1 Grammatical Relations

5.1 Dependency 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, 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.

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, employing 15 different relations. The relations uses in this tree are *auxpass* (passive auxiliary), *complm* (complementizer), *cop* (copula), *dep* (unclassified dependency), *det* (determiner), *expl* (expletive), *mwe* (multi-word expression), *nn* (noun compound

modifier), *npadvmod* (noun phrase as adverbial modifier), *parataxis*, *poss* (possession modifier), *possessive* (possessive modifier), *predet* (predeterminer), *prt* (phrasal verb particle), and *quantmod* (quantifier phrase modifier).

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.

5.2   Relations Described

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. An example [de Marneffe and Manning 2008] of this is:

auxpass

a. Kennedy has been killed

auxpass

b. Kennedy was/got killed.

The decision tree in Figure 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
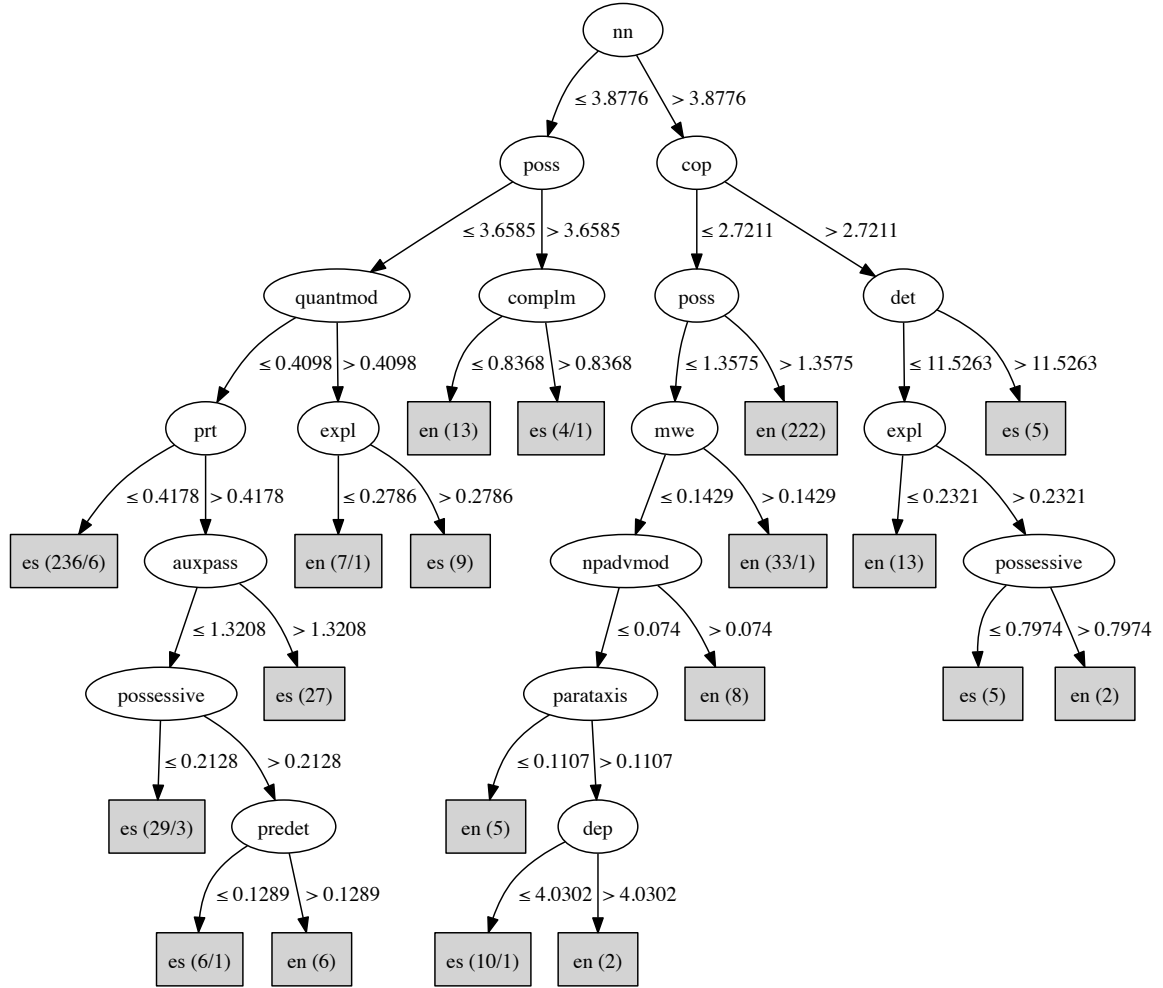

Figure 1: C4.5 Decision Tree Employing Relative Frequency of Relations
C4.5 Decision Tree Employing Relative Frequency of Dependency Relations

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.

Complementizer. A complementizer is a word that signals the beginning of a clausal complement. The Stanford Parser recognizes the complementizers *that* and *whether*. The governor of a complementizer dependency is the root of the clause, which is generally a verb or, in the case of copular clauses, the subject complement. The dependent is the complementizer itself. The following examples taken from the SULEC and WRICLE corpora illustrate this dependency:

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

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

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

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:

(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)

Butt and Benjamin [2004, p. 473] explain that this rule is occasionally broken, but generally only in two situations: business letters and nonstandard speech, and when the
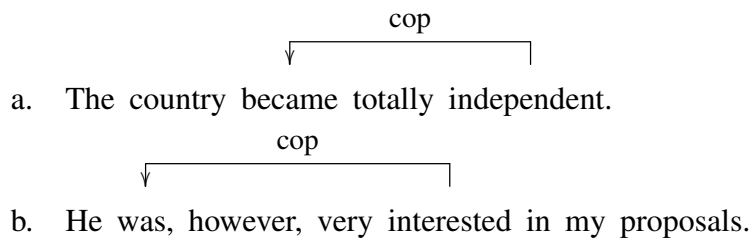
complementizer *que* appears close to other uses of the word *que*. Since these are restricted cases, it is reasonable to conclude that there would be L1-transfer in the construction of clausal complements, leading L1-Spanish learners to have some preference for (1a) over (1b), particularly considering 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 1 uses the $complm$ dependency once, and classifies cases with lower occurrences of $complm$ as native, and those with higher 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.

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. The following [Quirk et al. 1985, pp. 52-3] are examples of two different

copular verbs:

a.   The country became totally independent.

b.   He was, however, very interested in my proposals.

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

(2) *I have been in the garden.* [Quirk et al. 1985, p. 53]

The decision tree in Figure 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 best explanation for this is 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.

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. The following, a nonnative sample from the SULEC corpus, is an example of both determiners and a predeterminer in one sentence fragment:

For some people to make such a fuss about this matter ...
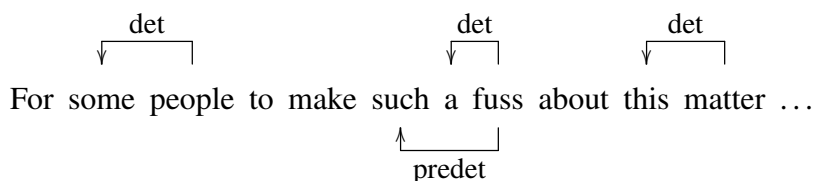
By the analysis of Quirk et al. [1985, p. 253], a determiner is an element which

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 1. It should not comes as a surprise that the $predet$ relation marks predeterminers, but it is worth 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 include certain words which modify quantity, such as *all*, *both*, *double*, and *half*, 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 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 some 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 usages 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:

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

b. *Tigers have four legs.*

c. *Los tigres tienen cuatro patas*

Though, in the right context, (3a) could refer to a particular tiger, it could also refer to tigers in general, as (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 (3c) [Whitley 1986, p. 157]. It is possible that the L1-Spanish learner of English, being accustomed to (3c), would choose the grammatically correct (3a) instead of the also grammatically correct but more common (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:

(4) a. *He did it in <u>a third</u> the time it took me.*

b. *He did it in <u>a third of</u> the time it took me.* [Quirk et al. 1985, p. 261]

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 (4b) [Butt and Benjamin 2004, pp. 122-3]. Spanish also tends to use constructions similar to (4b) to express multipliers:

(5) *El aire contiene <u>el doble de</u> óxido de nitrógeno que en Washington.* [Butt and Benjamin 2004, p. 125]

whereas English would use a simple predeterminer:

(6) *The air contains <u>double</u> 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.

Expletive. An existential *there* and the copular verb associated with it are connected with the expletive or $expl$ relation, as shown in the following [Quirk et al. 1985, pp. 126-7]:
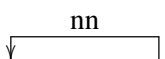
There was someone knocking at the door.

This relation is used twice in the decision tree, 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, pp. 429-32]. 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.

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.

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. An example of this, taken from the MICUSP corpus, is:

<div align="center">

nn

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

</div>

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 should 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:

(7) *un traje de baño*
    a   suit of bath
    'a bathing suit' [Butt and Benjamin 2004, p. 495]

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:

(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.

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.

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. The following sentences [de Marneffe and Manning 2008; Quirk et al. 1985, p. 921] each show an example of one of these types of parataxis:

parataxis

a. The guy, John said, left early in the morning.

parataxis

b. John plays the guitar; his sister, moreover, plays the piano.

The types of constructions marked by the *parataxis* 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 do 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 is being overused.

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. 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. The following example, taken from the MICUSP corpus, illustrates these dependencies:

poss

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

possessive

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 previously. Much like NNCs, the inflected genitive is usually translated into Spanish using a *de*-construction, as shown in:

12

(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, p. 497]

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

(10) a. *the mountain's peak*

      b. *the peak of the mountain*

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

      Phrasal Verb Particle. The phrasal verb particle relation (prt) ties the head word
of a phrasal verb to its particle. 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 following native example from the
MICUSP corpus shows this:

<div align="center">

prt

…the reduction of superfluous proteins will free up resources…

</div>

      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.

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 if in the following example de Marneffe and Manning [2008] the number *200* were replaced with the determiners *half the*, *about* would become the dependent of an $advmod$ dependency instead of the $quantmod$ dependency:

quantmod

About 200 people came to the party.

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.

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.

5.3   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 88.8% (see Table 1). 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 94.2% accuracy (see Table 2).

Table 1: Accuracy of C4.5 Classifier Using Dependency Attributes

	Mean	95% C.I.
Nonnative	89.7%	-
Native	87.9%	-
Overall	88.8%	86.3% — 91.2%

Table 2: Accuracy of Random Forest Classifier Using Dependency Attributes

	Mean	95% C.I.
Nonnative	96.0%	-
Native	91.9%	-
Overall	93.9%	92.1% — 95.8%

Some text

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.

PIERA, C. 1995. On compounding in English and Spanish. In *Evolution and Revolution in Linguistic Theory*, H. Campos, Ed. Georgetown University Press, 302–315.

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.