timation of the results of NLI when random splitting, particularly for groups of contributors that were presented with very different topics – e.g. students from Asia vs. students from Europe. We have analyzed the distribution of essays into topics using the essay titles, and observed that contributions from Europe (which are our focus) have similar distributions across the featured topics.

Growing the language tree To grow the language tree from the language vectors built from the English essays, we use a variation of the UPMGA – Unweighted Pair Group Method with Arithmetic Mean – algorithm. Starting with the language vectors V_{L_j}, we compute the distance between each pair of vectors using a distance metric algorithm. At each step we choose the closest pair (L_a, L_b) and combine them in a subtree, then combine their corresponding sub-collection of essays, and build the language vector for the "composite" language $L_{a,b}$, and compute its distance to the other language vectors.

4.2 Results

We test whether etymological information surfaces as native language interference that is detectible through the tasks of native language identification and reconstruction of the language family tree. Table 3 shows results on the multi-class classification of essays according to the native language of the author, in the form of F-score average results using SVM classification in 5-fold cross-validation (using Weka's SMO implementation[3] with polynomial kernel and default parameters). The baseline corresponds to the language distribution in the dataset. We use as additional comparison point another set of features used to reconstruct the language family tree – the (closed-class) word and POS 3grams Nagata and Whittaker (2013), such as *the NN of; a JJ NN; the JJ NN*. We build all such patterns for the data, and keep the top 1000 by overall frequency.

Adding etymological features that capture the distribution of etymological ancestors for each essay led to improved results for all languages, varying from a non-significant improvement of 0.2% point for Russian, to a significant and high 5.3% improvement for German. Using only words, the accuracy is 73.2%, which increases marginally to 73.7 when etymology information is added. Using a full complement of standard features – word, lemma and character ngrams (n=1..3) (built following (Lahiri and Mihalcea, 2013)) – gives an average accuracy (over 5 fold cross-validation) of 85.7%. Adding etymology does not lead to improvements when added to this set.

[3] http://www.cs.waikato.ac.nz/ml/weka/

Language	Baseline	Etym.	Patt	both
Bulgarian	8.52%	32.4	51.7	54.3
Czech	6.85%	21.9	53.4	54.4
Dutch	7.41%	11.7	50.4	51.1
French	9.78%	30.0	58.8	62.9
German	12.31%	45.4	47.4	52.7
Italian	11.04%	34.3	66.3	67.3
Norwegian	8.93%	35.5	57.0	59.3
Polish	10.28%	42.5	59.5	62.1
Russian	7.78%	12.7	46.9	47.1
Spanish	7.07%	24.6	57.9	59.6
Swedish	10.00%	23.1	44.8	45.7
Accuracy		31.7	54.2	56.3

Table 3: 5-fold cross-validation F-scores and accuracy for language classification

Despite the rather low results when etymology is used on its own for language identification, the cumulative evidence leads to a language family tree that closely matches the gold standard (Figure 2). The tree on top is the gold standard cf. (Nagata and Whittaker, 2013; Crystal, 1997). The tree is grown by computing the euclidean distance between pairwise vectors, and then iteratively grouping together the closest vectors at each step as described in Section 4.1.

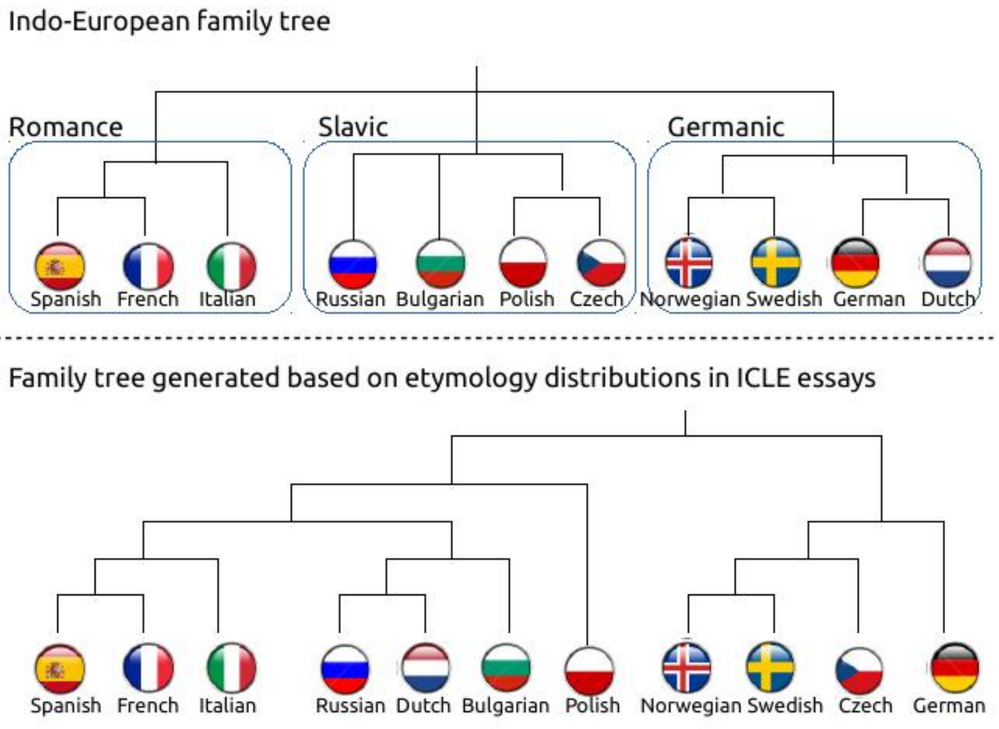

Figure 2: Language family trees – the gold standard and the automatically generated one

The two wrongly placed languages in our language family tree are Czech and Dutch. Czech