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

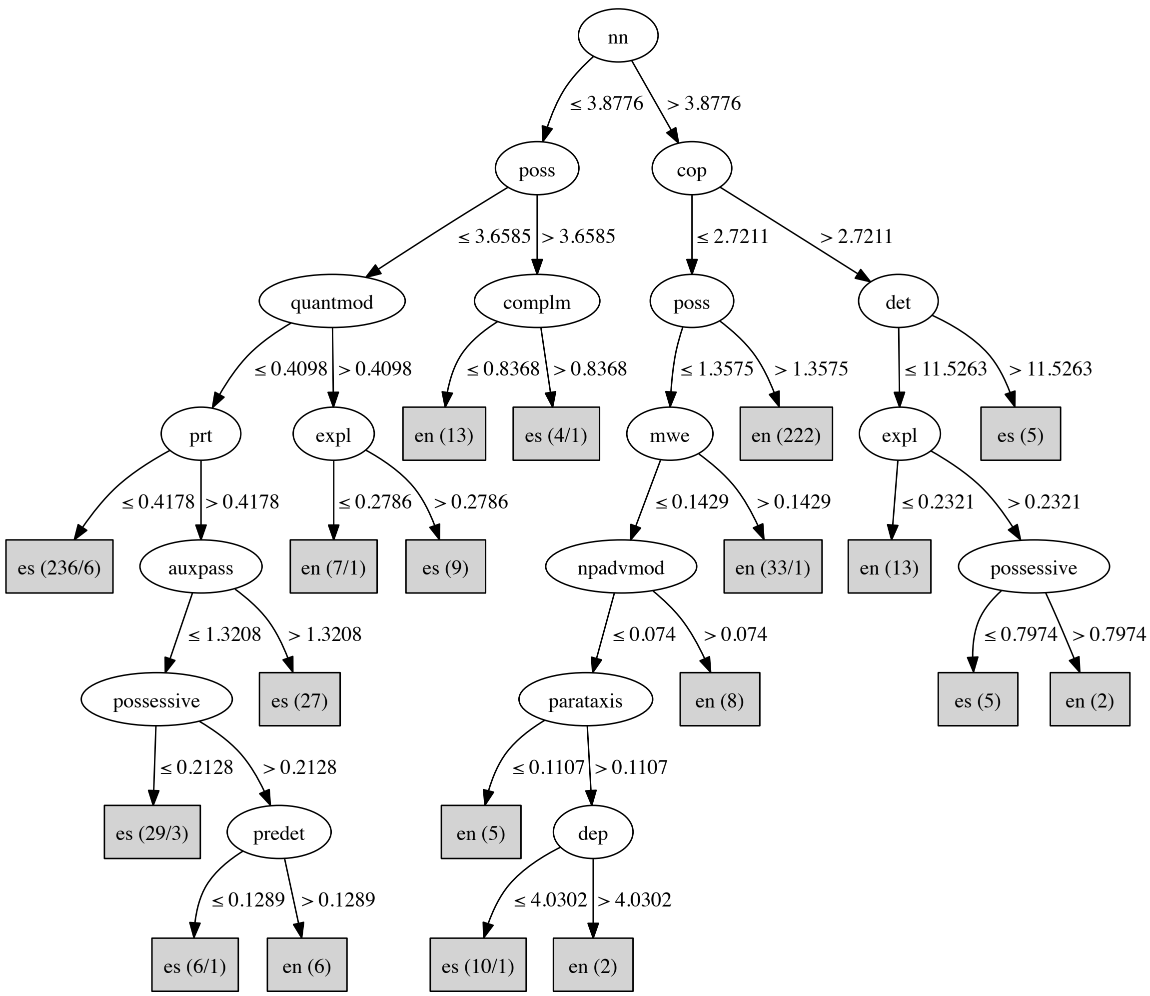

Figure 1.2: The dependencies $auxpass$(killed, been) and $auxpass$(killed, was/got). Taken from de Marneffe and Manning [2008].

auxpass

a. Kennedy has been killed

auxpass

b. Kennedy was/got killed

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

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 Example 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.

Figure 1.3: 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 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, 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 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 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.

(1.1) 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)

1.3 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.

1.4 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.

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 1.4 demonstrates this dependency.

Figure 1.4: 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 1.5. 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 1.5: 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

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 1.6 demonstrates this dependency.

Figure 1.6: 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 . . .

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 1.7. 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 1.7: 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 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 $\rightarrow$	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	

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 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.

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.

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