

# A System To Distinguish Native and Nonnative Written English

Philip White

April 5, 2012

## **1 Introduction**

## **2 History and Similar Efforts**

Automated document classification has been a highly productive area of study in the fields of text mining and machine learning. Initial attempts at document classification, and automated classification in general, were in the form of expert systems. Expert systems consist of human-compiled rules which are applied to a particular instance, such as a document, and make decisions about that instance based on occurrences of predefined features [Clifford et al. 1983]. Such a system relies entirely on the knowledge of the human expert who compiles the rules. The goal of such a system is to automate the classification process, but not necessarily to classify any more accurately than might an expert human. While such a system can theoretically approach the accuracy of the human who compiled the rules, it is incapable of exceeding that accuracy, and it comes at a high cost in terms of development time. Early attempts to solve these problems, in particular the latter, explored the automated creation of rule sets. Apté et al. [1994] was an early study into the effectiveness of computer-generated rule sets. Because of the enormous time savings afforded by such

a system, and because of the potential of such systems to exceed the accuracy of humans, classification systems using computer-generated rule sets have become very common in commercial products (consider, for instance, spam email filtering [Cormack 2008]) and continue to be a very active area of research. The reader who desires a more detailed account of the history of document classification and a survey of its current (as of 2004) state is referred to Berry [2004].

One of the goals of the current study is to explore how a text classification system can be used to develop a software tool which provides suggestions to language learners on how they can improve their written language. In light of that, it is worth exploring similar existing tools. Grammar checkers are one such tool. A great deal of work has been done in the field of automatic grammar checking and, in many ways, this a very mature field of computer science. English grammar checkers have been available in commercial word processors for nearly two decades [Vernon 2000], and there has been much progress, recently, on grammar checkers targeted towards language learners. One such effort is Microsoft Research's ESL Assistant. One element of this system, described in Gamon [2010], focuses on identifying common learner errors in article and preposition usage. The system uses maximum entropy classifiers in a novel approach to determine whether a particular location in a text should have an article or preposition and, if so, which specific article or preposition. The system looks at word boundaries within the text. For each boundary, it gathers features from the six words to the right and left of the boundary. It then applies the first of two classifiers to this feature set. The *presence classifier* determines the probability that the boundary is an appropriate location for an article or preposition. Then, if a location with a high probability is identified, a second classifier, the *choice classifier*, is applied to this set of features to identify the most likely appropriate articles or prepositions for use in that location. The system then compares the results of the classifiers with any articles or prepositions actually used in that location, and makes suggestions to the user. One benefit of this approach is that the classifiers need only be trained on native texts, which are more

plentiful than nonnative texts. The study also explains how meta-classifiers can be used to improve upon this approach. For this, a second error detection system is used as well, based on language models. Language models are systems that assign a probability to a sequence of words, indicating how likely it is for that sequence of words to occur in natural language. Gamon then uses both of these error detection systems as part of a meta-classifier trained on a corpus of learner texts in which all preposition and article errors have been marked. He finds that the meta-classifier provides considerably better accuracy than either system used alone.

Lee and Seneff [2006] explore a generative approach to grammar correction for language learners. In this system, one sentence or utterance is processed at a time. A number of permutations on the input are produced by modifying articles and preposition, inflecting nouns and verbs, and modifying auxiliary verbs. A language model is then used to choose a small subset of these permutations as candidates for correction. The study used both human evaluators and automatic evaluators to determine whether the system was producing output more appropriate than the learner's input.

Wagner et al. [2007] compare two automated grammatical error detection systems. Though the approaches they describe are not specific to learner English, the nature of the systems do make them well suited for such. The first approach they describe uses a *precision grammar*. A precision grammar is a set of grammatical rules (see Ch. 4.2) designed to parse only strictly grammatical language. This is in contrast to the grammars used in general purpose parsers, which tend to be designed to accommodate common errors. In this approach, the precision grammar is used to identify errors (i.e., unparseable passages). Though this study does not focus on providing corrections, the authors note that it is possible to include special "mal-rules" to identify specific types of malformed syntax. The second approach Wagner et al. explore uses part of speech n-grams (considering only the parts of speech of sequences of words) with language models to identify grammatical errors.

The studies highlighted here are but a small sample of the work that has been done in the areas of automated grammatical error detection and correction, but they are representative of the two most common approaches used: a shallow approach, using language models and n-grams to identify errors; and a deep approach, using parsers. The technologies used to generate corrections are more varied, and the reader interested in a broader review is referred to Lee [2009].

Little or no research has been done towards a system based on distinguishing native English and well-formed nonnative English, nor directly towards a system that offers grammatical suggestions to improve already grammatical learner English. It is likely that a system using language models could be adapted towards these ends, though this is not the approach taken here. Most of the work done on providing automatic grammar evaluation has been focused on the routine grammatical errors made by native speakers, or on the errors typical of novice and intermediate English learners. There appears to be little in the way of automated tools for advanced learners who wish to bring their writing skills to near-native levels of proficiency.

### **3 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 3.1 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 of each class is the same, 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 (MSUELI) 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 3.1:** 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

## 4 Parsing and Classification

### 4.1 Choice of Language

With very few exceptions, the code written in support of this thesis was done in Clojure, a dialect of LISP designed to work on top of the Java Virtual Machine (JVM). The choice of language was simple: a heavy dependence on the Stanford Parser and the WEKA package, both written in Java, necessitated a JVM-based language. The slowness of Java's compile/debug cycle eliminated that language as an option, leaving a handful of possible languages, from which Clojure was chosen for its speed, functional style, and elegance.

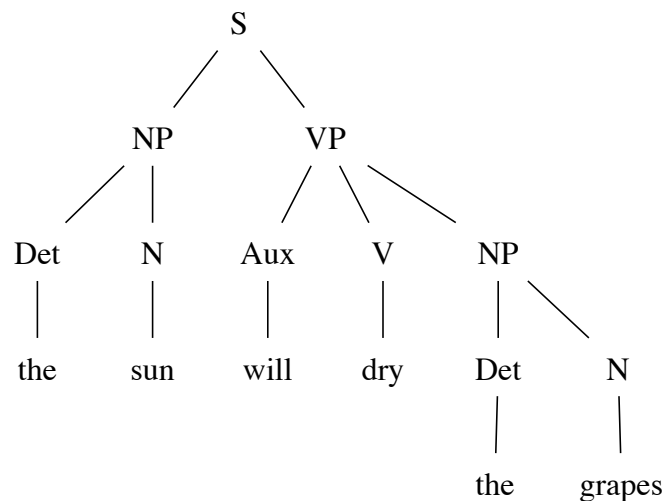
### 4.2 Parsing

The Stanford Parser software package, version 1.6.7, configured with the included probabilistic context-free grammar (PCFG) [Klein and Manning 2003], was used to generate all syntactic parse trees and grammar dependency graphs. A detailed description of PCFGs is beyond the scope of this paper, and the reader who desires such is referred to Booth and Thompson [1973]. Central to the PCFG is the context-free grammar (or *phrase structure grammar*). A context-free grammar consists of a number of rules, each of which describe the various possible compositions of a particular type of phrase. For instance, Figure 4.1 shows a very simple English grammar consisting of five rules. Rule (a) indicates that a sentence (S) is composed of a noun phrase (NP) followed by a verb phrase (VP). Rule (b) says that a noun phrase is composed of a noun (N) preceded by an optional determiner (Det),

- a. S  $\rightarrow$  NP VP
- b. NP  $\rightarrow$  (Det) N (PP)
- c. VP  $\rightarrow$  (Aux) V (NP) (AdvP)<sup>n</sup>
- d. PP  $\rightarrow$  P NP
- e. AdvP  $\rightarrow$   $\left\{ \begin{array}{l} \text{Adv} \\ \text{PP} \end{array} \right\}$

**Figure 4.1:** Simple Phrase Structure Rules. [Akmajian et al. 2010, Ch. 5.3]

and followed by an optional prepositional phrase (PP). Rule (c) says that a verb phrase contains an optional auxiliary verb (Aux), followed by a verb (V), followed by an optional noun phrase, and then by zero or more adverbial phrases (AdvP). A prepositional phrase is defined by rule (d) as being a preposition plus a noun phrase, and an adverbial phrase is defined by rule (e) as being either an adverb or a prepositional phrase. Again, these are only example rules and cover just a small subset of English grammar. Also, the symbols used by the Stanford parser are not necessarily the same used in these example rules. Using these rules, the sentence *the sun will dry the grapes* can be represented as the tree shown in Figure 4.2. Generating this tree from the original sentence requires a certain knowledge of



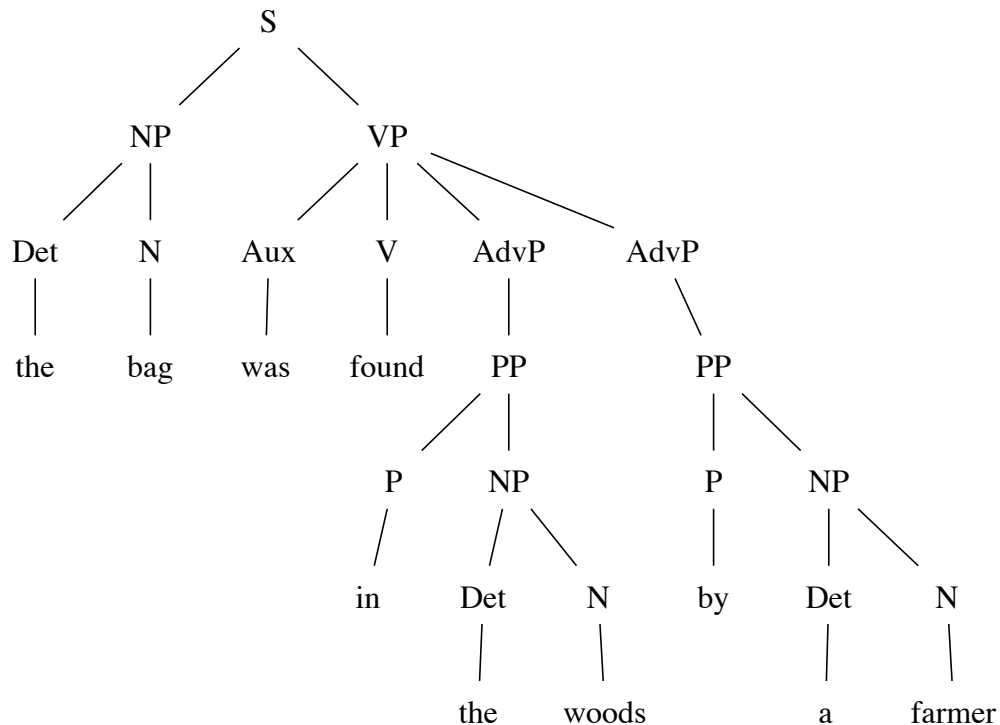
**Figure 4.2:** Parse of *the sun will dry the grapes* Generated from Rules in Figure 4.1.

the parts of speech of the words in the sentence. To some extent, it is not necessary to have all of this information. In fact, if any of the open-class<sup>1</sup> words in this example are replaced with nonsense words, most people would have no trouble parsing it, and would generate a tree with the same structure as that in Figure 4.2. Similarly, the Stanford parser successfully parses this sentence even if all of the open-class words are replaced with nonsense words. This does not always hold true for more complex sentences. Also, replacing the

<sup>1</sup>*Open-class words* are those belonging to a class which allows the admission of new words, such as nouns and verbs. *Closed-class* words belong to class which do not generally admit new entries, such as prepositions and articles. Open-class words are often called *content* words and closed-class words *function* words. In this example *sun*, *dry*, and *grapes* belong to open-classes and *the* and *will* to closed-classes.

closed-class words with gibberish makes the sentence much more difficult to parse. Also, replacing one word with another word of a different part of speech (e.g. *the sun will dry the warmly*) produces a sentence that is difficult to parse for both human and computer. Parsers guess the part of speech of unknown words the same way humans do, by choosing whatever part of speech produces the most common valid phrase structure. Parsers use a similar method for dealing with ambiguities.

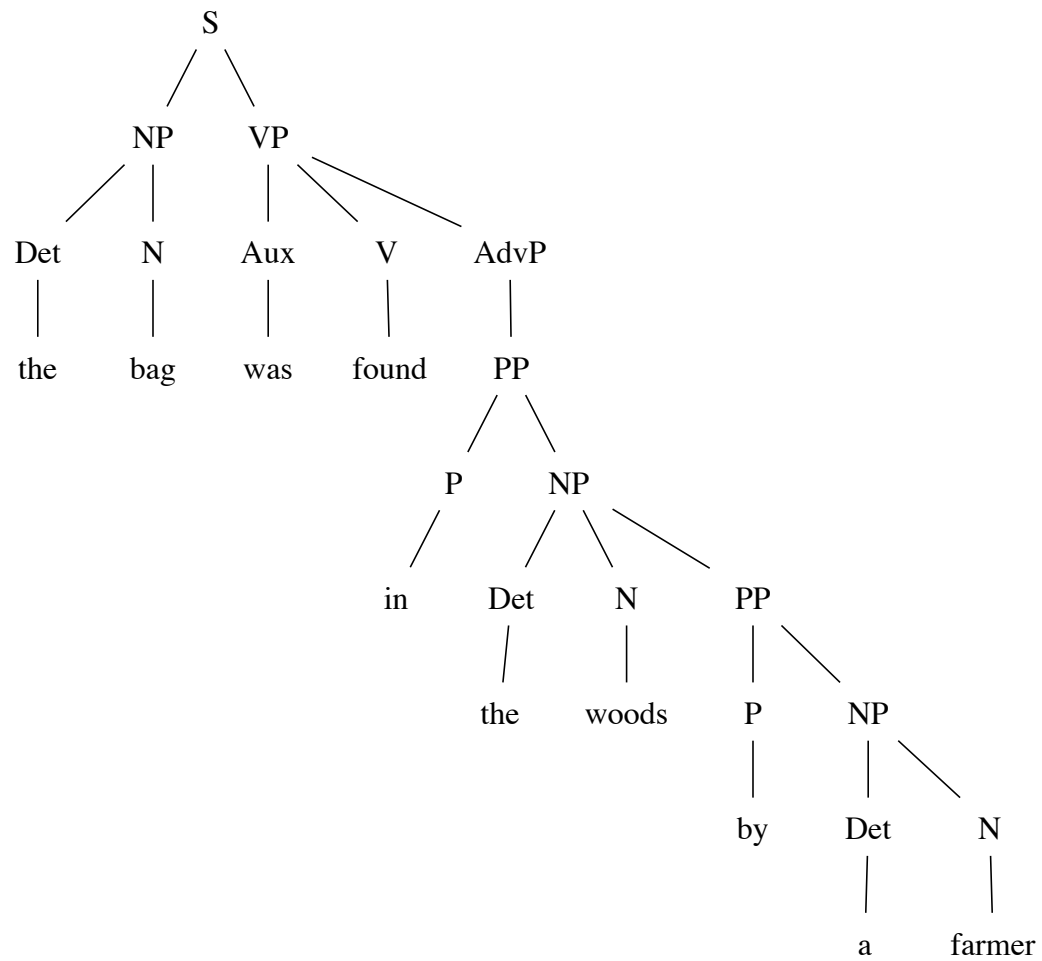
Consider the sentence *the bag was found in the woods by a farmer*. Using the phrase rule structures from Figure 4.1, one could generate either the parse shown in Figure 4.3 or that shown in Figure 4.4. In both parses, the PP *in the woods* is considered an adverbial



**Figure 4.3:** One Possible Parse of *the bag was found in the woods by a farmer* Generated from Rules in Figure 4.1.

phrase modifying the verb. However, the second PP can be parsed either the same way, or it can be placed within the NP containing *woods*. In other words, the sentence can either mean that a farmer found the bag, or that the bag was found in the woods and those woods were located near a farmer. The first interpretation is the one that most English speakers would





**Figure 4.4:** Another Possible Parse of *the bag was found in the woods by a farmer* Generated from Rules in Figure 4.1.

choose based on the semantic content of the two options. A rule-based parser, however, is unable to attempt semantic interpretation, and a purely rule-based parser might output all possible interpretations. A PCFG-based parser, however, will choose the parse that is *statistically* most likely to be correct. Such a parser requires a large database of correct parses with which it can make probability measurements. The Stanford parser, for instance, uses the Penn Treebank, a corpus of parsed English texts [Marcus et al. 1993]. Out of the various possible parses, the one that a probabilistic parser will choose is the one that has a structure, or a substructure, that is most common among parses in the database. The parser does not include only the phrase structure in its consideration, but certain words as well. For instance, given *the bag was found in the woods by a farmer*, the Stanford Parser will generate a parse very similar to Figure 4.3. However, when given the sentence *the bag was found under the bridge over the stream*, which has the same two possible parses, it chooses the other, semantically correct one. At first impression, one might think that the parser had located a semantic connection between *bridge* and *stream*, and had used this to choose the correct parse. However, it turns out that it is actually the prepositions (a group of closed-class words) that cause the difference in parsing. For instance, the sentence *the bag was found under the woods over a farmer*, is parsed like Figure 4.4 and *the bag was found in the bridge by the stream* is parsed like Figure 4.3. Hence the parser uses both the phrase structure and the closed-class words, but generally not the open-class words, when choosing the best parse.

### 4.3 The Tests

The crux of this project was the design and creation of a suite of tests, each of which identifies a number of closely-related grammatical characteristics of the text samples. These tests operate on the output from the Stanford parser — parse trees and grammatical dependencies. As output, they generate training or test cases to be used by the Weka classifier. Each of these cases consists of multiple attributes, corresponding to grammatical

features, each with continuous values indicating the relative frequency (probability) of that particular feature. For a case with  $n$  attributes where the number of occurrences of the grammatical feature associated with the  $i$ th attribute is  $g_i$ , the value  $f_i$  for that attribute is given by  $g_i / \sum_{i=1}^n g_i$ . For example, one test measures the relative frequencies of the various tense/aspect/voice combinations of finite verbs. English has twenty-four such combination, so the case generated by this test has twenty-four attributes.

In addition to the attributes, each case has a class which can be *es* or *en*, indicating that the class is associated with a text sample written by an L1-Spanish speaker or by a native English speaker, respectively. For training cases, the classes are known beforehand and are assigned to the cases manually. For test cases, the classes have missing values until such values are determined by a classifier, as discussed in the following section.

## 4.4 Classification

The Weka machine learning package, version 3.6 [Hall et al. 2009] was used to create, train and test classifiers based on the cases discussed above. Of the many classifiers included in the Weka package, two were used in this study: J48, which is Weka’s implementation of the C4.5 classifier [Quinlan 1993] and the RandomForest classifier, which is based on the random forest algorithm described by Breiman [2001]. The former is useful for its highly readable decision trees, which clearly indicate which attributes are involved in the classification and their roles. In later sections of this paper are found linguistic explanations for why these particular attributes should be useful in classification.

### 4.4.1 C4.5

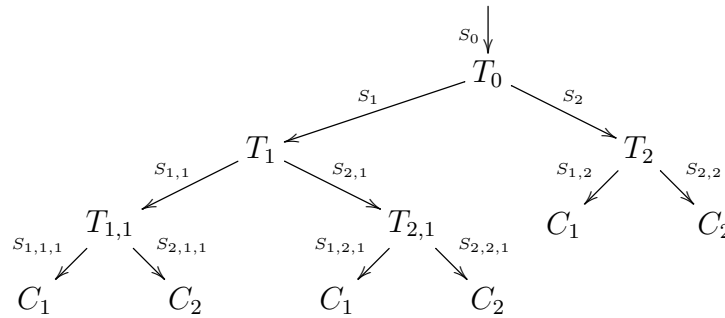
This section describes the C4.5 algorithm as it applies to this project. That is to say, C4.5 can deal with a number of circumstances that do not arise here. What is described here is a version of the C4.5 algorithm that is restricted to continuous attribute values and two class values, and which does not permit missing attribute values. That said, the C4.5 algorithm

consists of two phases, *tree construction* and *tree pruning*.

In the tree construction phase a decision tree is built which successively performs binary partitioning of a set of training cases. Consider a full binary tree where each edge represents a set of cases and each non-terminal node a partitioning operation, as shown in Figure 4.5. These partitioning operations take one set, represented by the parent edge, and divide it into two subsets, the daughter edges. The root node operates on an initial set  $S_0$ , and a leaf node simply indicates that its parent edge is a set consisting of cases of a single class. Let the first partitions of  $S_0$  be called  $S_1$  and  $S_2$ , where  $S_1 \cup S_2 = S_0$  and  $S_1 \cap S_2 = \emptyset$ , and of  $S_1$  let them be called  $S_{1,1}$  and  $S_{2,1}$  and so forth. Likewise, let the partitioning operation that operates on a particular set be designated by  $T$  with the same subscripts as that set.

Partitioning is performed by applying a binary test to each case within  $S$ , the set to be partitioned, and dividing the set based on the results. Each test considers a single attribute  $A$  and compares the value of that attribute,  $V_A$ , to a threshold value,  $V_C$ . All cases where the  $V_A \leq V_C$  will be put into one subset and all other cases into the other.

The attribute and threshold value for a particular test is determined using what Quinlan [1993] calls the *gain ratio criterion*, which is calculated as follows. If the probability of randomly drawing a case of class  $C_1$  from a set  $S$  is  $p_1$  and of drawing a case of the other class is  $p_2$ , where  $p_2 = 1 - p_1$ , then the average amount of information needed to identify



**Figure 4.5:** Decision Tree Showing the Partitioning of a Set of Training Cases  $S_0$  into Subsets Each Consisting of Elements of a Single Class.

the class of a case in  $S$  can be defined in terms of entropy as

$$\text{info}(S) = -p_1 \cdot \log_2(p_1) - p_2 \cdot \log_2(p_2). \quad (1)$$

A similar measure can be applied to the two partitions  $S_1$  and  $S_2$  created by applying the partitioning test  $T$  to  $S$ . The entropy after partition is given by taking a weighted sum of the entropy of the two sets as

$$\text{info}_T(S) = \frac{|S_1|}{|S|} \cdot \text{info}(S_1) + \frac{|S_2|}{|S|} \cdot \text{info}(S_2). \quad (2)$$

The decrease in entropy, expressed as a positive value (an information gain), due to partitioning  $S$  using the test  $T$  is then

$$\text{gain}(T) = \text{info}(S) - \text{info}_T(S). \quad (3)$$

Maximizing this gain can be and, in ID3, the predecessor to C4.5, was used as measurement of test fitness. However, in the more general case of C4.5, where one test can partition a set into more than 2 subsets, using this gain criterion to choose tests favors tests that partition sets into numerous subsets. To mitigate this, Quinlan added another factor to the criterion, the *split info*, which for this special case is given by

$$\text{split info}(T) = -\frac{|S_1|}{|S|} \cdot \log_2 \left( \frac{|S_1|}{|S|} \right) - \frac{|S_2|}{|S|} \cdot \log_2 \left( \frac{|S_2|}{|S|} \right). \quad (4)$$

Then the fitness of a test  $T$  can be measured using

$$\text{gain ratio}(T) = \frac{\text{gain}(T)}{\text{split info}(T)}. \quad (5)$$

It should be noted that in this special case where partitioning operations are always binary, the gain ratio criterion favors tests that split  $S$  into disparately sized sets, as split info is at

its maximum (unity) when  $|S_1| = |S_2|$ .

In choosing a test  $T$ , the C4.5 algorithm tries each attribute  $A$  from the set  $S$  of cases to be partitioned. For each, it orders the cases in  $S$  on the value of  $A$ . If the values of  $A$  corresponding to this ordered set are  $\{v_1, v_2, \dots, v_m\}$ , then any threshold between some  $v_i$  and  $v_{i+1}$  will result in the same partitions. From this it can be seen that the total number of possible partitions is  $m - 1$ . The algorithm tries all such partitioning schemes, measuring the gain ratio of each. When an optimal attribute and corresponding partitioning scheme has been chosen, the algorithm then chooses a threshold value that will produce this result. Again, to partition  $S$  into two sets where the values for  $A$  are  $\{v_1, v_2, \dots, v_i\}$  and  $\{v_{i+1}, v_{i+2}, \dots, v_m\}$ , respectively, a threshold value  $v_C$  must be chosen such that  $v_i \leq v_C < v_{i+1}$ . For this, it chooses the largest value for  $A$  from the entire training set  $S_O$  that does not exceed the midpoint of this range.

Partitioning operations are generated until a tree is produced which will divide the training set into subsets each of which contains cases of a single class. At this point a classification tree has been generated which, if fed a test case, will categorize that case into its correct class based on its attributes. In general this tree will be quite large, and, though it classifies the test cases with perfect accuracy<sup>2</sup>, it will tend to be overspecialized for the test cases and will not perform as well on other cases. Both of these problems are ameliorated in the final stage, known as *pruning*.

Pruning is performed by applying a statistical test to each non-terminal node to predict whether the tree would perform better in general if the subtree rooted at that node were replaced with a leaf node. To make this prediction, a certain confidence level  $p$  must first be chosen, the default for C4.5 being 25%. Then, for any pair of numbers  $N$  and  $E$ , where  $N$  is a number of classifications and  $E$  the number of those classifications which are incorrect, a value  $U_p(E, N)$  can be calculated by taking the upper limit of the confidence

---

<sup>2</sup>It is possible that the C4.5 algorithm will be unable to generate a tree that perfectly classifies the training cases. This can happen if the training set has two cases that have identical attributes and values but different classes. In this case no classifier would be able to distinguish the two.

interval for the binomial distribution that describes the probability of observing  $E$  events in  $N$  trials at a level of confidence  $p$ . The error of a leaf over  $N$  classifications can then be estimated by computing  $N \cdot U_p(E, N)$ . The C4.5 algorithm uses this to estimate the error of a subtree by calculating this quantity for each leaf in the tree, using as  $N$  the number of training cases that are classification at that node, and as  $E$ , the number that are classified incorrectly (generally zero), and summing these values. It then compares this estimated subtree error with the estimated error of a leaf replacing the entire subtree. In this case,  $N$  will be equal to the sum of the  $N$ s of the leaves in the deleted subtree, and  $E$  will generally be some nonzero value. If the predicted error of the leaf is less than that of the subtree, the replacement is made.

## 4.5 Random Forest

Random forest algorithms work by a certain number of decision trees, each trained on a random subset of the training case attributes. Classification is performed by applying all decision trees to a case and choosing the most popular class, a process generally called *voting*. This type of classifier was introduced by Breiman [2001], and can be generalized for use with any type of decision tree. The Weka implementation, RandomForest, uses a custom classifier for its constituent trees, the RandomTree classifier. This classifier is very similar to C4.5 without pruning, overfitting generally not being a concern with random forests [Breiman 2001]. The RandomForest classifier uses a customizable number of trees and attribute subset size. If the latter is not specified, the value defaults to  $\text{floor}(\log_2(n) + 1)$  where  $n$  is the number of trees. In this study, 100-tree forests were used (resulting in trees trained on 7 attributes), except when this would have resulted in an attribute subset size greater than the total number of attributes.

## 5 Grammatical Relations

The simplest classification approach used in this study considered the relative frequency of different grammatical relations. For this approach, the governor and the dependent of the dependencies were ignored, with only the relation itself being used.

Each data set instance contained attributes corresponding to dependency relations. The Stanford parser system, in its default configuration, does not generate the *punct* or punctuation dependency, which connects punctuation symbols to a key element in the associated clause. Since English punctuation is broadly similar to Spanish punctuation, aside from some stark differences such as Spanish’s inverted question and exclamation marks, which should be apparent to even the beginning learner, it did not seem to useful to activate this dependency. Additionally, the *abbrev* or abbreviation dependency was removed. This dependency marks the definition of an abbreviation, as in the example given by de Marneffe and Manning [2008], “Australian Broadcasting Corporation (ABC),” where the dependency would be *abbrev*(Corporation, ABC). This dependency has little to do with grammar, and thus was ignored for the purposes of this study. Having excluded these two dependencies, each data set instance contained 58 numerical attributes, one for each relation.

For each attribute  $A_r$  corresponding to the relation  $r$ , the corresponding value was the floating point number  $n_r/n_t$ , where  $n_r$  and  $n_t$  were the number of occurrences of the relation  $r$  and the total number of relations in the text, respectively. A C4.5 decision tree classifier trained on these instances produces the decision tree shown in Figure 5.1, employing 15 different relations. The full names for these relations are shown in Table 5.1. At each terminal node of the tree there is an integer or pair of integers in parentheses. These values indicate the number of the training cases that were categorized (correctly or not) at that node and the number of cases incorrectly categorized, this latter value only being shown when greater than zero. For any given test node, one can identify one branch as the predominately *en* branch and the other as the *es* branch. For test nodes where one or both branches lead to terminal nodes, this is trivial, as the terminal nodes themselves label



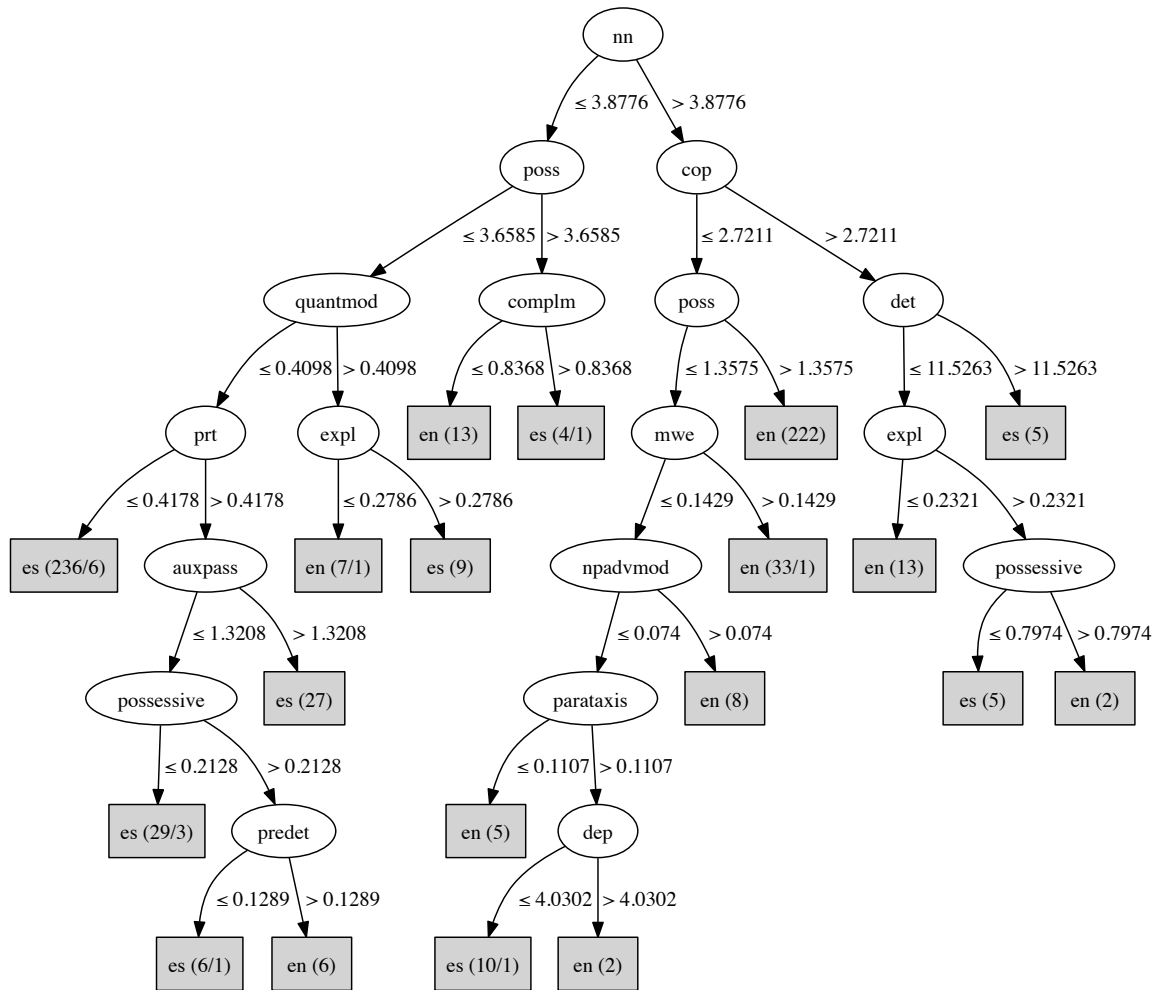
the branches. For any other test node, the branches can be identified by summing up the number of test cases at the terminal nodes of that branch. For instance, the root test node, which considers the relation *nn*, divides the training set of 642 cases into a subset of 337 cases, associated with the left branch, and another subset of 305 cases, associated with the right branch. Looking at the left branch, it can be seen that of these 337 cases, 301 of them are nonnative, i.e. of the class *es*, and only 36 are native. This indicates that this is a predominately nonnative branch. Conversely, the right hand branch consists of 285 native cases and only 20 nonnative cases, making it the native branch. This allows one to say, for instance, that cases with low occurrences of the *nn* relation tend to be nonnative samples. The following subsections explore the linguistic reasons why these relations are so useful in making such categorizations.

**Table 5.1:** Relation abbreviations

<i>auxpass</i>	passive auxiliary
<i>complm</i>	complementizer
<i>cop</i>	copula
<i>dep</i>	unclassified dependency
<i>det</i>	determiner
<i>expl</i>	expletive
<i>mwe</i>	multi-word expression
<i>nn</i>	noun compound modifier
<i>npadvmod</i>	noun phrase as adverbial modifier
<i>parataxis</i>	parataxis
<i>poss</i>	possession modifier
<i>possessive</i>	possessive modifier
<i>predet</i>	preconjunct
<i>prt</i>	phrasal verb particle
<i>quantmod</i>	quantifier phrase modifier

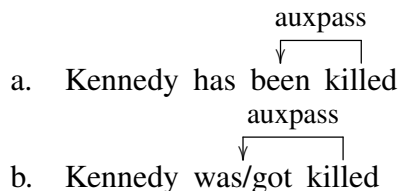
## 5.1 Passive Auxiliary

The passive auxiliary dependency *auxpass* marks an auxiliary verb which carries the passive information of the clause. In general, a parsed sample of text will contain one such dependency for every passive clause, and so a high relative frequency of this relation indi-



**Figure 5.1:** C4.5 Decision Tree Employing Relative Frequency of Dependency Relations.

cates heavy usage of the passive voice. Figure 5.2 illustrates this dependency.



**Figure 5.2:** The Dependencies *auxpass*(killed, been) and *auxpass*(killed, was/got). [de Marneffe and Manning 2008].

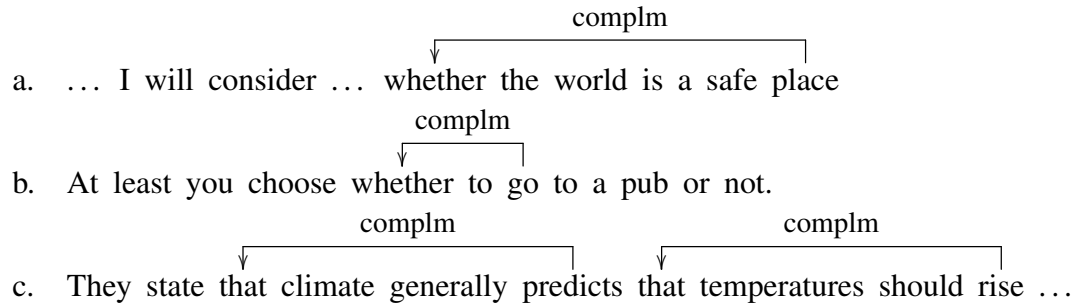
The decision tree in Figure 5.1 uses the *auxpass* attribute once, dividing the training set into cases with higher frequencies, which get classified immediately as nonnative, and cases with lower frequencies, which undergo additional testing. It is important to note that the large majority of this second set of cases (35 out of 41) are ultimately categorized as nonnative as well. So while high relative frequencies of the *auxpass* attribute are associated with learners, low frequencies are associated with both learners and native speakers. This indicates that the use of the passive is not a strong indicator of the nativeness of the text. Likely a slightly more aggressive pruning factor would have resulted in the elimination of the *auxpass* node.

## 5.2 Complementizer

A complementizer is a word that signals the beginning of a clausal complement. The Stanford Parser recognizes the complementizers *that* and *whether* as shown in Figure 5.3. The governor of a complementizer dependency is the root of the clause, which is generally a verb or, in the cause of copular clauses, the subject complement. The dependent is the complementizer itself.

Whitley [1986] points out that while English tends to allow the deletion of complementizers introducing clausal complements in the object position, Spanish generally does not, as shown in the following examples:

(5.1) a. *I say that he'll do it.*



**Figure 5.3:** The Dependencies *complm*(place, whether), *complm*(go, whether) *complm*(predicts, that), and *complm*(rise, that). Nonnative Samples from WRICLE (a and c) and SULEC (b).

b. *I say he'll do it.*

c. *Digo que lo hará.*

d. *\*Digo lo hará.* (Whitley 1986, p. 278)

Butt and Benjamin [2004, 33.4.6] explain that this rule is occasionally broken, but generally only in two situations: business letters and nonstandard speech, and when the complementizer *que* appears close to other uses of the word *que*. Since these are restricted cases, it is reasonable to conclude that there would be L1-transfer in the construction of clausal complements, leading L1-Spanish learners to have some preference for (5.1a) over (5.1b), particularly considering that they are both perfectly valid constructions.

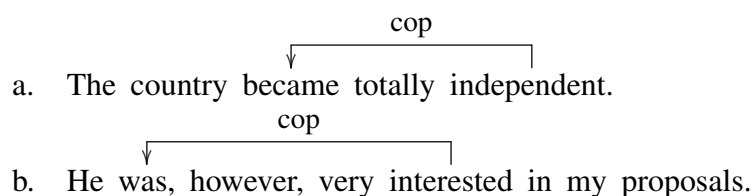
In a study on differences in complement clause usage between native and nonnative English speakers, Biber and Xeppen [1998] draw a number of conclusions relevant to the current study. First, they consider when native speakers omit the complementizer *that* and conclude that it is rarely omitted in academic prose and in opinion and descriptive essays. Since the vast majority of the corpus samples (both native and nonnative) fall into these categories, this provides encouraging evidence that the differences in complementizer usage identified by the classifier are not due to idiosyncrasies in the samples. Next, while considering four different groups of L1 speakers (French, Spanish, Chinese, and Japanese), Biber and Xeppen find that all groups show similar levels of *that* omission, and in general

these levels of omission are lower than the levels found in comparable types of native texts. They also find that L1-Spanish speakers use complement clauses, with and without omission of the complementizer, more often than either native speakers or the other groups of learners.

The decision tree shown in Figure 5.1 uses the *complm* dependency once, and classifies cases with lower occurrences of *complm* as native, and larger occurrences as nonnative, without further testing. Because this dependency only indicates the presence of a complement clause if it has a complementizer, the higher frequency among the learners may be due to either low rates of dropping the complementizer, or to high rates of complement clause usage. As shown above, both phenomena have linguistic backing, and very likely both are at play.

### 5.3 Copula

The copula or *cop* dependency marks copular verbs. This dependency takes as its governor the complement of the copular clause and the verb itself as the dependent. Figure 5.4 shows examples of two different copular verbs. However, the Stanford Parser does not recognize



**Figure 5.4:** The dependencies *cop*(independent, became) and *cop*(interested, was). [Quirk et al. 1985, Ch. 2.15, 2.16].

all copular clauses as such. In particular, copular clauses followed by adverbials are not identified with the *cop* dependency, for instance:

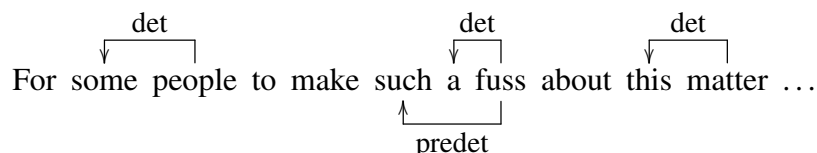
(5.2) *I have been in the garden.* [Quirk et al. 1985, Ch. 2.16]

The decision tree in Figure 5.1 contains one node which uses the *cop* dependency. This node divides the training set into two subsets such that the first, associated with lower

frequencies of *cop*, contains a smaller percentage of nonnative texts, and the second, with higher frequencies of the attribute, contains a larger percentage. The simplest explanation for this is simply that copular clauses tend to be the simplest type of clauses in language and thus are favored by learners. A detailed treatment of both of these points can be found in a study by Hinkel [2003], which investigates syntactic simplicity in L1 and L2 academic texts.

## 5.4 Determiner and Predeterminer

The determiner or *det* dependency connects a determiner to the NP it modifies, with the determiner being the dependent and the head of the NP the governor. Similarly, the *predet* dependency marks a predeterminer. Figure 5.5 shows examples of both determiners and a predeterminer in one sentence fragment. By the analysis of Quirk et al. [1985, Ch. 5.10],



**Figure 5.5:** The dependencies *det*(people, some), *predet*(fuss, such), *det*(fuss, a) and *det*(matter, this). Nonnative sample taken from SULEC.

a determiner is an element which modifies a NP, precedes any adjectives modifying the NP, and which expresses the type of reference made by that NP. Adjectives, on the other hand, indicate the attributes of a NP. Quirk et al. divide determiners into three classes: predeterminers, central determiners, and postdeterminers. Postdeterminers, which include quantifiers such as *many* and *few*, and both cardinal and ordinal numerals, are identified by the Stanford parser using relations not found in the decision tree in Figure 5.1. It should not come as a surprise that the *predet* relation marks predeterminers, but it is worthy noting that the *det* relation marks only *central* determiners. Perhaps the most common central determiners are the articles *the*, *a*, and *an*; but this class of words also includes a number of other words, many of which have separate roles as pronouns, such as *this*, *that*, *some*, and

so forth. The predeterminers consist of words which generally precede core determiners and which include certain words which modify quantity, such as *all*, *both*, *double*, *half*, etc. and others more difficult to define: *such*, *what*, and so forth. Note that the Stanford Parser only parses predeterminers as *predet* dependencies if they appear before a *det* dependency. Otherwise they get parsed as *det* dependencies.

Figure 5.1 shows one use each of these relations. The test node that considers *det* splits the training cases into two sets: cases with high frequencies of *det* which are immediately classified as nonnative, and cases with low frequencies, the majority of which are ultimately classified as native. For the *predet* test, both subsets are immediately classified, with low frequencies as nonnative and high frequencies as native. The implication then is that nonnative users overuse central determiners and underuse predeterminers. More specifically, nonnative speakers use predeterminers before central determiners with lower frequency than do native speakers.

For the most part, central determiners, especially articles, are closely parallel in English and Spanish. There are differences in article usage; in particular, the definite articles are frequently used in Spanish where no article is used in English, and, conversely, definite articles are frequently used in English where no article is used in Spanish. The rules governing these uses can certainly be trying for learners, but one would expect advanced learners to have mastered these concepts. Perhaps more difficult a concept, and one which may account for the overuse of central determiners by learners, is where English can express the same concept with a definite article or without an article at all. Consider the following examples:

(5.3) a. *The tiger has four legs.*

b. *Tigers have four legs.*

c. *Los tigres tienen cuatro patas*

Though, in the right context, (5.3a) could refer to a particular tiger, it could also refer to

tigers in general, as (5.3b) does. The former sounds a bit formal, or perhaps antiquated, while the latter is the more current. In Spanish, however, an article is generally required for generic reference, as shown in (5.3c) [Whitley 1986, Ch. 8.3.3]. It is possible that the L1-Spanish learner of English, being accustomed to (5.3c), would choose the grammatically correct (5.3a) instead of the also grammatically correct but more common (5.3b). Very likely there are other reasons behind the high frequency of the *det* relation in nonnative texts, but another study is needed to fully explore this issue.

The low frequency of the *predet* relation may simply be a matter of the learner preferring syntactically simple constructions. As mentioned above, the *predet* dependency is used when a predeterminer precedes a central determiner. This means that for every *predet* dependency, the parser has found a location with multiple determiners appearing together. By the very definition of complexity, such a construction is more complex than a construction with a single determiner, and thus likely to be avoided by the learner. Another likely source of this underuse may be that many of English's predeterminers do not have common predeterminer equivalents in Spanish. For instance, fractions in English can generally be expressed in two slightly different ways:

(5.4) a. *He did it in a third the time it took me.*

b. *He did it in a third of the time it took me.* [Quirk et al. 1985, Ch. 5.19]

The latter of these examples is not parsed as a predeterminer by the Stanford Parser. Except for a few common fractions, Spanish generally follows a format similar to (5.4b) [Butt and Benjamin 2004, Ch. 10.10]. Spanish also tends to use constructions similar to (5.4b) to express multipliers:

(5.5) *El aire contiene el doble de óxido de nitrógeno que en Washington.* [Butt and Benjamin 2004, Ch. 10.14]

whereas English would use a simple predeterminer:

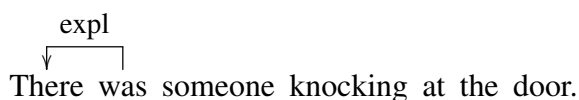


(5.6) *The air contains double the nitric oxide as Washington.*

This difference may encourage learners to use periphrastic constructions in English (e.g. *twice as much, two times the amount*) which are not parsed as predeterminers by the Stanford Parser.

## 5.5 Expletive

An existential *there* and the copular verb associated with it are connected with the expletive or *expl* relation, as shown in Figure 5.6. This relation is used twice in the decision tree,



**Figure 5.6:** The dependency *det*(was, there). [Quirk et al. 1985, Ch. 3.29]

both times associating higher frequencies with nonnative texts and lower frequencies with native texts. Spanish has a similar construction to the English existential *there + be*, using a 3rd-person singular form of the verb *haber* in any of its possible non-progressive forms. Spanish, which is a pro-drop language, does not use or permit a dummy subject analogous to the English *there*, nor does the verb *haber* agree in number with what follows except in very informal speech [Butt and Benjamin 2004, Ch. 30]. Otherwise, the existential *there* presents little difficulty for the L1-Spanish learner of English. The high rate of use is likely due to this: learners resort to existential *there* frequently because it is a “safe” expression, one they can generate correctly with little effort.

## 5.6 Multi-Word Expression

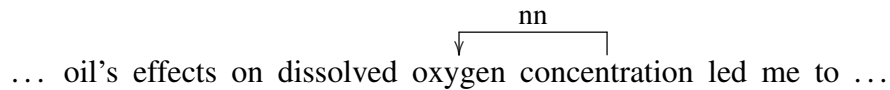
The Stanford typed dependency manual [de Marneffe and Manning 2008] defines multi-word expressions as being two or more words that are used together as a single unit such that the relationship between them is difficult to define. In the version of the Stanford parser used here, only the following expressions are considered multi-word expressions: *rather*

*than, as well as, instead of, such as, because of, in addition to, all but, due to.* As can be seen in the decision tree, higher rates of use are indicative of a native speaker. Since these tend to be idiomatic, or at least syntactically complex, it seems reasonable that they would be avoided by learners.

## 5.7 Noun Compound Modifier

Noun-noun compounds (NNCs) are marked with the relation *nn*. The governor of this dependency is the rightmost noun in the compound, and the dependent will be one of the nouns to the left. Note that since all dependencies only deal with pairs of words, a compound consisting of more than two nouns would be indicated by multiple dependencies, all sharing a common governor. Figure 5.7 demonstrates this dependency.

**Figure 5.7:** The dependency *nn*(concentration, oxygen). Native sample taken from MICUSP.



The *nn* relation occupies an important place in the decision tree, being the root test node and thus the relation with the highest information content. Summing leaf values will show that the *nn* node splits the training cases into a largely native set and a largely nonnative set, corresponding respectively to high and low frequencies of the dependency. That the underuse of noun compounds show be indicative of an L1-Spanish learner of English is understandable considering that Spanish has a much more restrictive system of noun compounding. While Spanish does have NNCs, they are far less common than in English and are not particularly productive [Piera 1995]. Most commonly, expressions in English using NNCs are translated into Spanish using the preposition *de*. Consider, for instance:

(5.7) *un traje de baño*

*a suit of bath*

‘*a bathing suit*’ [Butt and Benjamin 2004, Ch. 34.7]

English can also use the preposition *of* to express possession in a construction parallel to the Spanish *de*-possessive. Often times, the English speaker can switch between a NNC and the *of*-construction with little change in meaning, as in:

(5.8) a. *the mountain peak*

b. *the peak of the mountain*

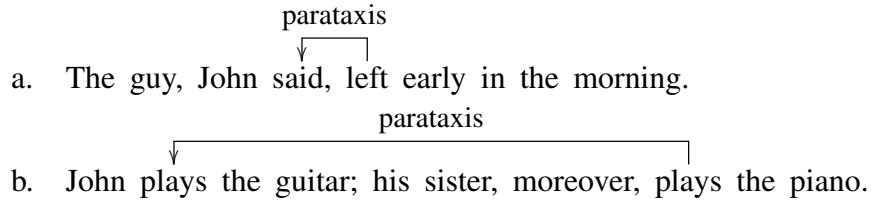
Considering this flexibility in English and the paucity of NNCs in Spanish, it seems very likely that L1-Spanish learners of English avoid NNCs in favor of the *of*-construction.

## 5.8 Noun Phrase as Adverbial Modifier

The *npadvmod* dependency marks where a NP is used like an adverbial modifier. In general it covers five types of constructions: phrases indicating measure (e.g. *the director is 65 years old*), certain phrases which expresses financial information (e.g. *IBM earned \$5 a share*), reflexive pronouns used for emphasis (e.g. *the silence itself is significant*), and a handful of other usages difficult to categorize. All of these tend to be idiomatic and syntactically complex, which would account for the use of this relation in the decision tree, where cases with high frequencies are categorized as native and with low frequencies as nonnative.

## 5.9 Parataxis

The *parataxis* relation ties the main verb of a clause to another element, generally a parenthetical or something appearing after a colon or semicolon. Figure 5.8 shown one example of each of these types of parataxis. The types of constructions marked by the *parataxis*

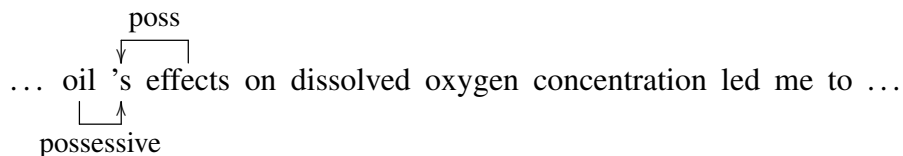


**Figure 5.8:** The dependencies *parataxis*(left, said) and *parataxis*(plays, plays). Taken from de Marneffe and Manning [2008] and Quirk et al. [1985, Ch. 13.7], respectively.

relation are syntactically complex, and so one might imagine that they would be underused by learners of English. However, the use of this relation in the decision tree indicates that learners actually use this more often than native speakers. There is little literature discussing parataxis in learner English nor do there appear to be any obvious qualities of Spanish that might explain this overuse. This analysis is particularly difficult considering that the available data does not indicate which of the two types of parataxis are being overused.

## 5.10 Possession and Possessive Modifiers

Inflected genitive constructions are marked by two dependencies: *poss*, which ties the head of a NP (the governor) to a genitive inflectional suffix ('s or '), indicating that the governor is the possessed element; and *possessive*, which connects a noun to its own genitive inflectional suffix. These two dependencies are illustrated in Figure 5.9. The *poss* dependency can also have as its dependent a possessive determiner such as *its* or *their*. In this type of construction, the *possessive* dependency is not used.



**Figure 5.9:** The dependencies *poss*(effects, oil) and *possessive*(oil, 's). Native sample taken from MICUSP.

These relations are both used twice in the decision tree. In all four instances, cases with

high frequencies tend to end up being classified as native, and cases with low frequencies as nonnative. Spanish lacks an inflected genitive, instead tending to use the preposition *de* to express possession, as was discussed in Section 5.7. Much like NNCs, the inflected genitive is usually translated into Spanish using a *de*-construction, as shown in:

(5.9) *una chica joven de vaqueros y chaqueta de hombre*

*a girl young of jeans and jacket of man*

*‘a young girl in jeans and a man’s jacket’* [Butt and Benjamin 2004, Ch. 34.7]

Also much like NNCs, it is often possible to use the *of*-construction in place of an inflected genitive:

(5.10) a. *the mountain’s peak*


b. *the peak of the mountain*

The reasonable conclusion then, is that learners of English are using the *of*-construction in place of the inflected genitive, resulting in fewer occurrences of the latter.

## 5.11 Phrasal Verb Particle

The phrasal verb particle relation (*pvt*) ties the head word of a phrasal verb to its particle as shown in Figure 5.10. The decision tree in Figure 5.1 contains this relation once. Relative frequencies of 0.4178% or less lead to the categorization of a text as nonnative, whereas larger values lead to a subtree. It can be seen that a very high percentage, 36.8%, of the training cases terminate at the left, or nonnative, branch of this test node, suggesting that this relation contributes a great deal of useful information to the categorization process.

...the reduction of superfluous proteins will free up resources ...



**Figure 5.10:** The dependency *pvt*(free, up). Native sample from MICUSP.

Phrasal verbs are multiword verbs consisting of a core word, which can generally stand alone as a distinct verb in other circumstances, and a preposition-like particle appearing

after, though in many cases not immediately after, the primary word [Celce-Murcia and Larsen-Freeman 1999]. These verbs appear to be rare in world languages, with few non-Germanic languages containing such constructions [Celce-Murcia and Larsen-Freeman 1999]. Liao and Fukuya [2004] conduct a review of the literature on phrasal verb avoidance in English language learners, starting with Dagut and Laufer [1985], a study which concluded that L1-Hebrew learners of English do avoid these verbs. They further asserted that the reason for this was syntactic differences between Hebrew and English, though others have questioned their bases for this assertion [Liao and Fukuya 2004]. The review continues with Hulstijn and Marchena [1989], who investigated the claims of Dagut and Laufer by applying the same data gathering techniques to a group of English learners whose first language was Dutch, a language which also uses phrasal verbs. These authors hypothesized that the first language has little influence on whether the learner avoids phrasal verbs in English. Contrary to their expectations, they found that the Dutch speakers did not avoid phrasal verbs in English, suggesting that L1-interference is, at least in part, the source of phrasal verb avoidance. Finally, the review cites the study of Laufer and Eliasson [1993], which performed a very similar study as Hulstijn and Marchena, but with native Swedish speakers, and drew much the same conclusions.

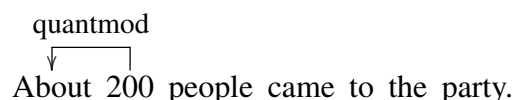
In their own study, Liao and Fukuya investigate L1-Chinese learners of English, and cautiously concluded that the syntactic features of Chinese lead to the avoidance of phrasal verbs in English. A later study, Alejo González [2010], uses the Spanish and Swedish subcorpora of ICLE, along with the British National Corpus (BNC), a corpus of native written English, to perform a quantitative study of phrasal verb usage. They found that the L1-Swedish learners used phrasal verbs 69% as often as the native speakers, and the L1-Spanish learners used phrasal verbs 45% as often. These numbers would seem to indicate that the syntax of the learner's L1 is an important, but not the only, contributing factor to phrasal verb avoidance.

Regardless of the reasons behind the avoidance of phrasal verbs shown by L1-Spanish

learners, Alejo González [2010] demonstrates that it is a reality of learner English. Considering this, it is not surprising that the C4.5 algorithm uses the *prt* relation with such success in the categorization process.

## 5.12 Quantifier Phrase Modifier

The Quantifier Phrase Modifier (*quantmod*) relation marks adverbs that modify certain determiners. In general, this relation is only used when the determiner being modified is a numeral. Most other determiners, including quantifying determiners such as *double* or *half*, would require the use of the *advmod* relation. Thus in Figure 5.11, if the number 200 were replaced with the determiners *half the*, *about* would become the dependent of an *advmod* dependency instead of the *quantmod* dependency.



**Figure 5.11:** The dependency *quantmod*(200, about). Taken from de Marneffe and Manning [2008].

The decision tree shows that high frequencies of this dependency are associated with native texts. Because the scope of this dependency is limited, and because Spanish grammar does not differ markedly from English grammar in the usage of adverbs as modifiers of determiners, little can be said in the way of linguistic analysis other than to suggest that these dependencies are less common in learner texts due to the complexity of the syntax that generates them.

## 5.13 The Unclassified Dependency

The final relation, *dep*, is used in any dependency that cannot be more exactly resolved by the parser, whether due to malformed grammar, parser limitations, or any other reason. Due to the nebulous nature of this dependency, no meaningful linguistic analysis is possible.

## 5.14 Classification Accuracy

Twenty fold cross-validation was used to test the real-world accuracy of the data. There being 642 cases in the data set, thirty-two unique cases were held out at a time and classified using a C4.5 classifier trained on the remaining 610 cases. This produced a correct classification rate of 89.72% with a mean absolute error (MAE) of 0.1139 and a  $\kappa$  value of 0.7944. Using a random forest classifier gave better results; performing 20 fold cross-validation on a 100 tree classifier where each tree was trained on six random features yielded 94.24% accuracy with MAE = 0.1707 and  $\kappa = 0.8847$ . Table 5.2 gives the confusion matrices for these two classifiers.

**Table 5.2:** Accuracy results for C4.5 and 100 tree Random Forest classifiers using 20 fold cross-validation on data set of 642 cases.

	C4.5		R. Forest	
	es	en	es	en
Classified as → es	309	12	291	30
en	25	296	36	285
% Correct	89.72		94.24	
MAE	0.1139		0.1707	
$\kappa$	0.7944		0.8847	



## 6 Argument Structure

In general, every verb in English takes one or more arguments, with a subject argument being required in normal speech and writing. The Stanford NLP system marks the arguments of verbs using the dependencies shown in Table 6.1. In the majority of non-copular clauses, the governors of these dependencies are the core verbs. In copular sentences, the governor is generally the subject complement (i.e. the argument generally appearing after the verb which is equated with the subject), though in the case of copular sentences with clausal subjects, the Stanford parser chooses the copula to be the governor. Figure 6.1 show examples of these dependencies.

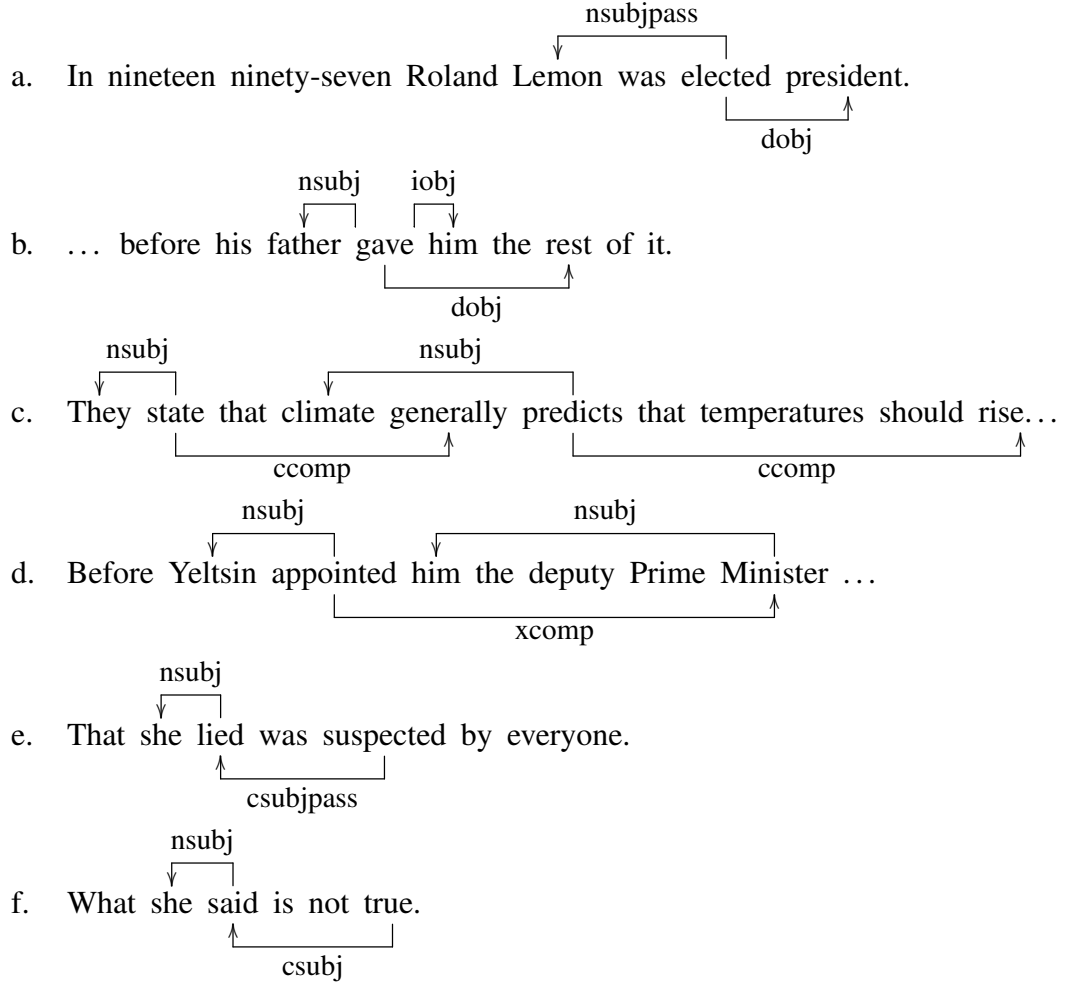
**Table 6.1:** The Dependencies Used to Identify Verbal Arguments.

ccomp	Clausal Complement
csubj	Clausal Subject
csubjpass	Passive Clausal Subject
doj	Direct Object
iobj	Indirect Object
nsubj	Nominal Subject
nsubjpass	Passive Nominal Subject
xcomp	Open Clausal Complement

Du Bois [2003] provides hints that verbal argument structure may differ between native and nonnative speakers. That study explores when speakers use lexical NPs rather than pronominal NPs as verbal arguments. Considering only native speakers but looking at a number of different languages, Du Bois finds that there is a very strong tendency for speakers to use no more than one lexical argument per finite clause (Du Bois's One Lexical Argument Constraint) and to avoid placing an argument in the subject role of transitive sentences (the Non-Lexical A Constraint<sup>3</sup>). He makes the case that these rules hold true for a number of world languages. However, in presenting data to show that several languages abide by the Non-Lexical A Constraint, he also shows that there are large differences between languages in the likelihood of a new argument appearing in the direct object and

---

<sup>3</sup>Du Bois [2003] uses the letters **A**, **I**, and **O** to refer to the subject, indirect object, and direct object arguments of a transitive verb, and **S** to refer to the sole argument of an intransitive verb.



**Figure 6.1:** Examples of the Dependencies Listed in Table 6.1. *a*, native sample from ICE-CAN; *b,c,d*, native samples from MICUSP; *e,f* from de Marneffe and Manning [2008].

intransitive subject roles [Du Bois 2003, Table 2.5]. A new argument is one presenting information that has not yet been presented in the discourse. A new argument generally cannot be a pronoun, as pronouns typically refer to something presented earlier in the discourse. Not all lexical arguments are new arguments of course, but it stands to reason that the factors affecting the ratio of new to old arguments will also affect the ratio of lexical to pronominal arguments. The data he gathers show that in English, on average, 21% of new arguments are found in intransitive clauses, versus 28% for Spanish, and 79% are found in direct object roles, versus 71% for Spanish. A  $\chi^2$  analysis shows that these differences are statistically significant considering his sample sizes ( $\chi^2 = 2.244$ ,  $df = 2$ ,  $p = 0.326$ ),

though admittedly at a relatively low confidence level (70%). As an aside, of the five languages for which data is presented in that table, English and Spanish are the most similar in new argument distribution. Of the three other languages considered, French, Hebrew, and Sakapultek, all show a greater probability of finding a new argument in the intransitive subject role, and a lower probability of finding one in the direct object role, than either Spanish or English. Seeing that there may be differences between Spanish and English in how new arguments are distributed among the various argument roles, and noting that other languages show such differences as well, it is worth investigating whether L1-Spanish learners of English differ in their usage of lexical arguments as compared to native speakers.

To investigate the feasibility of classifying language based on argument structure, Stanford dependency graphs were analyzed to identify 18 different types of finite clauses: intransitives with and without clausal subjects, copular clauses with and without clausal subjects, simple transitives with and without clausal subjects and objects, ditransitives with and without clausal subjects, complex transitives with and without clausal subjects, passives of simple transitives with or without clausal subjects, and passives of complex transitives with or without clausal subjects and complements. Then, for each of these 18 different types of clauses, all possible combinations of lexical and non-lexical arguments were considered, yielding 80 different attributes in total. This system, while providing a generous amount of data for the classifiers, yielded decision trees that were difficult to interpret linguistically. Fortunately, it was possible to generate from these attributes three much smaller and more coherent sets of attributes which, when combined and used to train a classifier, provided comparable accuracy.

The first of these attribute sets consisted of just three attributes, each corresponding to a particular verb valency (i.e. the number of arguments). The values associated with these attributes were the percentage of all clauses which used that number of arguments. The next attribute set consisted of four attributes, named *zero*, *one*, *two*, *three*, with corresponding values indicating what percentage of finite clauses had that number of lexical arguments.

To avoid lengthy periphrasis, it is convenient to call this metric *lexical argument density*. The last attribute set consisted of seven attributes corresponding to types of argument roles. The values associated with these attributes were the percentage of lexical arguments found in that type of argument role. The seven types were intransitive, transitive, passive, and copular subjects, indirect and direct objects, and subject complements. This latter category included both the complement of copular clauses and what is usually the third argument in a complex transitive (e.g. *deputy Prime Minister* in Table 6.1d). This set of attributes will be referred to as the *lexical argument role* set.

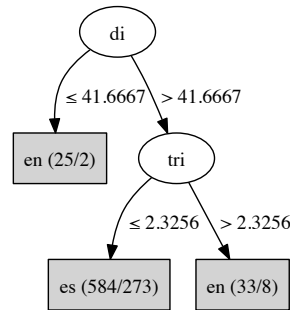
To determine what was a lexical argument and what was not, a list of pronominal forms, largely taken from Celce-Murcia and Larsen-Freeman [1999, Ch. 16], and shown in Table 6.2, was compared with each argument, and if no match was found, the argument was considered lexical. By this standard, clausal arguments were always considered lexical arguments.

**Table 6.2:** Pronouns Used to Determine Non-Lexical Status of Arguments.

other	another	else	same	one	
this	that	these	those	what	
myself	yourself	herself	himself	itself	
ourselves	yourselves	themselves	oneself		
mine	yours	hers	his	ours	theirs
me	you	her	him	it	us
them	I	she	he	we	they

Figure 6.2 shows a C4.5 decision tree trained on cases consisting of the three verb valency attributes. It is noteworthy that the tree considers only two of the three attributes, and appears to indicate that native English has a larger proportion of trivalent verbs than does nonnative English, and that the opposite is true of divalent verbs. However, an extremely large number of training cases are misclassified by this tree, particularly by the middle leaf. In fact, Table 6.3 shows that the tree on average only classifies slightly more than half of all test cases correctly, and, worse still, the confidence interval for that accuracy encompasses 50%, indicating that the tree may be doing no better than a random classifier. It is a bit

surprising that the C4.5 algorithm was unable to construct a useful decision tree, as Du Bois presents data showing that there are statistically significant differences in the usage frequency of transitive and intransitive verbs between English and Spanish [Du Bois 2003, Table 2.3]. He found that in English 58% of finite verbs were intransitive and 42% transitive, whereas in Spanish the numbers were 63% and 37%. Furthermore, calculating from the data he gives, this difference is statistically significant at a high level of confidence:  $\chi^2 = 4.693, df = 1, p < 0.05$ . Based on Du Bois’s data, one might expect a decision tree to classify texts as native based on a low ratio of transitive (di- and trivalent) verbs to intransitive (monovalent) verbs. However, there are a number of possible reasons why this is not the case, the most obvious being that perhaps this characteristic is not involved in L1-transfer.



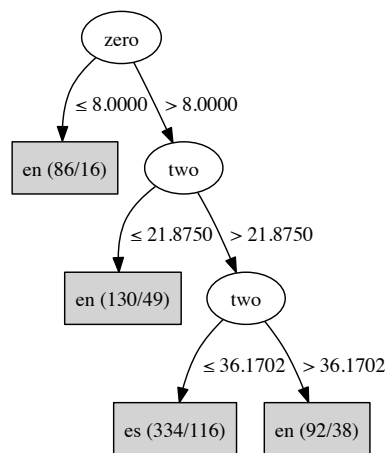
**Figure 6.2:** Verbal Clause Valency C4.5 Decision Tree.

**Table 6.3:** Accuracy of Verbal Clause Valency C4.5 Classifier

Nonnative	84.4%
Native	17.4%
Overall	50.9%
Overall C.I. 95%	47.1% — 54.8%

Figure 6.3 shows a decision tree generated by the C4.5 algorithm using cases with attributes indicating lexical argument density. Though Du Bois [2003] did gather lexical argument density data, he only appears to have done so for English and the Mayan language Sakapultek. The differences between those two languages in this regard are pronounced, but suggest little about any such differences between English and Spanish. The tree itself

is unusual, as it considers only two of the four attributes, one of which is considered twice. The first node in the tree checks whether fewer than 8% of clauses have no lexical arguments and, if so, immediately classifies that case as native. Du Bois's data shows that such clauses are common in spoken English, accounting for a startling 87% of all clauses. In written English, they are decidedly less common. In the 321 native training cases, finite verbs with no lexical arguments account for only 17% of all verbs. There does not seem to be much available research on lexical argument density in nonnative language, but it might be conjectured that learners of English have not developed clear distinctions between the written and spoken registers, and thus tend to show some speech-like elements in their written language, such as a high incidence of finite verbs without lexical arguments. The other two nodes in Figure 6.3 are more opaque to interpretation. These nodes both consider the two-lexical-argument attribute, essentially performing a ternary test on it, classifying cases with intermediate values as nonnative and cases with high and low values as native. It is hard to imagine that there is a convincing linguistic reason behind this, particularly considering the high rate of error associated with these nodes. Table 6.4 shows the accuracy of this classification system. As the confidence interval shows, it is an improvement over the valency system, though is still error-prone.



**Figure 6.3:** Lexical Argument Density C4.5 Decision Tree.

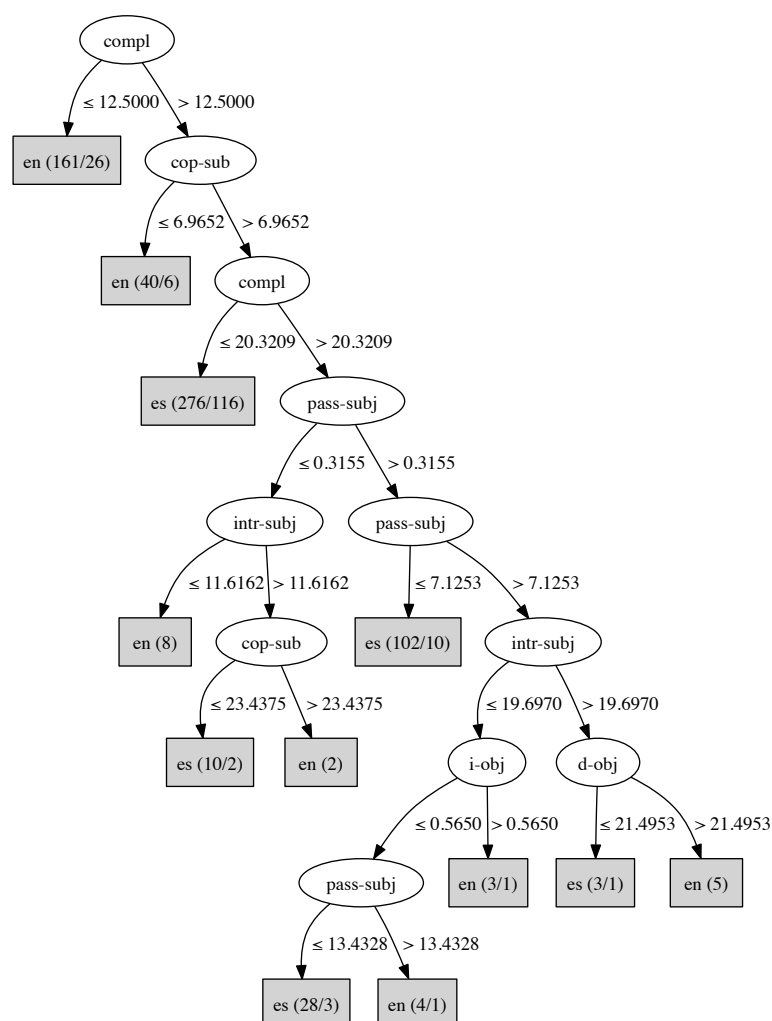
Training a C4.5 classifier on the lexical argument role attribute set yields the decision

**Table 6.4:** Accuracy of Lexical Argument Density C4.5 Classifier

Nonnative	70.7%
Native	48.6%
Overall	58.7%
Overall C.I. 95%	55.9% — 63.5%

tree shown in Figure 6.4. This tree is considerably more complicated than those shown in Figures 6.2 and 6.3, and is more accurate, as well. Table 6.5 shows that it correctly classifies test cases approximately two thirds of the time. Considering this tree in light of Du Bois’s data shows some interesting parallels. Of the three roles that he considered, Du Bois found that there was not a significant difference between English and Spanish in the number of new arguments appearing as transitive subjects, and, in fact, found very few such arguments. It is not surprising then, that the classifier excluded the transitive subject attribute when constructing the decision tree. Du Bois also gathered data for the intransitive subject role, finding that English tends to have fewer new core arguments in this role than does Spanish. As he explains in an earlier paper [Du Bois 1987], he lumps the subjects of copular verbs into this role as well. However, in the decision tree, intransitive and copular subjects are considered separately, as are the subjects of transitive verbs in the passive voice, which Du Bois presumably considered to be intransitive as well. Considering first the intransitive subject (*intr-subj*), it can be seen that this attribute appears twice in the decision tree. Where it appears closer to the root, the node splits the test cases at a value of 11.6162%. The lower values are immediately classified as native, while the others undergo an additional test, with the majority ultimately being classified as nonnative. The other usage of this attribute is in a separate branch, deeper in the tree. Here, using a higher comparison point, low values lead to a branch where the majority are ultimately classified as nonnative, while most high values are classified as native. With these conflicting branches, it is difficult to draw conclusions, but it can be seen that the latter-mentioned branch, the one that appears to coincide with Du Bois’s data, does deal with more training cases (43 with 86% accuracy) than does the other (20 with 90% accuracy). This suggests that, while

a minority of learners overuse lexical intransitive subjects, a large group of learners underuse them. A similar pattern holds with the copular subject. Early in the decision process, one test splits the training cases based on the *cop-sub* attribute and the group with the lower value is immediately classified as native. The other group, which is quite large, undergoes several more tests. The second use of the *cop-sub* is much deeper in the tree and is a direct descendant of the first, meaning that the information content of the attribute was much higher during the first test. This is convenient, as the first test is the one that coincides with Du Bois's data.



**Figure 6.4:** Lexical Argument Role C4.5 Decision Tree.

The passive subject attribute (*pass-subj*), is used three times in the classifier. The first



**Table 6.5:** Accuracy of Lexical Argument Role C4.5 Classifier

Nonnative	78.5%
Native	54.8%
Overall	66.7%
Overall C.I. 95%	63.0% — 70.3%

test that considers it splits off a small number of test cases with low values and passes them on to further tests. This group consists of approximately equal numbers of native and nonnative cases. The other, much larger group is passed onto another test, which also looks at the passive subject attribute. Here, the splitting causes a large number of cases with lower values to be immediately classified as nonnative. Deeper in the tree a pre-terminal node again considers this attribute, and once again classifies low values as nonnative. This is not what one would expect based on Du Bois's data for intransitive subjects. The most likely explanation is that the avoidance of lexical passive subjects by English learners is due to the avoidance of passives altogether. Butt and Benjamin [2004, 28.2.3] note that the passive in Spanish, particularly the form that is exactly parallel to the English passive, employing the copular verb plus a non-finite verb form, is rarely used, and that freedom of word order in Spanish allows simple fronting of an object, which is often why the passive is used in English. It is very likely that this is the source of L1-transfer, resulting in under-usage of the passive, and hence, the lexical passive subject, in learner English.

Du Bois also shows that the direct object role is more likely to be occupied by a NP bearing new information in English than in Spanish. The direct object attribute (*d-obj*) appears once in the decision tree. Cases with low values are immediately classified as nonnative, and those with high values as native. This suggests that L1-interference leads English learners into underusing lexical direct objects. The other two attributes used in the tree, the complement role (*compl*) and the indirect object role (*i-obj*), will not be closely analyzed here owing to a lack of data with which to compare them and to the difficulty of interpretation of their usages within the tree.

To gauge the potential accuracy of these attributes in constructing data models for clas-

sification, a 10-tree random forest classifier was trained using the lexical argument role attributes (the other attribute sets are too small for use with a random forest classifier). Another such classifier was trained on all 3 attribute sets combined. The results from these classifiers are shown in Tables 6.6 and 6.7, respectively. Considering first Table 6.6, it does appear that the random forest classifier is able to better take advantage of the data, the average accuracy being boosted from 66.7% to 70.0%. However, the overlapping of the confidence intervals means that there is still margin to doubt whether the random forest classifier is better than the C4.5 classifier. Table 6.7 shows that little is gained by including the other sets of attributes. The valency attributes, as was already shown, are an ineffective basis for classification, and do not seem to improve when combined with other attributes. The failure of the addition of the lexical argument density attributes to improve classification must mean that these attributes do not contribute information that is not already contained in the lexical role arguments.

**Table 6.6:** Accuracy of Lexical Argument Role Random Forest Classifier

Nonnative	76.6%
Native	63.2%
Overall	70.0%
Overall C.I. 95%	66.4% — 73.5%

**Table 6.7:** Accuracy of Combined Random Forest Classifier

Nonnative	76.6%
Native	63.9%
Overall	70.2%
Overall C.I. 95%	66.7% — 73.8%

## 7 Verbs

The experiments described in this section explore the suitability of using verbal features for language classification. The English verb shows limited grammatical inflection, in contrast to the Spanish verb, which is heavily inflected for tense, mood, number, and person. English, nevertheless, does have a great deal of complexity in its verbal system, employing a wide range of auxiliary verbs and particles to indicate the various possible tense, aspect, and mood combinations, as well as other subtleties of meaning. It is not possible to provide a detailed description of the English verbal system here, but an attempt will be made to touch on the most salient aspects. As they occur rarely in written texts, particularly in the corpora used in this study, question and imperative forms are not discussed here.

### 7.1 Grammar

Not counting the verb *to be*, English verbs have three finite inflected forms, as demonstrated here by the forms of the verb *to walk*: a past tense form (*walked*), a present third person singular form (*walks*), and a form for all other present person/number combinations (*walk*). This last form is also the base form of the verb, appearing in the infinitive after the particle *to* (*to walk*), and along with various auxiliary words to form the future (*will walk*) and numerous other verbal constructions (*should walk*, *am able to walk*, *have to walk*, etc.) In addition to the three finite forms, there is also a present participle (*walking*) and a past participle (*walked*), which is often identical to the past form.

The verb *to be* has five finite forms, with three in the present: the first person singular *am*, the third person singular *is*, and *are* for the other person/number combinations; and two in the past: *was* for both first and third person singular, and *were* for other cases. This latter form is also used in all persons and numbers for what is variously called the conditional or past subjunctive mood: (e.g. *if I were rich...*). In addition, there is the base form *be*, the present participle *being*, and the past participle *been*.

In addition to these basic inflected forms, the English verbal system relies on a broad array of auxiliary words. One such class of words are the modal auxiliaries or, simply, the modals. The nine English modals are shown in Table 7.1. These modals express concepts as basic as the future and the subjunctive, and others more subtle, such as intention, obligation, ability, and so forth. It is important to note that these modals, whether they reflect a change of tense or not, do not inflect to agree with the subject.

**Table 7.1:** English Verbal Forms Employing Modals

<i>will</i> walk	<i>can</i> walk	<i>should</i> walk
<i>may</i> walk	<i>could</i> walk	<i>must</i> walk
<i>might</i> walk	<i>shall</i> walk	<i>would</i> walk

Very similar to the modals are the phrasal modals. The thirteen phrasal modals considered in this experiment are those listed by Quirk et al. [1985, Ch. 3], and are shown in Table 7.2. Of these, *dare to*, *need to*, *have got to*, *have to*, and all those beginning with the verb *be*, can be inflected to show tense, person, and number, and can generally be used in conjunction with modals. The remaining phrasal modals do not inflect to agree with the subject, and are restricted to the present tense, with the exception of *used to*, which is restricted to the past tense. It is worth noting that Quirk et al. [1985] do not group all of these into one category, but into three separate categories of modal-like auxiliaries: *marginal modals*, *modal idioms*, and *semi-auxiliaries*. Because a more complicated analysis of these constructions would add little to this study, the term phrasal modal is here applied to any multiword modal-like construction.

**Table 7.2:** English Verbal Forms Employing Phrasal Modals

<i>dare to</i> walk	<i>used to</i> walk	<b>be</b> <i>about to</i> walk
<i>need to</i> walk	<i>had better</i> walk	<b>be</b> <i>able to</i> walk
<i>ought to</i> walk	<i>have got to</i> walk	<b>be</b> <i>bound to</i> walk
<i>have to</i> walk	<b>be</b> <i>supposed to</i> walk	<b>be</b> <i>willing to</i> walk
		<b>be</b> <i>obliged to</i> walk

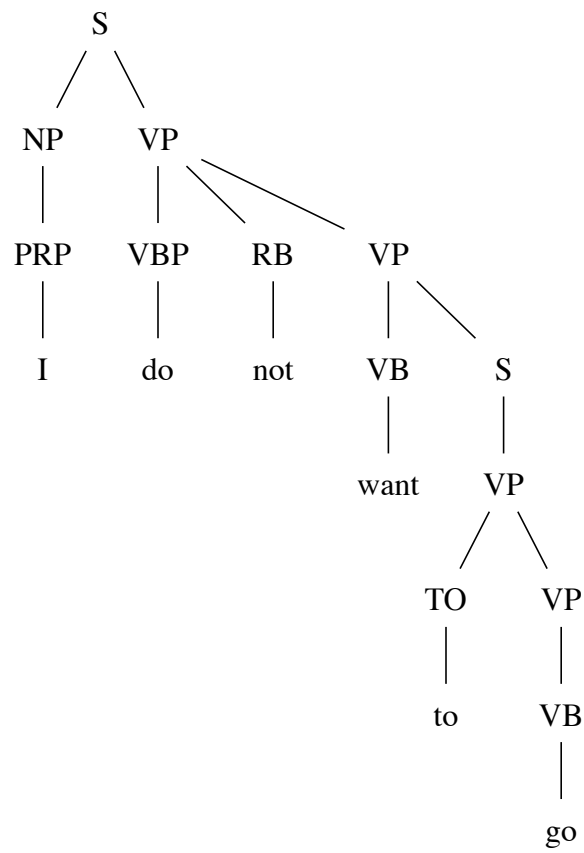
The words *have*, *be*, *get*, and *do*, which are full verbs in their own right, play an im-

portant role in English as auxiliary verbs. When used as such, they become the inflected element of the verb, being used in conjunction with a nonfinite form of the main verb. A form of the verb *have* followed by a past participle indicates the perfective aspect (e.g., *he has walked/had walked/will have walked/etc.*) *Be*, in any of its forms, plays two roles as an auxiliary. Followed by a present participle it forms the progressive aspect (e.g., *he is walking/was walking/will be walking/etc.*) Followed by a past participle, it forms the passive mood (e.g., *he was pursued.*) A passive can also be formed using *get* plus a past participle (e.g. *he got hurt*) A finite form of *do* is used to add emphasis to a sentence and to form questions, negatives, and affirmative responses to questions, but only when no modal is present (e.g., *you do not drink wine* but *you will not drink wine*).

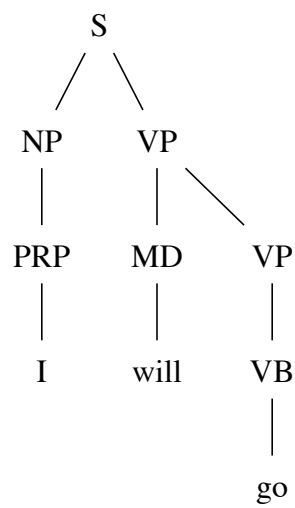
## 7.2 Parsing of Verbs

The Stanford Parser marks verbs, but does not explicitly mark all of the verbal attributes discussed above. The parse trees it generates indicate the structure of verb phrases (VPs), and distinguishes the various inflected forms of a verb. In general, it correctly distinguishes the finite and nonfinite uses of the base form of a verb. An example can be seen in Figure 7.1, which shows the parse of the sentence *I do not want to go*. Here the three verbs *do*, *want*, and *go* are tagged with the labels “VBP,” “VB,” and “VB,” respectively. “VBP” indicates a present form other than 3rd- person singular (which is indicated by “VBZ”). “VB” marks a base form in a nonfinite usage. The parser also marks past-tense forms, *ing*-forms, and *ed*-forms, using “VBD,” “VBG,” and “VBN,” respectively.

Any verbal information beyond that provided by these six tags must be determined from the shape and content of the VP subtrees in which the verbs are found. The parser directly marks modals, using the “MD” tag, as shown in Figure 7.2. However, it does not mark phrasal modals in a consistent manner. Figures 7.3 and 7.4 show how two phrasal modals are parsed differently. In Figure 7.3, *he is able to go* is parsed with *is* as the main verb, and with *able to go* being an adjectival phrase. In Figure 7.4, the sentence *he is going to go*



**Figure 7.1:** Typical Parse Tree Showing Verb Types.



**Figure 7.2:** Parse Tree Showing Modal *will*

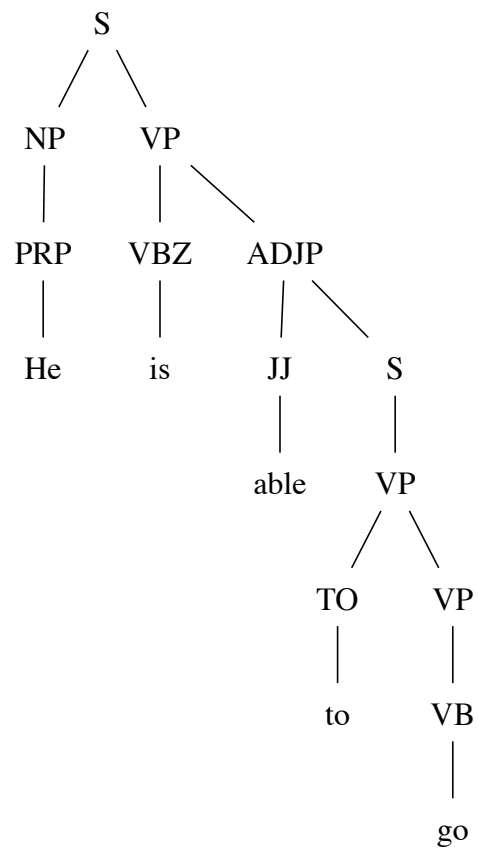
is also parsed as a copular sentence, but with the predicate nominative being treated as a nonfinite clause. Indeed, the parser treats all phrasal modals of the form *be* + particle + *to* as *be* verbs followed by an adjectival or participial construction. Perhaps this is not surprising as the distinction between these constructions and phrasal modals is somewhat blurry in English grammar. Quirk et al. [1985, footnote, p. 144] indicate that the main criterion for distinguishing these is whether what follows *be* is able to stand alone at the beginning of a sentence. Consider, for instance:

- (7.1) a. *Compelled to take stern measures, the administration lost popularity.*  
       b. *?Bound to take stern measures, the administration lost popularity.*  
       c. *Unable/Unwilling to resist, Matilda agreed to betray her country.*  
       d. *?Able/?Willing to resist, Matilda declined to betray her country.*  
       (Quirk et al. 1985, footnote, p. 144)

When fronted, the phrasal modals produce questionable sentences, as in (7.1b) and (7.1d). The non-phrasal modals, however, produce clearly acceptable sentences, shown in (7.1a) and (7.1c). Interestingly, by this criterion the negated phrasal modals (when negated on the lexical level) do not appear to be true phrasal modals and are not included as such in this study.

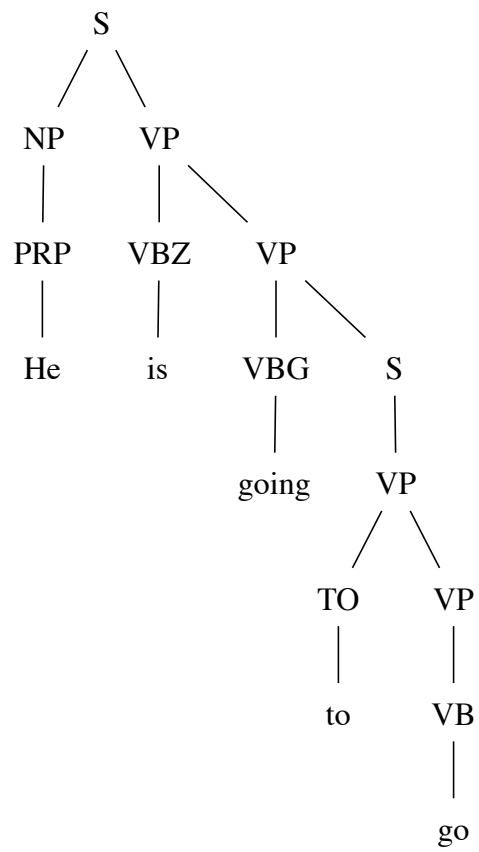
Phrasal modals are identified by searching parse trees for subtrees that match the basic form of the VP subtrees shown in Figures 7.3 and 7.4, and of other similar trees, while allowing for variation where appropriate. For instance, to match an instance of *be going to*, a subtree must be found that matches the most dominant VP subtree in Figure 7.4, with the exception that where the leaf [*go*] is found in the model, there may be any terminal node, and where the subtree [*VBZ — is*] is found, there may be any subtree representing a conjugation of *be* (e.g., [*VBD — was*], [*VBP — am*], etc.) The other phrasal modals and constructions using the auxiliary verbs *do*, *have*, and *be* are similarly identifiable using other distinctive subtrees.

Whenever the conjunction *and* is encountered in a VP subtree, multiple distinct but



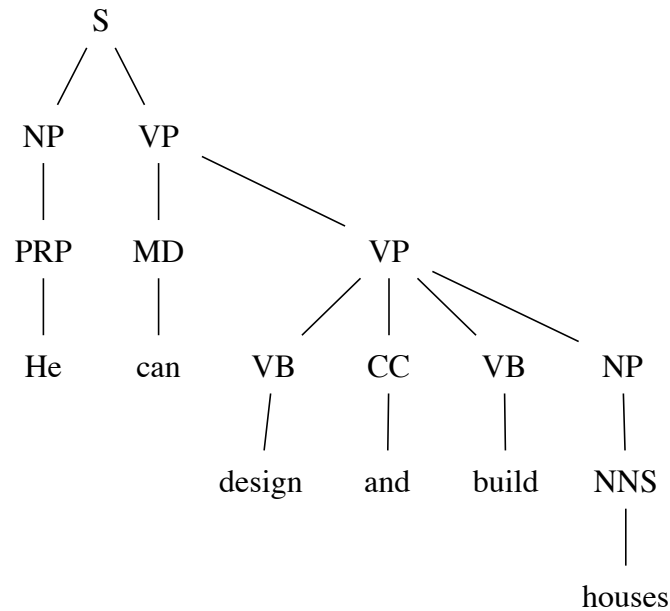
**Figure 7.3:** Parse Tree Showing Phrasal Modal *be able to*.





**Figure 7.4:** Parse Tree Showing Phrasal Modal *be going to*.

overlapping subtrees are generated from this and treated independently. For instance, the parse of the sentence *he can design and build houses* shown in Figure 7.5 is processed to generate the two separate parses *he can design houses* and *he can build houses* shown in Figure 7.6. The exception to this is when the conjunction is used to apply multiple modals to a verb (e.g. *I can and will...*).

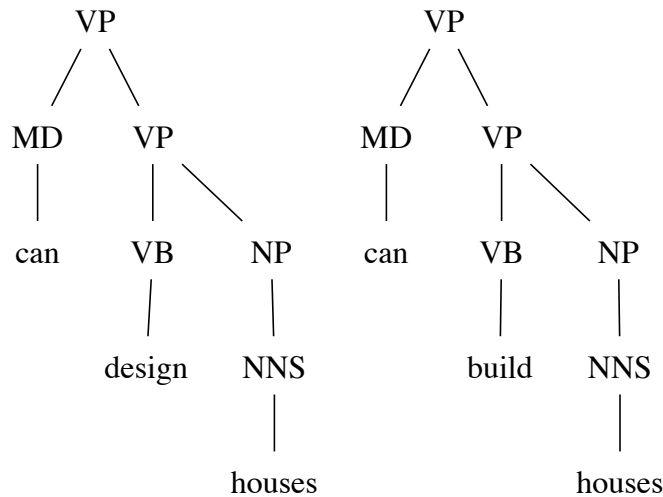


**Figure 7.5:** Parse Tree Showing a Verb with Embedded Conjunction

In all, this system uses 169 model subtrees to match the various possible supported verb configurations. It identifies the following binary independent attributes: the past tense, the perfect aspect, the progressive aspect, the passive voice, the presence of the auxiliary *do*, and whether it is negated with *not*. It also identifies any and all modals, the presence and identity of a phrasal modal, and the core verb.

### 7.3 Classification

Being able to identify so many attributes gives a wealth of data with which a classifier may be trained. The challenge, as always, is choosing subsets of this data that maximize information content while still producing decision trees comprehensible to a human. The first



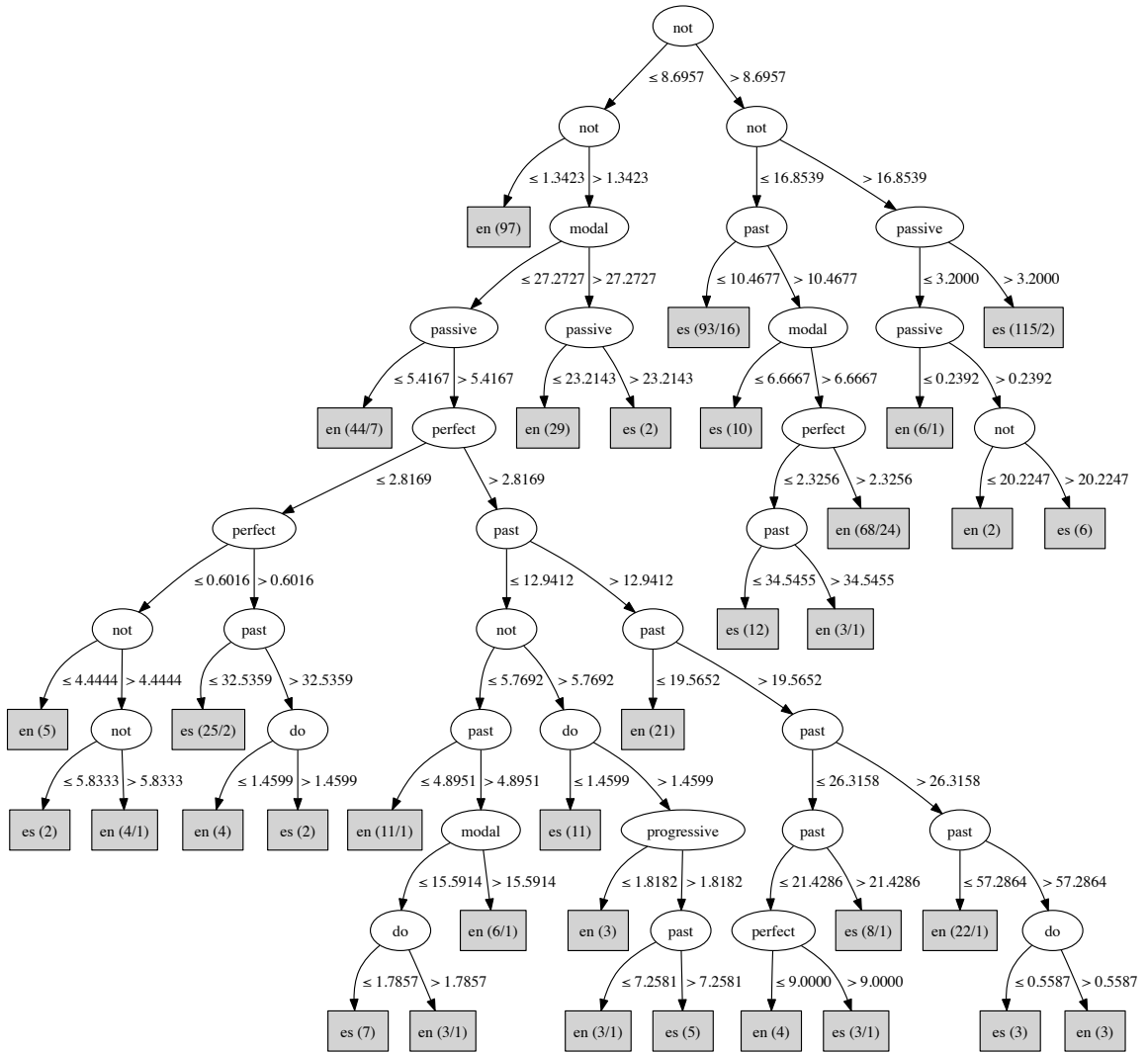
**Figure 7.6:** Parse Trees Showing the VP from Figure 7.5 Split into Two VPs

such subset examined here consists of eight attributes, each with a value indicating the relative frequency with which such verbal qualities are found in a text. These eight attributes are named “not,” “modal,” “progressive,” “past,” “perfect,” “passive,” “quasimodal,” and “do.” For the most part these are self-explanatory, with “not” being the relative frequency of the negating adverb *not* and so forth (“quasimodal” indicates a phrasal modal). It is worth pointing out that with the exception of “modal,” these verbal attributes can only occur once per verbal construction, but of course all can appear many times in a given text.

**Table 7.3:** Accuracy of C4.5 Classifier Using Various Verbal Attributes With Default Pruning

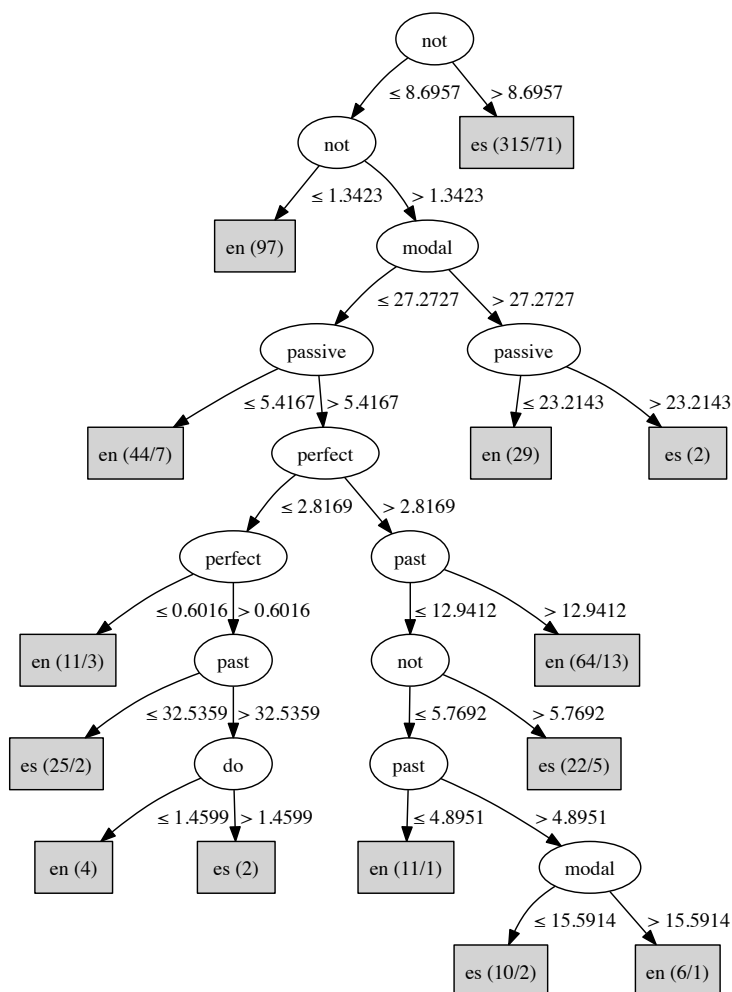
Nonnative	71.7%
Native	81.9%
Overall	76.8%
Overall C.I. 95%	73.5% — 80.1%

A C4.5 decision tree generated from this data set is shown in Figure 7.7 and the accuracy of such classifiers, calculated using 20-fold cross-validation, is shown in Table 7.3. As can be seen, this is quite a large tree for purposes of analysis, with many attributes appearing multiple times. A somewhat simpler tree, with only a modest lost of accuracy, can be had by using more aggressive pruning following the tree construction phase. Figure 7.8



**Figure 7.7:** C4.5 Decision Tree Using Various Verbal Attributes With Default Pruning

shows such a tree, and Table 7.4 its accuracy. This tree consists of 13 decision nodes considering 6 different attributes, none of which are considered more than three times.



**Figure 7.8:** Aggressively Pruned C4.5 Decision Tree Using Various Verbal Attributes

Of the eight attributes used in training, “quasimodal” and “progressive” are not found in the aggressively pruned tree, nor does “quasimodal” appear in the normally-pruned tree. “Modal,” however, does appear in both trees. In the aggressively-pruned tree it appears twice, once very near the root of the tree and the other time as the deepest decision node. The first of these splits the training cases into a predominately nonnative subset, those with frequencies higher than 27.2727%, and a predominately native subset with lower values. The second decision node, however, uses a lower comparison value and splits the cases the other way, with low values being classified as nonnative. This would seem to indicate that

**Table 7.4:** Accuracy of Agressively-Pruned C4.5 Classifier Using Various Verbal Attributes

Nonnative	71.3%
Native	79.8%
Overall	75.5%
Overall C.I. 95%	72.2% — 78.9%

extremes in modal usage are associated with nonnative usage. Before attempting to explain this, it is worth exploring the role of modal verbs in Spanish. According to the analysis of Butt and Benjamin [2004], Spanish has a small array of modal verbs, in general similar syntactically to English phrasal modals, as shown in the following examples:

- (7.2) a. *No debiste hacerlo.*  
*Not you-should-have done-it.*  
*‘You shouldn’t have done it.’*
- b. *Hubo de repetir el experimento.*  
*(S)he-had-to repeat the experiment.*  
*‘(S)he had to repeat the experiment’*  
[Butt and Benjamin 2004, Ch. 21.3,21.4]

Some of these are usually translated into English using modals or phrasal modals, and some require the use of other constructions [Butt and Benjamin 2004, Ch. 21]. Conversely, many English modals and phrasal modals can be translated into Spanish using that language’s modals, while others translate as verbal inflections. In addition, for those Spanish modals which can be translated into English modals or phrasal modals, the correspondence is rarely one-to-one, with there being a great deal of overlap and inexactness in meaning both ways. This should not be surprising, considering the close but imperfect correspondence between many of English’s modals and phrasal modals (e.g. *can* and *be able to*) [Celce-Murcia and Larsen-Freeman 1999, Ch. 8].

As mentioned above, the Spanish modals have much in common with the English phrasal modals. In general, the pattern of usage of the Spanish modals is *conjugated modal verb + one or zero particles + infinitive*. As was seen in Table 7.2, many of the English phrasal modals follow this pattern as well. With a few exceptions, the Spanish modals

can take the full range of verbal inflections, as can the majority of English phrasal modals. From this one might assume that L1-Spanish learners of English would take naturally to the English phrasal modal. Indeed, the decision trees shown in Figures 7.7 and 7.8 would seem to support this, or at least support the proposition that the learners neither overuse nor underuse phrasal modals relative to their native counterparts. A quick experiment in which the C4.5 classifier was run on the same data set, but with the “modal” attribute removed, showed that the “quasimodal” attribute was still not used in the resultant decision tree. This means that the reason for the algorithm’s excluding the “quasimodal” attribute cannot be attributed to its containing no information beyond what the “modal” attribute contains.

Returning to the “modal” attribute, it was mentioned above that the decision tree uses overuse and underuse of the English simple modals as an indication that the text being analyzed was written by a learner. Unfortunately, there is surprisingly little literature on the acquisition of English modals by Spanish speakers, but it is not difficult to imagine situations that would lead learners to either overuse or underuse modals. A learner might avoid the English modal, it being syntactically unusual from a Spanish grammar perspective, or, having discovered the relative simplicity of the English modal, which requires no verbal inflection, may use it to excess. The data also suggests that some learners avoid all types of modals.

In Figure 7.8, it could be seen that the attribute with the highest information content is “not.” This is used three times, with higher frequencies tending to lead to classification as nonnative and lower frequencies to native. It is tempting to attribute this to Spanish’s double negative construction [Butt and Benjamin 2004, ch. 23.3], but it seems doubtful that advanced English learners would not have grasped this basic difference between English and Spanish grammar. More likely is that native speakers, with their presumably larger vocabularies, have a greater number of lexical negatives (e.g. *he is unkind* versus *he is not kind*) at their disposal.

Considering next the “passive” attribute, it can be seen that this attribute, too, is used

in a consistent manner in the decision tree, with lower and higher frequencies leading to classification as native and nonnative, respectively. Spanish expresses passives in primarily two ways: using the copular verb *ser* plus a past participle in a construction similar to the English passive, or, more commonly, using the reflexive pronoun *se*:

(7.3) a. *Las muestras les serán devueltas.*  
*The samples to-you (pl.) will-be returned.*

b. *Se les devolverán las muestras.*  
*to-you (pl.) will-return the samples.*  
*‘The samples will be returned to you.’*  
 [Butt and Benjamin 2004, p. 402,8]

Based on these constructions, it is not surprising that L1-English learners of Spanish tend to overuse the *ser*-passive [Butt and Benjamin 2004, ch. 28.2.3]. That L1-Spanish learners of English do the same with the *be*-passive is more surprising. Indeed, this author was unable to find any literature that investigates, or even acknowledges, this phenomenon. One possibility has to do with the connection between the English passive and the presentation of information at the discourse level. The English passive is frequently used to reverse the order of what, in an active sentence, would be the subject and object. Such a reversal is often necessary to preserve the tendency in language to present old information before new information. However, the passive is only one of a number of constructions, the least marked such construction, perhaps, that allow the fronting of a particular element in a sentence [Ward and Birner 2008]. Other such constructions (e.g. preposing, inversion), being more complicated and alien, may be avoided by the learner, leading to an overuse of the passive.

The remaining three attributes used in the decision tree are rather resistant to analysis. The perfect aspect constructions in English and Spanish, for instance, are remarkably similar, with English using a form of the verb *to have* plus the past participle and Spanish using a form of the verb *haber* (cf. Latin *habere*, *to have*), plus a past participle. That the frequency of usage of the perfect should be a useful metric in classification is surprising.



Similarly, English and Spanish both have inflected past tenses which, combined with the complex role the “past” attribute plays in the decision tree, makes deciphering that attribute difficult. Finally, while Spanish has nothing quite like English’s *do* auxiliary, the extremely limited role which the “do” attribute plays in the classification process (it is only involved in classifying six out of 642 training cases) suggests that it is of little utility.

The next set of attributes deals with the relative frequencies of the various possible modals and phrasal modals. Table 7.2 shows the phrasal modals parsed in this system, of which there are 13; and Table 7.1 shows the modals, of which there are 9. This yields a total of 22 attributes. Training a C4.5 classifier on this data set produces a decision tree which is rather opaque in terms of interpretation, but the accuracy of such a classifier can be seen in Table 7.5. A derivative data set, which sacrifices accuracy for interpretability, pro-

**Table 7.5:** Accuracy of C4.5 Classifier Using Modal Attributes

Nonnative	76.1%
Native	73.8%
Overall	74.9%
Overall C.I. 95%	71.6% — 78.3%

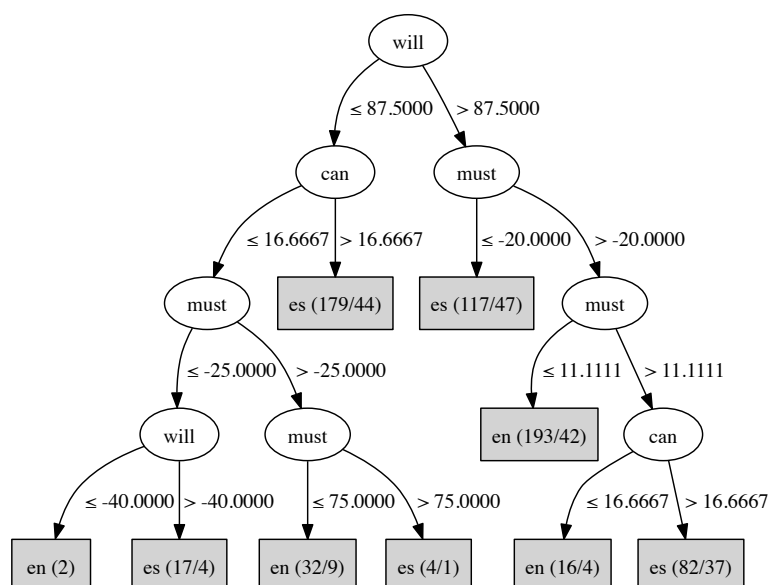
duces the C4.5 tree shown in Figure 7.9. The attributes used in this tree indicate whether there is a tendency in a text to use a modal over a phrasal modal with a similar meaning. Only four attributes were used, derived from a total of eleven modals and phrasal modals. Table 7.6<sup>4</sup> shows the modals used and the corresponding phrasal modals. The values for these attributes were calculated by subtracting from the number of occurrences of a particular modal the number of occurrences of the corresponding phrasal modals, and dividing the difference by the sum of these two quantities. This yields a real number ranging from  $-1$  to  $1$ , with the most negative number indicating all phrasal modals and no modals, zero indicating an even number of each, and  $1$  indicating all modals and no phrasal modals. For the purposes of display in Figure 7.9, these numbers are scaled by a factor of 100.

<sup>4</sup>For unspecified reasons, Butt and Benjamin [2004] does not include *ir a* with the modals. This is likely for the same reason that many English grammars do not consider *will* a modal, presumably because it modifies tense and not mood.

The labels given to the attributes are the names of the modal, these being unique for each correspondence.

**Table 7.6:** Correspondence Between Modals and Phrasal Modals

Modal	Phrasal Modal	Spanish Modal
<i>can</i>	<i>be able to</i>	<i>poder</i>
<i>must</i>	<i>have to</i>	<i>tener que</i>
	<i>have got to</i>	<i>deber</i>
	<i>need to</i>	<i>haber que</i>
<i>should</i>	<i>ought to</i>	<i>deber</i>
	<i>be supposed to</i>	<i>haber de</i>
	<i>be obliged to</i>	
<i>will</i>	<i>be going to</i>	<i>ir a</i>



**Figure 7.9:** C4.5 Decision Tree Using Modal vs Phrasal Attributes

The resultant tree uses three of the four attributes, with the “should” attribute not being included. The tree indicates that there is a strong tendency for the learners to use core modals more often relative to phrasal modals than the native writers. This is despite the similarity between English phrasal modals and Spanish modals. Table 7.6 also shows common Spanish modal equivalents to the English modals and phrasal modals. As can be seen

**Table 7.7:** Accuracy of C4.5 Classifier Using Modal vs Phrasal Attributes

Nonnative	70.7%
Native	61.4%
Overall	66.0%
Overall C.I. 95%	62.3% — 69.7%

from this table, each of these English modals can be expressed using a common Spanish modal, including *will/be going to*, which has a Spanish modal equivalent that is used alongside Spanish's inflected future [Butt and Benjamin 2004, ch. 14.6.4]. The best explanation for this is that learners prefer the core modals due to their syntactic simplicity.

The third set of attributes considers various common English verbs which are shown in Table 7.8. These verbs are identified as common verbs that tend to be overused by

**Table 7.8:** High Frequency Verbs

<i>make</i>	<i>use</i>	<i>take</i>	<i>see</i>	<i>say</i>
<i>go</i>	<i>become</i>	<i>believe</i>	<i>give</i>	<i>feel</i>
<i>come</i>	<i>find</i>	<i>think</i>	<i>know</i>	<i>look</i>
<i>seem</i>	<i>want</i>	<i>get</i>	<i>live</i>	<i>work</i>

English learners in a study by Ringbom [1998]. This study uses an earlier version of the ICLE, examining the French, Spanish, Finnish, Swedish, Dutch, and German subcorpora for usages of these verbs, and comparing the frequency of usage to that found in a native subcorpus of the ICLE<sup>5</sup>. Ringbom gives the breakdown per word, and finds that not all of the verbs show overuse in the Spanish subcorpus, with only *use*, *believe*, *feel*, and *come* showing overuse. Ringbom does not attempt to establish statistical significance, however. These verbs were used to construct a data set consisting of one attribute per verb with the value of each being equal to the relative frequency of that verb when used as a main verb in a finite clause. The accuracy of a C4.5 classifier trained on this data set is shown in Table 7.9.

Finally, to gauge the efficacy of verbal attributes in general, classifiers were trained on

<sup>5</sup>The version of ICLE used in the present study contains no native subcorpus.

**Table 7.9:** Accuracy of C4.5 Classifier Using High Frequency Verb Attributes

Nonnative	65.4%
Native	78.8%
Overall	72.1%
Overall C.I. 95%	68.6% — 75.6%

a combined data set consisting of all four sets of verbal attributes discussed here. Both a C4.5 classifier and a random forest classifier was tried. The accuracy of these classifiers is shown in Table 7.10 and Table 7.11. Combining the various attributes sets results

**Table 7.10:** Accuracy of C4.5 Classifier Using All Verbal Attributes

Nonnative	80.4%
Native	77.3%
Overall	78.8%
Overall C.I. 95%	75.7% — 82.0%

**Table 7.11:** Accuracy of Random Forest Classifier Using All Verbal Attributes

Nonnative	90.3%
Native	85.4%
Overall	87.9%
Overall C.I. 95%	85.3% — 90.4%

in an approximately 4% improvement in accuracy over using just the modal attributes, which have the highest accuracy of any of the attribute sets. However, one should note that the confidence intervals for the two classifiers do overlap somewhat, which would lead the cautious statistician to conclude that there is insufficient evidence to assert that one is better than the other at the 95% confidence level. At slightly lower confidence levels, however, there would be no overlap. Using a random forest classifier improves that accuracy by nearly 10%.

## 8 Discussion

The previous chapters describe several systems for classifying texts as native and nonnative based on grammatical features. The accuracy of these systems varies from approximately 70% to 90%. While such classifiers are interesting and may have some utility in and of themselves, it is worth exploring whether these can form the foundation of an educational tool for learners of English. An effort has been made, when examining the decision trees of the various classification systems, to find linguistic reasons behind the classifiers' choice of attributes. It has been found that some attributes are used in ways which are very easy to interpret, based on the grammars of English and Spanish and the concept of L1-transfer, and others seem incomprehensible. This linguistic analysis is important, as it gives clues as to which attributes might play a useful part of a learning tool and which should be excluded. This chapter seeks to further analyze the attributes used by the classifiers in light of their utility in a learner's tool, and to sketch out a possible design of such a tool.

### 8.1 Learner's Tool Design

Much of the complexity of any piece of user-oriented software is in the interface, but that will not be discussed in detail here. Suffice it to say that the program would allow the user, a learner of English, to enter a selection of English text that he or she has written. Because this system would be based on the frequency of textual attributes, the text would need to be long enough for the system to measure these frequencies with some level of statistical significance. Having entered a text, the user would then be presented with the output. This output would inform the user that certain features identify the text as nonnative, or that the system could not distinguish it from native text. In the former case, the system would show the user which features of the text are responsible for this classification and, when possible, would display all passages of the text that exhibited these features.

Identifying these features could be easily done if the system is using decision trees,

whether generated by the C4.5 algorithm or by a similar algorithm. Consider any of the decision trees shown in the previous chapters. When one of these trees labels a text as nonnative, it associates the text with a particular leaf on the tree. From the leaf to the root there is a unique path that passes through all of the decision nodes which were used in classifying the text. Each of these decision nodes considers a single attribute, and these attributes are those that most strongly mark the text as nonnative. It would not be difficult, using Weka, for instance, to implement a version of C4.5 that generates decision trees which output all attributes used in a particular classification, in addition to the class itself. The system could use a single tree, trained on all attributes, or it could use multiple trees. These trees might correspond to grammatically distinct attribute sets, as do the trees shown in the previous chapters, or they might be the various trees of a Random Forest classifier. Using multiple trees could enable the system to present the user with a wider range of features.

The passages which exhibit the telltale features could then be located using tables generated when the system initially identified the features. Many of these textual features would be examples of overuse of various grammatical constructions, and the user could focus on the displayed passages as he or she revises the text. Others, however, would be examples of underuse. In this case, the system would only be able to identify positive examples in the text, or perhaps no examples at all, and it would be up to the user to find the appropriate places in which to use the relevant construction. A more sophisticated system might group certain features together in complementary or nearly complementary pairs (e.g. modals and phrasal modals), and then use the complement of the underused feature to suggest areas for revision. The user could then edit the text and let the system reevaluate it. Eventually, if the user was able to make appropriate changes, there would be a change in how the text was classified. The system might still classify it as nonnative, but would do so at a different leaf node in the tree, with a different set of responsible attributes. The cycle of revision and reevaluation could continue until the system classifies the text as native.

## 9 Conclusion

## References

- AKMAJIAN, A., DEMERS, R. A., FARMER, A. K., AND HARNISH, R. M. 2010. *Linguistics: An Introduction to Language and Communication* 6 Ed. The MIT Press, Cambridge, Massachusetts.
- ALEJO GONZÁLEZ, R. 2010. L2 spanish acquisition of english phrasal verbs. In *Corpus-Based Approaches to English Language Teaching*, M. C. Campoy-Cubillo, B. Bellés-Fortuño, and M. L. Gea-Valor, Eds. Continuum International Publishing, Chapter 11.
- APTÉ, C., DAMERAU, F., AND WEISS, S. M. 1994. Automated Learning of Decision Rules for Text Categorization. *ACM Trans. Inf. Syst.* 12, 3, 233–251.
- BERRY, M. 2004. *Survey of Text Mining: Clustering, Classification, and Retrieval*. Number v. 1. Springer.
- BIBER, D. AND XEPPEN, R. 1998. Comparing Native and Learner Perspectives on English Grammar: a Study of Complement Clauses. In *Learner English on Computer*, S. Granger, Ed. Addison Wesley Longman, Chapter 11, 148–158.
- BOOTH, T. L. AND THOMPSON, R. A. 1973. Applying Probability Measures to Abstract Languages. *IEEE Transactions on Computers* C-22, 5, 442–449.
- BREIMAN, L. 2001. Random Forests. In *Machine Learning*. 5–32.
- BUTT, J. AND BENJAMIN, C. 2004. *A New Reference Grammar of Modern Spanish* Fourth Ed. McGraw-Hill.
- CELCE-MURCIA, M. AND LARSEN-FREEMAN, D. 1999. *The Grammar Book, an ESL/EFL Teacher's Course* Second Ed. Heinle and Heinle Publishers.

- CLIFFORD, J., JARKE, M., AND VASSILIOU, Y. 1983. A Short Introduction to Expert Systems. *SSRN eLibrary*.
- CORMACK, G. V. 2008. Email Spam Filtering: A Systematic Review. *Found. Trends Inf. Retr.* 1, 4, 335–455.
- DAGUT, M. AND LAUFER, B. 1985. Avoidance of Phrasal Verbs — A Case for Contrastive Analysis. *Studies in Second Language Acquisition* 7, 01, 73–79.
- DE MARNEFFE, M.-C. AND MANNING, C. D. 2008. Stanford typed dependencies manual.
- DU BOIS, J. W. 1987. The discourse basis of ergativity. *Language* 63, 4, 805–855.
- DU BOIS, J. W. 2003. Discourse and grammar. In *The New Psychology of Language: Cognitive and Functional Approaches to Language Structure*, M. Tomasello, Ed. Vol. 2. Lawrence Erlbaum Associates, Chapter 2, 47–87.
- GAMON, M. 2010. Using mostly native data to correct errors in learners’ writing: A meta-classifier approach. In *2010 Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.
- HALL, M., FRANK, E., HOLMES, G., PFAHRINGER, B., REUTEMANN, P., AND WITTEN, I. H. 2009. The WEKA data mining software: An update. *SIGKDD Explorations* 11, 1.
- HINKEL, E. 2003. Simplicity without elegance: Features of sentences in l1 and l2 academic texts. *TESOL Quarterly* 37, 2, 275–301.
- HULSTIJN, J. H. AND MARCHENA, E. 1989. Avoidance. *Studies in Second Language Acquisition* 11, 03, 241–255.
- KLEIN, D. AND MANNING, C. D. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Meeting of the Association for Computational Linguistics*. 423–430.



- LAUFER, B. AND ELIASSON, S. 1993. What causes avoidance in L2 learning. *Studies in Second Language Acquisition* 15, 01, 35–48.
- LEE, J. AND SENEFF, S. 2006. Automatic grammar correction for second-language learners. In *INTERSPEECH-2006*. 1978–1981.
- LEE, J. S. Y. 2009. Automatic correction of grammatical errors in non-native english text. Ph.D. thesis, Massachusetts Institute of Technology.
- LIAO, Y. AND FUKUYA, Y. J. 2004. Avoidance of phrasal verbs: The case of chinese learners of english. *Language Learning* 54, 2, 193–226.
- MARCUS, M. P., MARCINKIEWICZ, M. A., AND SANTORINI, B. 1993. Building a large annotated corpus of english: the penn treebank. *Computational Linguistics* 19, 2, 313–330.
- PIERA, C. 1995. On compounding in English and Spanish. In *Evolution and Revolution in Linguistic Theory*, H. Campos, Ed. Georgetown University Press, 302–315.
- QUINLAN, J. R. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers.
- QUIRK, R., GREENBAUM, S., LEECH, G., AND SVARTVIK, J. 1985. *A Comprehensive Grammar of the English Language*. Longman.
- RINGBOM, H. 1998. High-frequency verbs in the iclc corpus. In *Explorations in corpus linguistics*, A. Renouf, Ed. Language and Computers: Studies in Practical Linguistics. Rodopi.
- VERNON, A. 2000. Computerized grammar checkers 2000: capabilities, limitations, and pedagogical possibilities. *Computers and Composition* 17, 3, 329 – 349.
- WAGNER, J., FOSTER, J., AND VAN GENABITH, J. 2007. A comparative evaluation of deep and shallow approaches to the automatic detection of common grammatical errors.

In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language processing and Computational Natural Language Learning*. Association for Computational Linguistics, 112–121.

WARD, G. AND BIRNER, B. 2008. *Information Structure and Non-canonical Syntax*. Blackwell Publishing Ltd, 152–174.

WHITLEY, M. S. 1986. *Spanish/English Contrasts: A Course in Spanish Linguistics*. Georgetown University Press.