1 Introduction

Fluency is the goal of most language learners. While mastering the spoken language is usually the primary goal, becoming a proficient writer is of importance to many learners, particularly those whose learn languages for professional or academic reasons. Although it is generally less obvious than in the case of spoken language, most learners will reach a stage where they produce grammatically-correct language, but which is still identifiable as nonnative. At this stage, further improvement of their language is no longer a matter of error correction, but of changing subtleties in their production which may not be apparent to the learner. A learning tool that could analyze written language and give feedback to the user on which aspects of their language need to be modified to reach a native level would be useful. Such a tool would only be possible if there existed an automated system that could analyze language, classify it as native or nonnative, and give the reasons for the classification. This paper describes such a system.

Most computer-based ESL tools have been for beginning and intermediate learners, focusing on grammatical mistakes. Such tools analyze text and offer suggestions on aspects such as correct article usage or noun pluralization. In many ways, these tools are very similar to the grammar checkers found in most modern word-processing packages, though those tend to be targeted towards native speaker, or towards no particular group of users. In general, these tools are not of any use in the analysis of syntactically well-formed texts.

The system explored in this study relies on modern language parsing systems. Probabilistic parsers are capable of generating grammatical parse trees of texts in a number of languages, particularly English, with a high degree of accuracy. Such systems, in particular the Stanford Parser, used in this study, are often capable of presenting additional syntactic information in the form of dependency graphs, which indicate various relationships between words. This study uses automatically generated parse trees and

dependency graphs to identify grammatical features. The relative frequencies of each feature are calculated, and this information is used to train automatic classifiers, which are then tested on other text samples.

Chapter 2 of this paper is a literature review and a description of similar and relevant existing technologies. Chapter 3 describes the corpora from which the text samples were gathered. The fourth chapter gives a brief description of the parsing and classification systems used, and describes the general experimental setup used in this study. The next three chapters describe the three experiments performed. Each of these experiments deals with a different set of grammatical features. These chapters describe the methods by which each set of grammatical features is extracted from the output of the Stanford Parser, and analyze the effectiveness of classifiers trained on these features. Two of these experiments further break down feature sets into smaller subsets to ease in the analysis of their efficacy. In these chapters, the decision trees generated by the classifiers are also analyzed to identify the linguistic significance of the features used in the trees. Finally, in the eighth chapter, this paper proposes a design for a learner's tool based on this system, and suggests other possible applications as well.