$n_{il_k}$ is the number of words with etymological ancestor $l_k$ in the $i$-th essay, and $N_{i,etym}$ is the number of words with etymological information in essay $i$.

**Language vectors** For each subcollection corresponding to one (student native) language $L_j$, we build the language vectors by averaging over the essay vectors in the subcollection:

$$V_{L_j} = \langle p_{L_j l_1}, ..., p_{L_j l_m} \rangle$$

$$\textit{where} \quad p_{L_j l_k} = \frac{\sum_{lang(e_i)=L_j} p_{il_k}}{|\{e_i | lang(e_i)=L_j\}|}$$

$p_{L_j l_k}$ is the proportion of etymological ancestor language $l_k$ in all essays whose author has as native language $L_j$.

The essay and language vectors are filtered by removing etymological languages whose corresponding values in the language vectors are less than $10^{-4}$.

## 4 Experiments

We investigate the strength of the etymological "fingerprint" of individual and collective essays written by non-native speakers of English, through two tasks – native language identification and language family tree construction. Towards this end, we work with a collection of essays written by contributors whose native language is an Indo-European language. The dataset is described in Section 4.1. For etymological information we rely on an etymological dictionary, described briefly in Section 3.1. Data modeling and the experiments conducted are described in Section 3.2.

### 4.1 Data

We used the ICLE dataset (Granger et al., 2009), consisting of English essays written by non-native English speakers. We filter out those that were written by people whose mother tongue is not from the Indo-European family (i.e. Chinese, Japanese, Turkish and Tswana). Table 1 shows a summary of the data statistics, including the number of words for which we have found ancestors in the etymological dictionary used. The corpus consists entirely of essays written by students. Two types of essay writing are present: argumentative essay writings and literature examination papers. Table 2 displays a list of topics in the corpus. The essays should be at least 500 words long and up to 1,000, and contain all the spelling mistakes made by their authors.

Following Nagata and Whittaker (2013), who also built the Indo-European family tree based on n-grams composed of function words and open-class parts of speech, essays that do not respect one of the following rules are filtered out: (i) the writer has only one native language, (ii) the writer has only one language at home; (iii) the two languages in (i) and (ii) are the same as the native language of the subcorpus to which the essay belongs. Table 1 shows a summary of the data statistics after filtering, including the number of words for which we have found ancestors in the etymological dictionary used.

| *Native language* | *# essays* | *# tokens (with etym)* |
|---|---|---|
| Bulgarian | 302 | 226,407 (149,151) |
| Czech | 243 | 226,895 (148,391) |
| Dutch | 263 | 264,981 (169,040) |
| French | 347 | 256,749 (161,136) |
| German | 437 | 259,967 (170,056) |
| Italian | 392 | 253,798 (165,500) |
| Norwegian | 317 | 238,403 (156,764) |
| Polish | 365 | 263,223 (172,319) |
| Russian | 276 | 259,510 (167,938) |
| Spanish | 251 | 225,341 (139,565) |
| Swedish | 355 | 224,948 (146,143) |

Table 1: Statistics on the subset of ICLE dataset used.

| | |
|---|---|
| 1 | Crime does not pay. |
| 2 | The prison system is outdated. No civilized society should punish its criminals: it should rehabilitate them. |
| 3 | Most university degrees are theoretical and do not prepare students for the real world. They are therefore of very little value. |
| 4 | A man/woman's financial reward should be commensurate with their contribution to the society they live in. |
| 5 | The role of censorship in Western society. |
| 6 | Marx once said that religion was the opium of the masses. If he was alive at the end of the 20th century, he would replace religion with television. |
| 7 | All armies should consist entirely of professional soldiers : there is no value in a system of military service. |
| 8 | The Gulf War has shown us that it is still a great thing to fight for one's country. |
| 9 | Feminists have done more harm to the cause of women than good. |
| 10 | In his novel Animal Farm, George Orwell wrote "All men are equal: but some are more equal than others". How true is this today? |
| 11 | In the words of the old song "Money is the root of all evil". |
| 12 | Europe. |
| 13 | In the 19th century, Victor Hugo said: "How sad it is to think that nature is calling out but humanity refuses to pay heed. "Do you think it is still true nowadays ? |
| 14 | Some people say that in our modern world, dominated by science technology and industrialization, there is no longer a place for dreaming and imagination. What is your opinion ? |

Table 2: Topics in the ICLE dataset.

The suitability of the dataset above for NLI was questioned by Brooke and Hirst (2012). They have shown that the fact that the corpus consists of sets of essays on a number of topics causes an overes-