1 Conclusion

Building an accurate classifier to distinguish between native and nonnative texts is very feasible with today's parsing technologies. Even a text lacking syntactic errors can be classified as nonnative through the identification of overused or underused grammatical constructions. This tendency for the learner to use certain grammatical constructions more or less frequently than a native English speaker is likely also part of what make nonnative texts appear as such to native readers, though human readers undoubtedly take semantics into consideration as well when making such a determination. As shown in the previous chapter, it is feasible to develop a tool based on the classification principles shown here which would help a learner to become aware of the features of his or her writing that identify it as nonnative.

This study considered three approaches to gathering grammatical features for use in the classification process. The first of these used grammatical relations of all types generated by the Stanford Parser. By the nature of grammatical relations, this approach looked at a broad range of grammatical constructions, including such diverse elements as the use of the inflected genitive and the use of phrasal verbs. This breadth undoubtedly accounts for its high accuracy. The next two approaches delved more deeply into particular aspects of grammar. The second classification method explored using grammatical dependencies and lexemes to explore the use of pronominal and lexical verbal arguments. The third method extracted information on verbal constructions from parse trees. These three methods are only examples of the types of features sets that can be extracted from text. Other possible feature set could be derived from vocabulary (perhaps focusing on closed class words), syntactic complexity (e.g., the depth of parse trees, nesting of various types of phrases, and so forth), use of contractions, use of indirect objects (e.g., whether prepositions are used or not), the use of copular verbs (e.g., considering words such as *seem*, *become* as compared to *be*), and so forth. The greater the variety of feature sets used, the more useful a learner's tool would be.

While this study focused on L1-Spanish learners of English, none of the feature sets are actually specialized for that particular class of nonnative English. By retraining on different nonnative corpora, the classification systems used here could be extended to any type of nonnative English. Undoubtedly some feature sets will prove more or less useful for other L1s, but little change would be needed. A learner's tool might be tuned for a particular L1-background so as to provide the most informative feedback, but such tuning would not require changes to the fundamentals of the system.

This system could be extended to other applications as well. For instance, instead of training binary (i.e., native vs. nonnative) classifiers, one could train on a number of classes of texts to develop a classifier that would identify the particular L1 of the writer. Such a tool could be used to detect plagiarism by learners, or could even used in the field of forensics. Also, the system could be used to distinguish human-generated text from machine-translated text, for similar ends.

In the past, most approaches towards the automated analysis of learner texts have been focused on identifying malformed syntax. In general, such systems are of great use to beginning and intermediate learners, but of little use to advanced learners. It has been shown here that automated analysis of the language of advanced learners is very feasible, and has applications both in education and elsewhere. It seems very likely that in the coming years there will be greater focus on this field, both in academia and in the commercial sector.