1 Corpora

The data used in this study was drawn from nine different corpora. Of these, three contained only native texts, four only nonnative texts, and two texts of both types. Table 1.1 shows the number of tokens contributed by each corpus. A token is a unit parseable by the Stanford parser, the large majority of which are simply words but which also include punctuation and the genitive suffixes *'s* and *'*. As can be seen in the table, the two classes of texts (native and nonnative) are very closely matched in size. Furthermore, the number of samples in each class is identical, 321, giving a total of 642 instances or cases. All classification methods used in this study operated on these same 642 instances.

The following corpora contributed native samples: the Brown University Standard Corpus of Present-Day American English subcorpus of letters-to-the-editor and editorials (BROWN), the International Corpus of English-Hong Kong (ICE-HK), the Michigan Corpus of Upper-level Student Papers (MICUSP), the Open American National Corpus (OANC), and the International Corpus of English-Canada (ICE-CAN). MICUSP and ICE-CAN contributed nonnative samples as well, and the remainder of the nonnative texts came from the International Corpus of Learner English, Spanish Subcorpus (SPICLE), the Santiago University Learner Corpus (SULEC), and the Written Corpus of Learner English (WRICLE). One additional student paper supplied by Missouri State University's English Language Institute rounded out the nonnative samples. All nonnative samples were written by individuals whose first language was Spanish and who were judged, by the compilers of the corpora, to be advanced English learners. Many of the individuals had a language in addition to English and Spanish. In the cases of the SULEC and WRICLE corpora, both of which were compiled at Spanish universities, a large number of the learners spoke other Romance languages in addition to Spanish, in particular Catalan and Galician. Many of the samples in the ICE-HK corpus were written by individuals whose second language was Cantonese, and a number of the contributors to ICE-CAN had some French as well. Any sample written by an individual who knew a Germanic language (other than English) was

not included.

Table 1.1: Corpora Composition

Corpus	Tokens Native	Tokens Nonnative
BROWN	57,809	0
ICE-HK	59,674	0
MICUSP	163,218	29,897
MSUELI	0	538
OANC	84,0522	0
SPICLE	0	216,879
SULEC	0	39,254
WRICLE	0	96,247
ICE-CAN	25,225	2,070
Total	389,978	384,885