Thesis

Philip White

February 8, 2012

1 Corpora

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

The following corpora contributed native samples: the Brown University Standard Corpus of Present-Day American English subcorpus of letters-to-the-editor and editorials (BROWN), the International Corpus of English-Hong Kong (ICE-HK), the Michigan Corpus of Upper-level Student Papers (MICUSP), the Open American National Corpus (OANC), and the International Corpus of English-Canada (ICE-CAN). MICUSP and ICE-CAN contributed nonnative samples as well, and the remainder of the nonnative texts came from the International Corpus of Learner English, Spanish Subcorpus (SPICLE), the Santiago University Learner Corpus (SULEC), and the Written Corpus of Learner English (WRICLE). One additional student paper supplied by Missouri State University's English

Language Institute rounded out the nonnative samples. All nonnative samples were written by individuals whose first language was Spanish and who were judged, by the compilers of the corpora, to be advanced English learners. Many of the individuals had a language in addition to English and Spanish. In the cases of the SULEC and WRICLE corpora, both of which were compiled at Spanish universities, a large number of the learners spoke other Romance languages in addition to Spanish, in particular Catalan and Galician. Many of the samples in the ICE-HK corpus were written by individuals whose second language was Cantonese, and a number of the contributors to ICE-CAN had some French as well. Any sample written by an individual who knew a Germanic language (other than English) was not included.

Table 1.1: Corpora Composition

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

2 Parsing and Classification

2.1 Choice of Language

With very few exceptions, the code I wrote 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 a 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 I chose Clojure for its speed, functional style, and elegance.

2.2 Parsing

The Stanford Parser software package, version 1.6.7, configured with the probabilistic context-free grammar (PCFG) [Klein and Manning 2003], was used to generate all syntactic parse trees and grammar dependency graphs. In brief, PCFGs have their origins in the work of

2.3 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, i.e. parse trees and grammatical dependencies. As output they generate training or testing 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 ith attribute is g_i, the value f_i for that attribute is given by $g_n / \sum_{i=1}^{n} g_i$. For instance, 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 testing cases, the classes have missing values, until such values are determined by a classifier, as discussed in the following section.

2.4 Classification

I used the Weka machine learning package, version 3.6 [Hall et al. 2009], to create, train and test classifiers based on the cases discussed above. I primarily used two classifiers: 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.

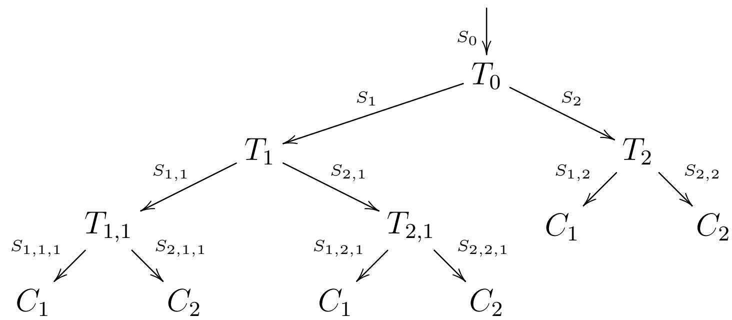
2.4.1 C4.5

This section describes the C4.5 partition 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 to exactly two class values, and which does not permit missing attribute values. That having been 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 2.1. 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.

The partitioning operations are 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

Figure 2.1: A decision tree showing the partitioning of a set of training cases S_0 into subsets $S_{1,2}$, $S_{1,1,1}$, and $S_{1,2,1}$ whose elements are of class C_1, and $S_{2,2}$, $S_{2,1,1}$, and $S_{2,2,1}$ whose elements are of class C_2. The nodes T_0, T_1, etc. are partitioning operations such that for any operation T operating on a set S_a the generated sets are $S_{1,a}$ and $S_{2,a}$ where $S_{1,a} \cup S_{2,a} = S_a$ and $S_{1,a} \cap S_{2,a} = \emptyset$.



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 decision of the attribute and threshold value for a particular test is determined using what Quinlan 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 the class of a case in S can be defined in terms of entropy as

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

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

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

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

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

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

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

Then the fitness of a test T can be measured using

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

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 $\{v_1, v_2, \ldots, v_m\}$, then any threshold between 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 than chooses a threshold value that will produce this result. Again, to partition S into two sets where the values for A are $\{v_1, v_2, \ldots, v_i\}$ and $\{v_{i+1}, v_{i+2}, \ldots, v_m\}$, 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.

3 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 $abbrev(\text{Corporation}, \text{ABC})$. 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 used.

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 3.1, employing 15 different relations. The full names for these relations are shown in Table 3.1. 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 336 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 205 native cases and only 20 nonnative cases, making it the native branch. This allows one to say, for instance, that fewer occurrences of the *nn* relation are associated with nonnative samples. The following subsections explore the linguistic reasons why these relations should be so useful in making such categorizations.

Table 3.1: Relation abbreviations

$auxpass$	passive auxiliary
$complm$	complementizer
cop	copula
det	determiner
$expl$	expletive
mwe	multi-word expression
nn	noun compound modifier
$npadvmod$	noun phrase as adverbial modifier
$parataxis$	parataxis
$poss$	possession modifier
$possessive$	possessive modifier
$predet$	preconjunct
prt	phrasal verb particle
$quantmod$	quantifier phrase modifier
rel	relative

Figure 3.1: C4.5 decision tree employing relative frequency of dependency relations. Relative frequencies are shown as percentages. Values in parentheses are the number of training case classified at that point and, following the slash when present, the number of those cases which were incorrectly classified.

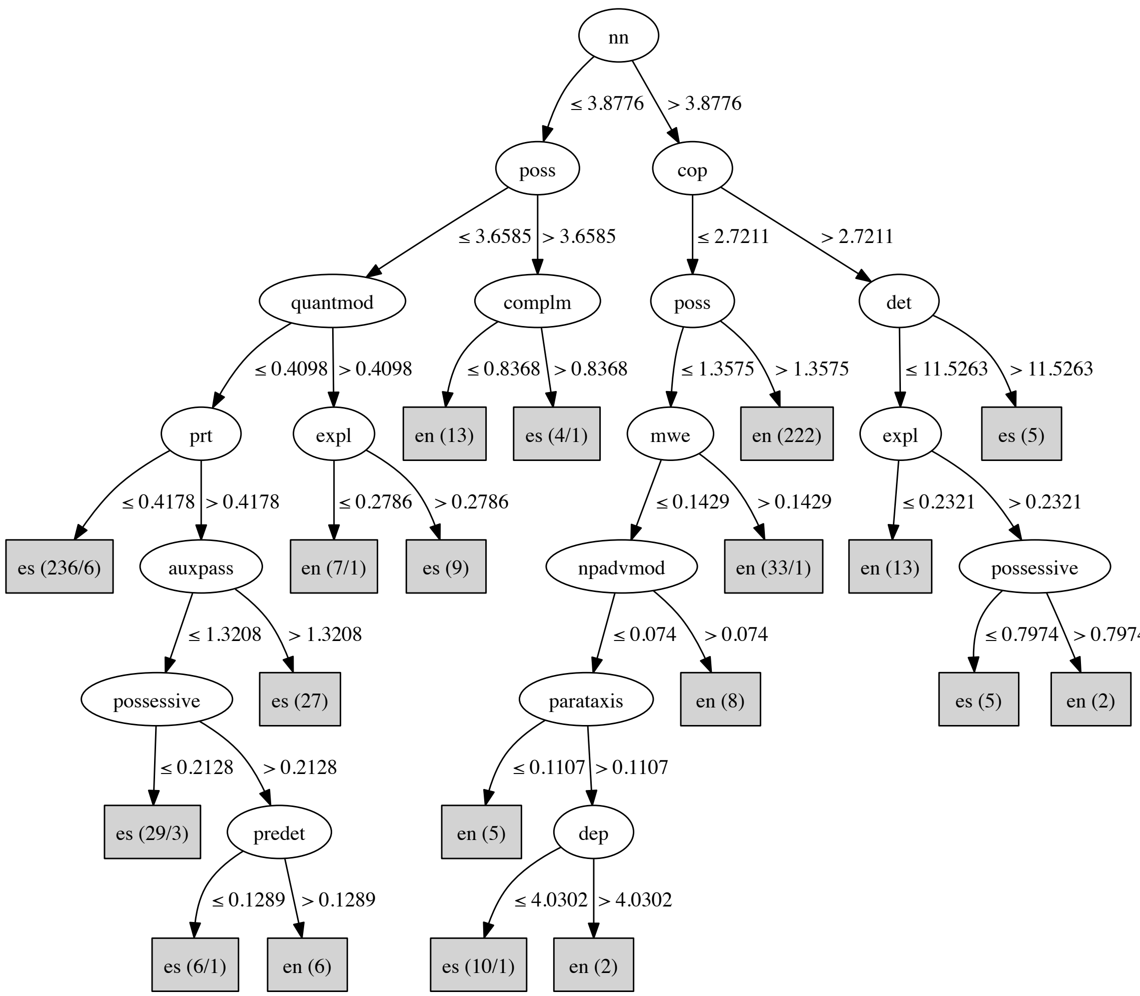


Figure 3.2: The dependencies $auxpass$(killed, been) and $auxpass$(killed, was/got). Taken from de Marneffe and Manning [2008].

auxpass

a. Kennedy has been killed

auxpass

b. Kennedy was/got killed

3.1 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 indicates heavy usage of the passive voice. Example 3.2 illustrates this dependency.

3.2 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 Example 3.3. 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.

Figure 3.3: 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*).

complm

a. … I will consider … whether the world is a safe place

complm

b. At least you choose whether to go to a pub or not.

complm complm

c. They state that climate generally predicts that temperatures should rise …

Whitley [1986] points out that while English tends to allow complementizers introducing clausal complements in the object position to be deleted, Spanish generally does not (see Example 3.1). Butt and Benjamin [2004, 33.4.6] explain that this rule is occasionally broken, but generally only in two situations, business letters and substandard 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 to L1-Spanish learners to have some preference for Example 3.1a over 3.1b, 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] make 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 shows similar levels of *that* omission, and in general these levels of omission are lesser than the levels found in comparable types of native texts. They also find, interesting 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 3.1 uses the $complm$ dependency once, and classifies cases with lower occurrences of $complm$ as native and larger occurrences as nonnative, without further testing. This this dependency does not part necessarily indicate the presence of a complement clause, but rather the presence of a complementizer, the higher frequency among the learners may be due either to low rates of dropping the complementizer, or

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

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

c. *Digo que lo hará.*

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

high rates of complement clause usage. As shown above, both phenomena have linguistic backing and very likely both are at play.

3.3 Copula

The copula or cop dependency marks the copular verb. This dependency takes as its governor the complement of the copular clause and the verb itself as the dependent.

3.4 Determiner

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.

3.5 Expletive

An existential *there* and the copular verb associated with it are connected with the expletive or $expl$ relation.

3.6 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 expression are considered multi-word expressions: *rather than, as well as, instead of, such as, because of, in addition to, all but, due to*.

3.7 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. Example 3.4 demonstrates this dependency.

Figure 3.4: The dependency $nn(\text{concentration}, \text{oxygen})$. Native sample taken from MICUSP.

nn

… oil's effects on dissolved oxygen concentration led me to …

3.8 Noun Phrase as Adverbial Modifier

3.9 Parataxis

3.10 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 3.5. The $poss$ dependency can also have as its dependent a possessive determiner such as *its* or *their*. In this type of construction, the $possession$ dependency is not used.

Figure 3.5: The dependencies $poss(\text{effects}, \text{oil})$ and $possessive(textoil, \text{'s})$. Native sample taken from MICUSP.

poss

… oil 's effects on dissolved oxygen concentration led me to …

possessive

3.11 Preconjunct

The preconjunct ($preconj$) dependency connects the head of a phrase employing a conjunction to a word that emphasizes or brackets that conjunction, such as *either*, *neither*, or *both*. Figure 3.6 demonstrates this dependency.

Figure 3.6: The dependencies $preconj(\text{boys}, \text{both})$ and $preconj(\text{emotionally}, \text{neither})$. (a) taken from de Marneffe and Manning [2008] and (b) from WRICLE (nonnative).

preconj

a. Both the boys and the girls are here.

preconj

b. ... are neither emotionally nor economically prepared ...

3.12 Phrasal Verb Particle

The phrasal verb particle relation (prt) ties the head word of a phrasal verb to its particle as shown in Example 3.7. The decision tree in Figure 3.1 contains this relation once. Relative frequencies of less than or equal to 0.4178% 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.

Figure 3.7: The dependency $prt(\text{free}, \text{up})$. Native sample from MICUSP.

prt

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

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 their same data gathering techniques to a group of English learners whose first language was Dutch, a language which also uses phrasal verbs. Contrary to their expectations, they found that the Dutch speakers did not avoid phrase 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 made 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 the English of those learners. 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 L1-Spanish learners avoidance of phrasal verbs, 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.

Table 3.2: 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	
Classified as $\rightarrow$	es	en	es	en
es	309	12	291	30
en	25	296	36	285
% Correct	89.72		94.24	
MAE	0.1139		0.1707	
κ	0.7944		0.8847	

3.13 Quantifier Phrase Modifiers

3.14 Relative

3.15 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 κ 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 3.2 gives the confusion matrices for these two classifier.

4 Argument Structure

The classification systems discussed in this section considered the argument structure of verbs. In general, every finite verb in English takes one or more arguments, with a subject argument always being required. The Stanford NLP system marks the arguments of verbs using the dependencies shown in Table 4.1. In the majority of non-copular sentences, the

Figure 4.1: The dependencies $complm(\text{place}, \text{whether})$, $complm(\text{go}, \text{whether})$ $complm(\text{predicts}, \text{that})$, and $complm(\text{rise}, \text{that})$. Nonnative samples from WRICLE (*a* and *c*) and SULEC (*b*).

nsubj dobj

a. In nineteen ninety-seven Roland Lemon was elected president.

nsubj iobj

b. … before his father gave him the rest of it.

dobj

complm complm

c. They state that climate generally predicts that temperatures should rise …

governors of these dependencies are the core verb. In copular sentences the governor will be the subject complement, i.e. the argument generally appearing after the verb which is equated with the subject. Table 4.1 show examples of these dependencies.

Table 4.1: The dependencies used to identify verbal arguments.

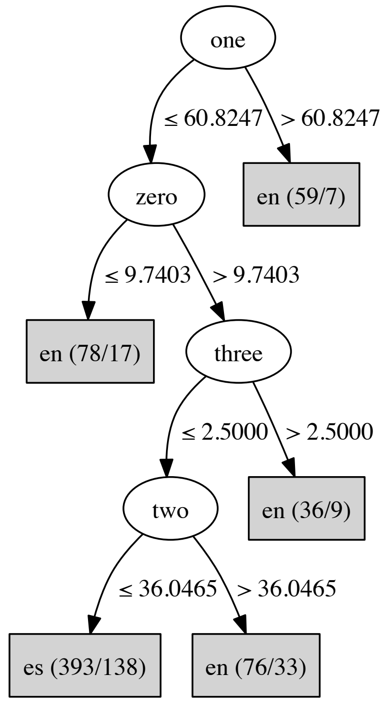
csubj	Clausal Subject
ccomp	Clausal Complement
dobj	Direct Object
iobj	Indirect Object
nsubj	Nominal Subject
nsubjpass	Passive Nominal Subject
xcomp	Open Clausal Complement

References

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.

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.

Figure 4.2: Lexical argument quantity C4.5 decision tree. The four attributes have values equal to the percentage of finite verbal clauses with zero, one, two, or three lexical arguments.



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.

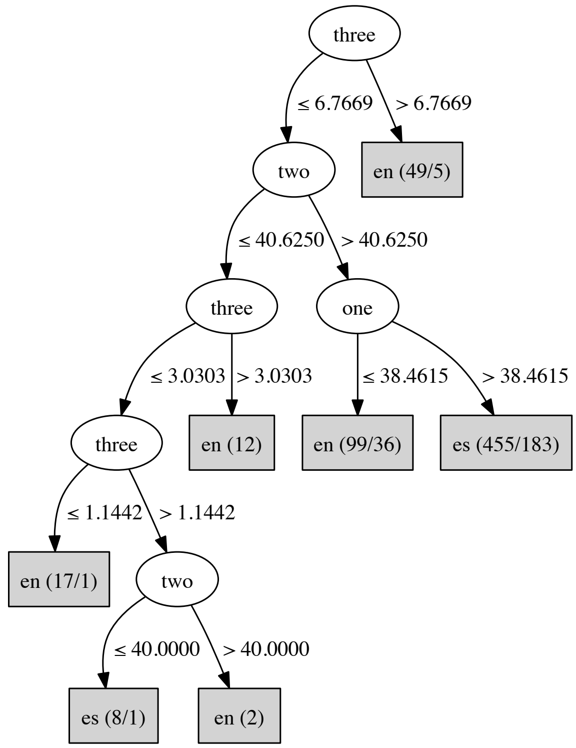
CELCE-MURCIA, M. AND LARSEN-FREEMAN, D. 1999. *The Grammar Book, an ESL/EFL Teacher's Course* Second Ed. Heinle and Heinle Publishers.

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., MACCARTNEY, B., AND MANNING, C. D. 2006. Generating typed dependency parses from phrase structure parses. In *LREC 2006*.

DE MARNEFFE, M.-C. AND MANNING, C. D. 2008. Stanford typed dependencies manual.

Figure 4.3: Verbal clause valency C4.5 decision tree. The three attributes have values equal to the percentage of finite verbal clauses which take one, two, or three arguments.



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.

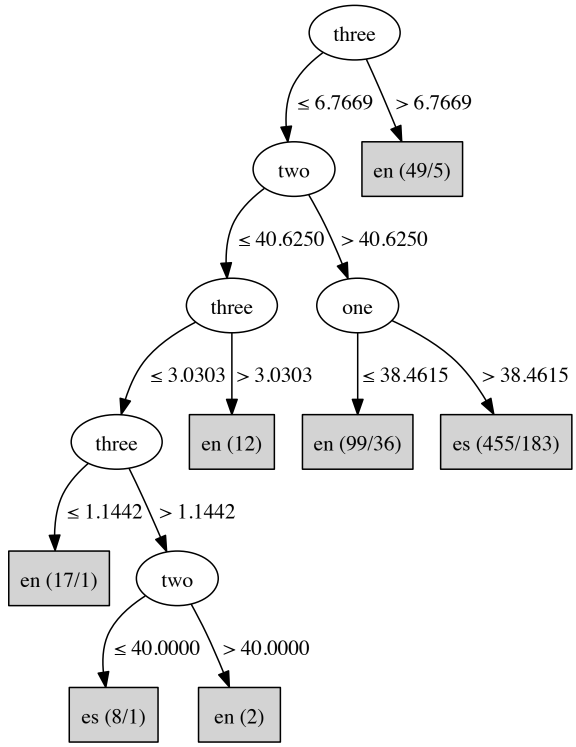
HULSTIJN, J. H. AND MARCHENA, E. 1989. Avoidance. *Studies in Second Language Acquisition 11,* 03, 241–255.

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.

LIAO, Y. AND FUKUYA, Y. J. 2004. Avoidance of phrasal verbs: The case of chinese learners of english. *Language Learning 54,* 2, 193–226.

Figure 4.4: Lexical argument role C4.5 decision tree. Attributes have values indicating the percentage of all lexical arguments which are found in that role.



MOORE, A. W. 1994. Efficient algorithms for minimizing cross validation error. In *Proceedings of the Eleventh International Conference on Machine Learning*. Morgan Kaufmann, 190–198.

QUINLAN, J. R. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers.

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

WITTEN, I. H. AND FRANK, E. 2005. *Data Mining: Practical Machine Learning Tools and Techniques* Second Ed. Morgan Kaufmann.