1 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 $abbrev(\text{Corporation, ABC})$. 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 1.1, employing 15 different relations. The full names for these relations are shown in Table 1.1. 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 1.1: 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

1.1   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-
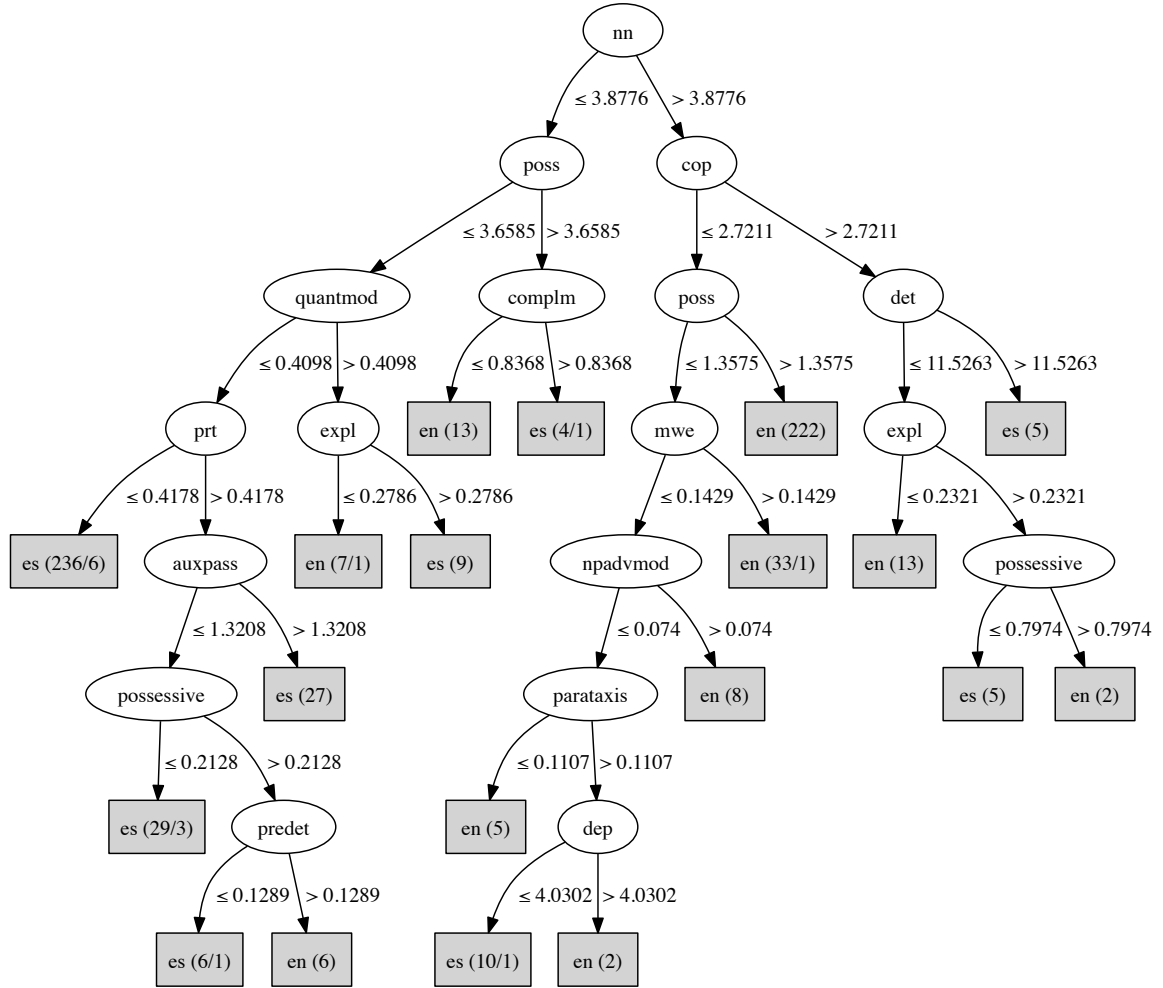
Figure 1.1: C4.5 decision tree employing relative frequency of dependency relations. Relative frequencies are shown as percentages. Values in parentheses are the number of training case classified at that point and, following the slash when present, the number of those cases which were incorrectly classified.
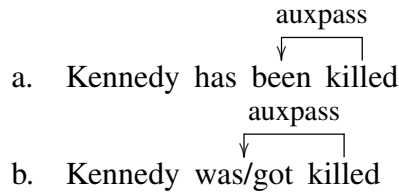
$$\overset{\text{auxpass}}{\overbrace{}}$$

a.   Kennedy has been killed

$$\overset{\text{auxpass}}{\overbrace{}}$$

b.   Kennedy was/got killed

Figure 1.2: The dependencies $auxpass(\text{killed}, \text{been})$ and $auxpass(\text{killed}, \text{was/got})$. Taken from de Marneffe and Manning [2008].

cates heavy usage of the passive voice. Figure 1.2 illustrates this dependency.

The decision tree in Figure 1.1 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 addition 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 the nativeness of the text. Likely a slightly more aggressive pruning factor would have resulted in the elimination of the $auxpass$ node.

1.2   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 1.3. 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 complementizers introducing clausal complements in the object position to be deleted, Spanish generally does not (see Example 1.1). 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,
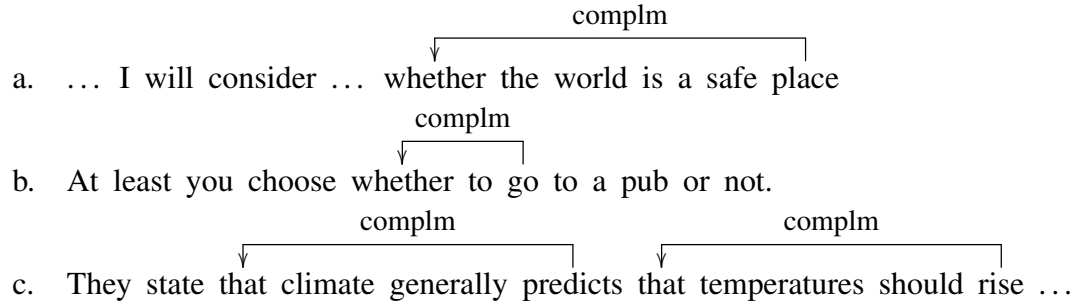
$$\overset{\text{complm}}{\frown}$$

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

$$\overset{\text{complm}}{\frown}$$

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

$$\overset{\text{complm}}{\frown} \qquad \overset{\text{complm}}{\frown}$$

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

Figure 1.3: 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*).

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 Example 1.1a over 1.1b, 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 1.1 uses the *complm* dependency once, and classifies

(1.1) 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)

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

1.3　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 1.4 shows examples of two different copular verbs. However, the Stanford parser does not
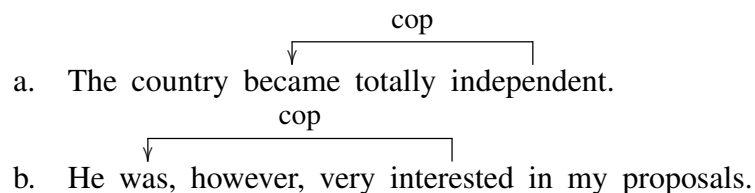
　　　　　　　　　　　　　　cop
　　　　a. The country became totally independent.
　　　　　　　　　　　cop
　　　　b. He was, however, very interested in my proposals.

Figure 1.4: The dependencies $cop(\text{independent}, \text{became})$ and $cop(\text{interested}, \text{was})$. [Quirk et al. 1985, Ch. 2.15, 2.16].

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

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

　　The decision tree in Figure 1.1 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 sentential simplicity in L1 and L2 academic texts.

1.4   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 1.5 shows examples of both determiners and a predeterminer in one sentence fragment. By the analysis of Quirk et al., a determiner is
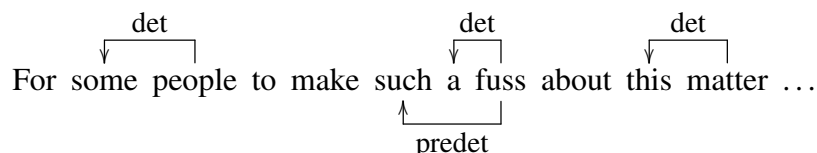

Figure 1.5: The dependencies $det(\text{people}, \text{some})$, $predet(\text{fuss}, \text{such})$, $det(\text{fuss}, \text{a})$ and $det(\text{matter}, \text{this})$. Nonnative sample taken from SULEC.

an element which modifies a NP, precedes any adjectives modifying the NP, and which expresses the type of reference made by that NP [Quirk et al. 1985, Ch. 5.10]. Adjectives, on the other hand, modify the attributes of a NP. Quirk divides 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 1.1. It should not comes 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*, etc.

The predeterminers consist of words which generally precede core determiners and which includes certain words which modify quantity, such as *all*, *both*, *double*, *half*, etc. and others more difficult to define: *such*, *what*, etc. 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 1.1 shows one usage each of the 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 for the indefinite article. The rules governing these uses can certainly be trying for learners, but for the most part one would expect advanced learners to have mastered these concepts. Perhaps more difficult a concept, and one which may account, at least in part, 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:

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

   b. *Tigers have four legs.*

   c. *Los tigres tienen cuatro patas*

Though, in the right context, 1.3a could refer to a particular tiger, it could also refer to tigers in general, as 1.3b 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 1.3c [Whitley 1986, Ch. 8.3.3]. It is possible that the L1-Spanish learner of English, being accustomed to 1.3c, would choose the grammatically correct 1.3a instead of the more common 1.3b. 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 use 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 and thus very 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:

(1.4) 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 1.4b Butt and Benjamin [2004, Ch. 10.10]. Spanish also tends to use constructions similar to 1.4b to express multipliers:

(1.5) *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:

(1.6) *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.

1.5 Expletive

An existential *there* and the copular verb associated with it are connected with the expletive or *expl* relation.
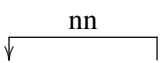
1.6 Multi-Word Expression

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*.

1.7 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 1.6 demonstrates this dependency.

Figure 1.6: The dependency $nn(\text{concentration}, \text{oxygen})$. Native sample taken from MICUSP.

nn

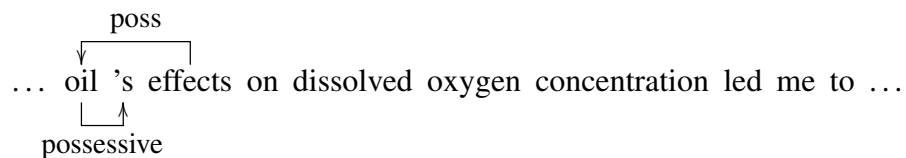... oil's effects on dissolved oxygen concentration led me to ...

1.8   Noun Phrase as Adverbial Modifier

1.9   Parataxis

1.10   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 1.7. 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.
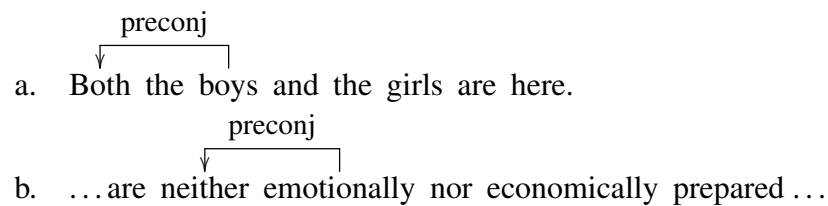
Figure 1.7: The dependencies $poss(\text{effects}, \text{oil})$ and $possessive(textoil, \text{'s})$. Native sample taken from MICUSP.

$$
\begin{array}{c}
\text{poss} \\
\dots \quad \text{oil} \quad \text{'s} \quad \text{effects on dissolved oxygen concentration led me to } \dots \\
\text{possessive}
\end{array}
$$

1.11   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 1.8 demonstrates this dependency.

Figure 1.8: 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).

$$
\begin{array}{c}
\text{preconj} \\
\text{a.} \quad \text{Both the boys and the girls are here.}
\end{array}
$$

$$
\begin{array}{c}
\text{preconj} \\
\text{b.} \quad \dots \text{are neither emotionally nor economically prepared} \dots
\end{array}
$$

1.12 Phrasal Verb Particle

The phrasal verb particle relation (prt) ties the head word of a phrasal verb to its particle as shown in Example 1.9. The decision tree in Figure 1.1 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 1.9: The dependency $prt(\text{free}, \text{up})$. Native sample from MICUSP.

... 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.

1.13 Quantifier Phrase Modifiers

1.14 Relative

1.15 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 1.2 gives the confusion matrices for

Table 1.2: 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 →	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	

these two classifier.

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.

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. AND MANNING, C. D. 2008. Stanford typed dependencies manual.

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.

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.

QUIRK, R., GREENBAUM, S., LEECH, G., AND SVARTVIK, J. 1985. *A Comprehensive Grammar of the English Language*. Longman.

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