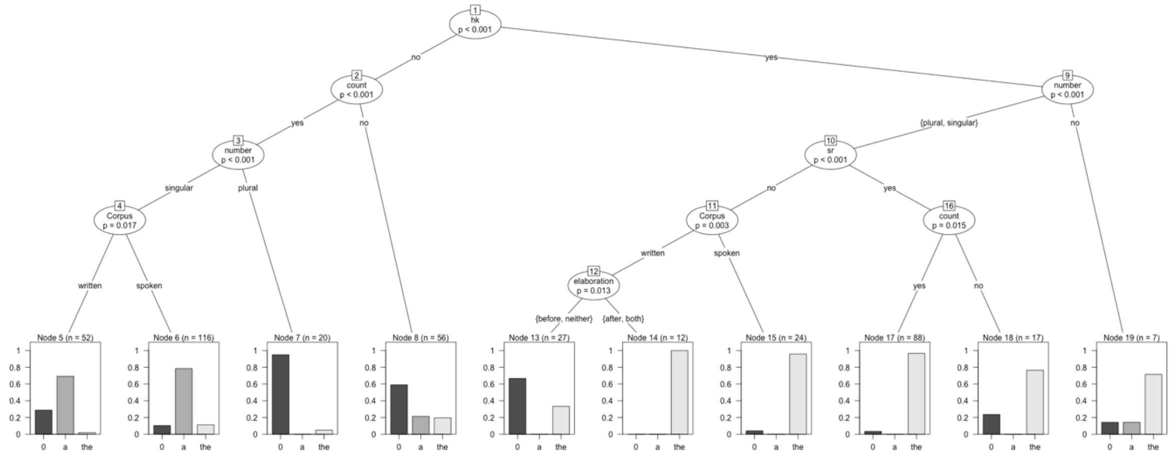

Figure (4.2): Classification tree for articles in set phrases

In terms of variable importance, HK triggers the largest mean decrease in the Gini index (decrease = 68.7), followed after a drop in values by SR (decrease = 22.9), and then Count (decrease = 16.6), Number (decrease = 13.2) and Elaboration (decrease = 10.5). Corpus contributes least (decrease = 8.4).

Table (4.3): Confusion matrix for one Tree on Set Phrases

	zero	**a**	**the**
zero	**70**	12	21
a	27	**127**	14
the	9	1	**138**

Table (4.4): Confusion matrix for a Random Forest on Set Phrases

	zero	**a**	**the**
zero	**60**	28	18
a	12	**127**	1
the	13	14	**146**