1 Verbs

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

7.1 Grammar

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

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

addition, there is the base form *be*, the present participle *being*, and the past participle *been*.

In addition to these basic inflected forms, the English verbal system relies on a broad array of auxiliary words. One such class of words are the modal auxiliaries or, simply, the modals. The nine English modals are shown in Table 1. These modals express concepts as basic as the future and the subjunctive, and others more subtle, such as intention, obligation, ability, and so forth. It is important to note that these modals, whether they reflect a change of tense or not, do not inflect to agree with the subject.

Table 1: English Verbal Forms Employing Modals

will walk	*can* walk	*should* walk
may walk	*could* walk	*must* walk
might walk	*shall* walk	*would* walk

Very similar to the modals are the phrasal modals. The thirteen phrasal modals considered in this experiment are those listed by Quirk et al. [1985, pp. 136-47], and are shown in Table 2. Of these, *dare to*, *need to*, *have got to*, *have to*, and all those beginning with the verb *be*, can be inflected to show tense, person, and number, and can generally be used in conjunction with modals. The remaining phrasal modals do not inflect to agree with the subject, and are restricted to the present tense, with the exception of *used to*, which is restricted to the past tense. It is worth noting that Quirk et al. [1985] do not group all of these into one category, but into three separate categories of modal-like auxiliaries: *marginal modals*, *modal idioms*, and *semi-auxiliaries*. Because a more complicated analysis of these constructions would add little to this study, the term phrasal modal is here applied to any multiword modal-like construction.

The words *have*, *be*, *get*, and *do*, which are full verbs in their own right, play an important role in English as auxiliary verbs. When used as such, they become the inflected element of the verb, being used in conjunction with a nonfinite form of the main verb. A form of the verb *have* followed by a past participle indicates the perfective aspect (e.g., *he*

Table 2: English Verbal Forms Employing Phrasal Modals

dare to walk	*used to* walk	**be** *about to* walk
need to walk	*had better* walk	**be** *able to* walk
ought to walk	*have got to* walk	**be** *bound to* walk
have to walk	**be** *supposed to* walk	**be** *willing to* walk
		be *obliged to* walk

*has walked/had walked/will have walked/*etc.). *Be*, in any of its forms, plays two roles as an auxiliary. Followed by a present participle it forms the progressive aspect (e.g., *he is walking/was walking/will be walking/*etc.). Followed by a past participle, it forms the passive mood (e.g., *he was pursued.*). A passive can also be formed using *get* plus a past participle (e.g. *he got hurt*). A finite form of *do* is used to add emphasis to a sentence and to form questions, negatives, and affirmative responses to questions, but only when no modal is present (e.g., *you do not drink wine* but *you will not drink wine*).

7.2 Parsing of Verbs

The Stanford Parser marks verbs, but does not explicitly mark all of the verbal attributes discussed above. The parse trees it generates indicate the structure of verb phrases (VPs), and distinguishes the various inflected forms of a verb. In general, it correctly distinguishes the finite and nonfinite uses of the base form of a verb. An example can be seen in Figure 1, which shows the parse of the sentence *I do not want to go.* Here the three verbs *do*, *want*, and *go* are tagged with the labels "VBP," "VB," and "VB," respectively. "VBP" indicates a present form other than 3rd- person singular (which is indicated by "VBZ"). "VB" marks a base form in a nonfinite usage. The parser also marks past-tense forms, *ing*-forms, and *ed*-forms, using "VBD," "VBG," and "VBN," respectively.

Any verbal information beyond that provided by these six tags must be determined from the shape and content of the VP subtrees in which the verbs are found. The parser directly marks modals, using the "MD" tag, as shown in Figure 2. However, it does not

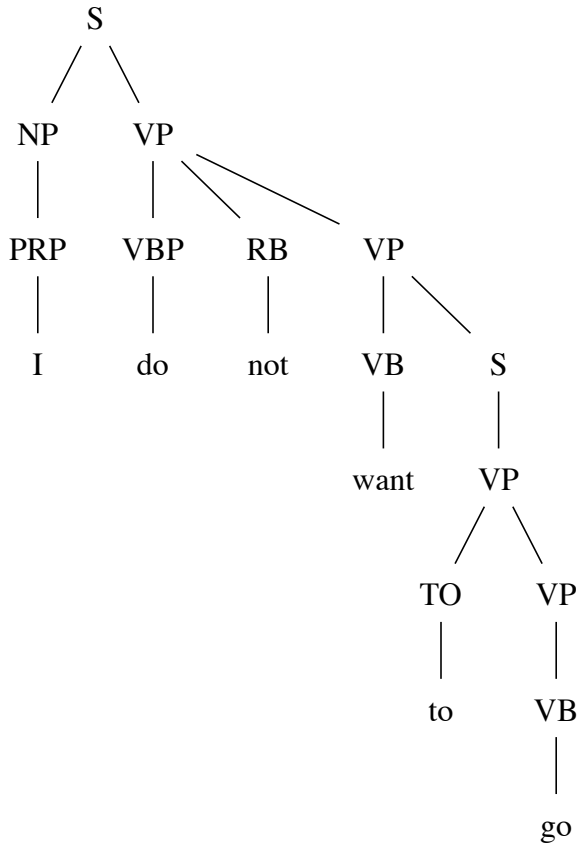

Figure 1: Typical Parse Tree Showing Verb Types

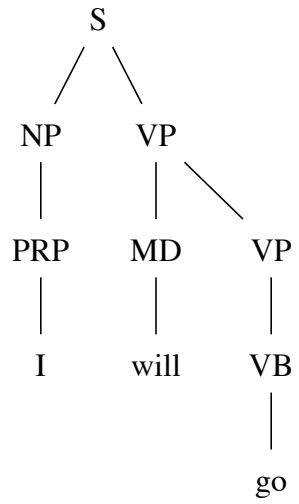

Figure 2: Parse Tree Showing Modal *will*

mark phrasal modals in a consistent manner. Figures 3 and 4 show how two phrasal modals are parsed differently. In Figure 3, *he is able to go* is parsed with *is* as the main verb, and with *able to go* being an adjectival phrase. In Figure 4, the sentence *he is going to go* is also parsed as a copular sentence, but with the predicate nominative being treated as a nonfinite clause. Indeed, the parser treats all phrasal modals of the form *be* + particle + *to* as *be* verbs followed by an adjectival or participial construction. Perhaps this is not surprising as the distinction between these constructions and phrasal modals is somewhat blurry in English grammar. Quirk et al. [1985, footnote, p. 144] indicate that the main criterion for distinguishing these is whether what follows *be* is able to stand alone at the beginning of a sentence. Consider, for instance:

(1) a. *Compelled to take stern measures, the administration lost popularity.*

b. *?Bound to take stern measures, the administration lost popularity.*

c. *Unable/Unwilling to resist, Matilda agreed to betray her country.*

d. *?Able/?Willing to resist, Matilda declined to betray her country.*
(Quirk et al. 1985, footnote, p. 144)

When fronted, the phrasal modals produce questionable sentences, as in (1b) and (1d). The non-phrasal modals, however, produce clearly acceptable sentences, shown in (1a) and (1c). Interestingly, by this criterion the negated phrasal modals (when negated on the lexical level) do not appear to be true phrasal modals and are not included as such in this study.

Phrasal modals are identified by searching parse trees for subtrees that match the basic form of the VP subtrees shown in Figures 3 and 4, and of other similar trees, while allowing for variation where appropriate. For instance, to match an instance of *be going to*, a subtree must be found that matches the most dominant VP subtree in Figure 4, with the exception that where the leaf [*go*] is found in the model, there may be any terminal node, and where the subtree [VBZ — *is*] is found, there may be any subtree representing a conjugation of *be* (e.g., [VBD — *was*], [VBP — *am*], etc.) The other phrasal modals and

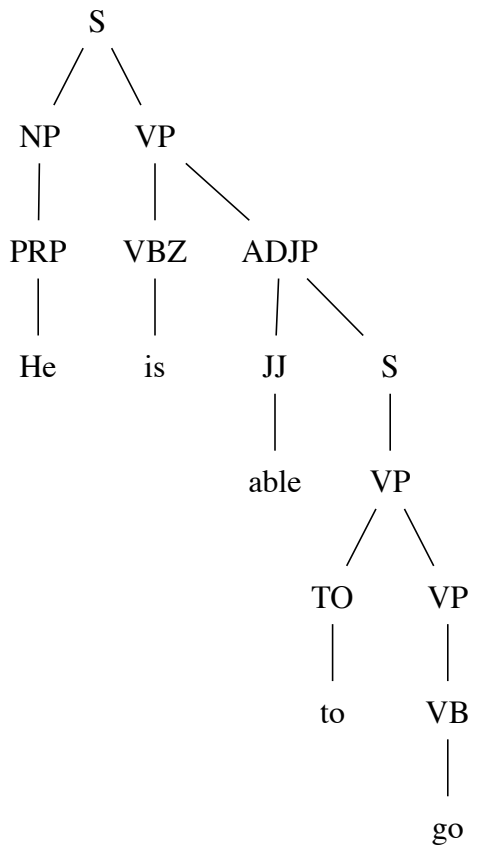

Figure 3: Parse Tree Showing Phrasal Modal *be able to*

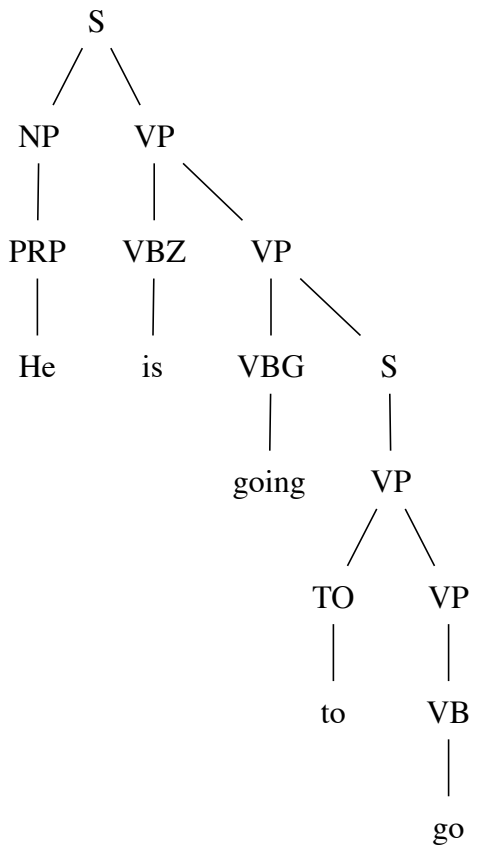

Figure 4: Parse Tree Showing Phrasal Modal *be going to*

constructions using the auxiliary verbs *do*, *have*, and *be* are similarly identifiable using other distinctive subtrees.

Whenever the conjunction *and* is encountered in a VP subtree, multiple distinct but overlapping subtrees are generated from this and treated independently. For instance, the parse of the sentence *he can design and build houses* shown in Figure 5 is processed to generate the two separate parses *he can design houses* and *he can build houses* shown in Figure 6. The exception to this is when the conjunction is used to apply multiple modals to a verb (e.g. *I can and will...*).

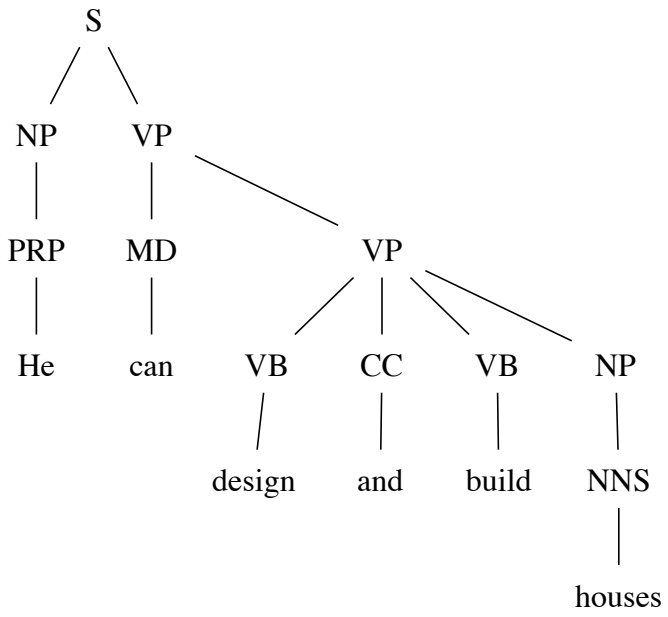

Figure 5: Parse Tree Showing a Verb with Embedded Conjunction

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

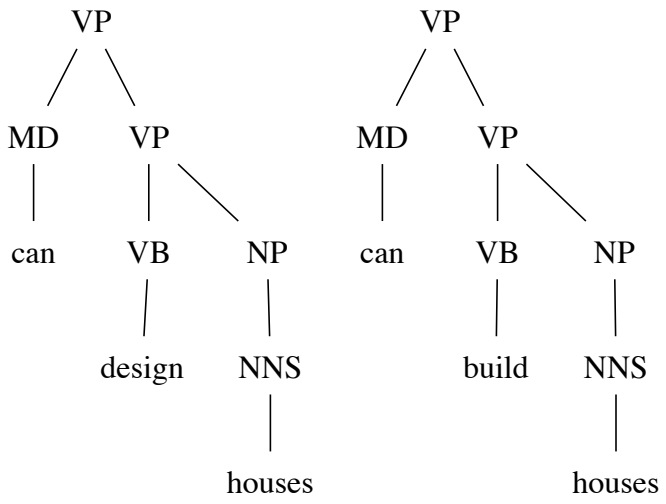

Figure 6: Parse Trees Showing the VP from Figure 5 Split into Two VPs

7.3 Classification Accuracy

Being able to identify so many attributes gives a wealth of data with which a classifier may be trained. The challenge, as always, is choosing subsets of this data that maximize information content while still producing decision trees comprehensible to a human. The first such subset examined here consists of eight attributes, each with a value indicating the relative frequency with which such verbal qualities are found in a text. These eight attributes are named "not," "modal," "progressive," "past," "perfect," "passive," "quasimodal," and "do." For the most part these are self-explanatory, with "not" being the relative frequency of the negating adverb *not* and so forth ("quasimodal" indicates a phrasal modal). It is worth pointing out that with the exception of "modal," these verbal attributes can only occur once per verbal construction, but of course all can appear many times in a given text.

A C4.5 decision tree generated from this data set is shown in Figure 7; and the accuracy of such classifiers, calculated using 20-fold cross-validation, is shown in Table 3. As can be seen, this is quite a large tree for purposes of analysis, with many attributes appearing multiple times. A somewhat simpler tree, with only a modest lost of accuracy,

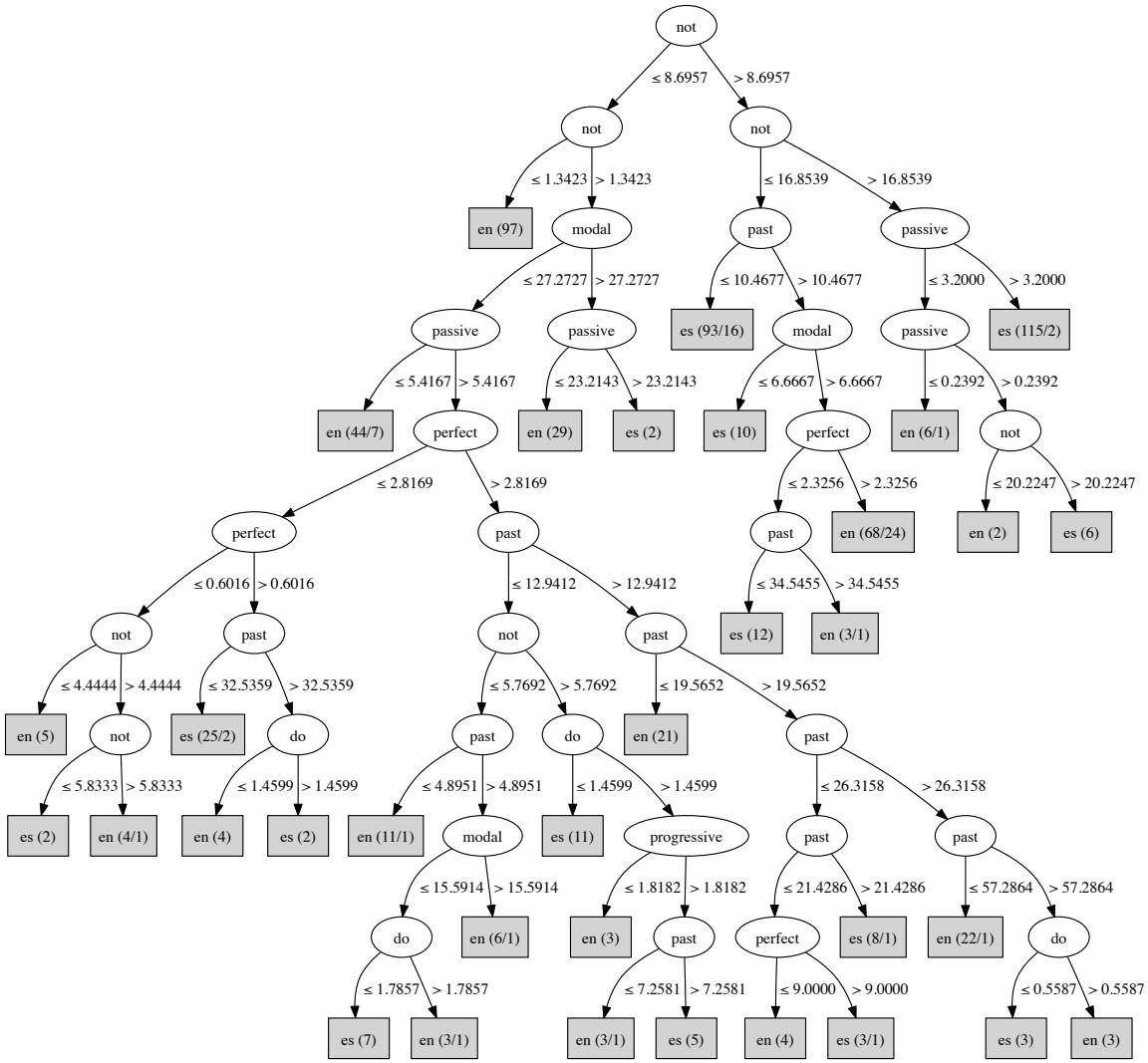

Figure 7: C4.5 Decision Tree Using Various Verbal Attributes With Default Pruning

Table 3: Accuracy of C4.5 Classifier Using Various Verbal Attributes With Default Pruning

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

can be had by using more aggressive pruning following the tree construction phase. Figure 8 shows such a tree, and Table 4 its accuracy. This tree consists of 13 decision nodes considering 6 different attributes, none of which are considered more than three times. Of the eight attributes used in training, "quasimodal" and "progressive" are not

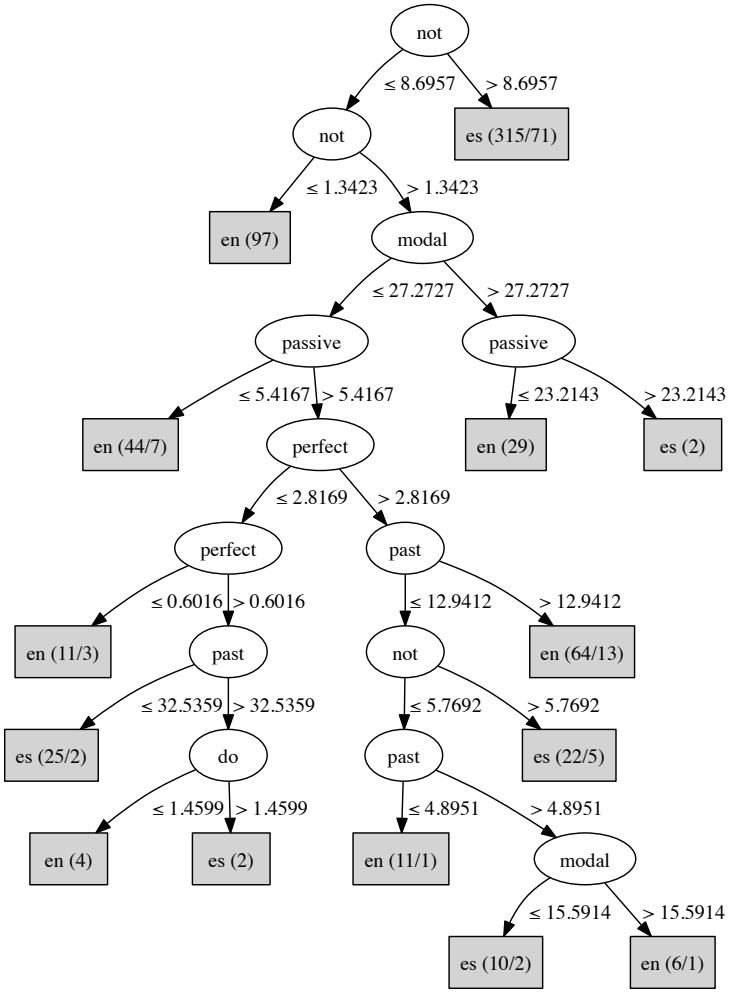

Figure 8: Aggressively Pruned C4.5 Decision Tree Using Various Verbal Attributes

found in the aggressively pruned tree, nor does "quasimodal" appear in the normally-pruned tree. "Modal," however, does appear in both trees. In the aggressively-pruned tree it appears twice, once very near the root of the tree and the other time as the deepest decision node. The first of these splits the training cases into a predominately nonnative subset, those with frequencies higher than 27.2727%, and a predominately native subset with lower values. The second decision node, however, uses a

Table 4: Accuracy of Agressively-Pruned C4.5 Classifier Using Various Verbal Attributes

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

lower comparison value and splits the cases the other way, with low values being classified as nonnative. This would seem to indicate that extremes in modal usage are associated with nonnative usage. Before attempting to explain this, it is worth exploring the role of modal verbs in Spanish. According to the analysis of Butt and Benjamin [2004], Spanish has a small array of modal verbs, in general similar syntactically to English phrasal modals, as shown in the following examples:

(2) a. *No debiste hacerlo.*
Not you-should-have done-it.
'*You shouldn't have done it.*'

b. *Hubo de repetir el experimento.*
(S)he-had-to repeat the experiment.
'*(S)he had to repeat the experiment'*
[Butt and Benjamin 2004, pp. 327-30]

Some of these are usually translated into English using modals or phrasal modals, and some require the use of other constructions [Butt and Benjamin 2004, pp. 325-32]. Conversely, many English modals and phrasal modals can be translated into Spanish using that language's modals, while others translate as verbal inflections. In addition, for those Spanish modals which can be translated into English modals or phrasal modals, the correspondence is rarely one-to-one, with there being a great deal of overlap and inexactness in meaning both ways. This should not be surprising, considering the close but imperfect correspondence between many of English's modals and phrasal modals (e.g. *can* and *be able to*) [Celce-Murcia and Larsen-Freeman 1999, pp. 137-157].

As mentioned above, the Spanish modals have much in common with the English phrasal modals. In general, the pattern of the Spanish modals is *conjugated modal verb* +

one or zero particles + infinitive. As was seen in Table 2, many of the English phrasal modals follow this pattern as well. With a few exceptions, the Spanish modals can take the full range of verbal inflections, as can the majority of English phrasal modals. From this one might assume that L1-Spanish learners of English would take naturally to the English phrasal modal. Indeed, the decision trees shown in Figures 7 and 8 would seem to support this, or at least support the proposition that the learners neither overuse nor underuse phrasal modals relative to their native counterparts. A quick experiment in which the C4.5 classifier was run on the same data set, but with the "modal" attribute removed, showed that the "quasimodal" attribute was still not used in the resultant decision tree. This means that the reason for the algorithm's excluding the "quasimodal" attribute cannot be attributed to its containing no information beyond what the "modal" attribute contains.

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

In Figure 8, it could be seen that the attribute with the highest information content is "not." This is used three times, with higher frequencies tending to lead to classification as nonnative and lower frequencies to native. It is tempting to attribute this to Spanish's double negative construction [Butt and Benjamin 2004, pp. 344-5], but it seems doubtful that advanced English learners would not have grasped this basic difference between English and Spanish grammar. More likely is that native speakers, with their presumably larger vocabularies, have a greater number of lexical negatives (e.g. *he is unkind* versus *he*

is not kind) at their disposal.

Considering next the "passive" attribute, it can be seen that this attribute, too, is used in a consistent manner in the decision tree, with lower and higher frequencies leading to classification as native and nonnative, respectively. Spanish expresses passives in primarily two ways: using the copular verb *ser* plus a past participle in a construction similar to the English passive, or, more commonly, using the reflexive pronoun *se*:

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

b. *Se les devolverán las muestras.*
to-you will-return the samples.
'The samples will be returned to you.'
[Butt and Benjamin 2004, pp. 402,8]

Based on these constructions, it is not surprising that L1-English learners of Spanish tend to overuse the *ser*-passive [Butt and Benjamin 2004, p. 406]. That L1-Spanish learners of English do the same with the *be*-passive is more surprising. Indeed, this author was unable to find any literature that investigates, or even acknowledges, this phenomenon. One possibility has to do with the connection between the English passive and the presentation of information at the discourse level. The English passive is frequently used to reverse the order of what, in an active sentence, would be the subject and object. Such a reversal is often necessary to preserve the tendency in language to present old information before new information. However, the passive is only one of a number of constructions, the least marked such construction, perhaps, that allow the fronting of a particular element in a sentence [Ward and Birner 2008]. Other such constructions (e.g. preposing, inversion), being more complicated and alien, may be avoided by the learner, leading to an overuse of the passive.

The remaining three attributes used in the decision tree are rather resistant to analysis. The perfect aspect constructions in English and Spanish, for instance, are remarkably similar, with English using a form of the verb *to have* plus the past participle

and Spanish using a form the verb *haber* (cf. Latin *habere*, *to have*), plus a past participle. That the frequency of usage of the perfect should be a useful metric in classification is surprising. Similarly, English and Spanish both have inflected past tenses which, combined with the complex role the "past" attribute plays in the decision tree, makes deciphering that attribute difficult. Finally, while Spanish has nothing quite like English's *do* auxiliary, the extremely limited role which the "do" attribute plays in the classification process (it is only involved in classifying six out of 642 training cases) suggests that it is of little utility.

The next set of attributes deals with the relative frequencies of the various possible modals and phrasal modals. Table 2 shows the phrasal modals parsed in this system, of which there are 13; and Table 1 shows the modals, of which there are 9. This yields a total of 22 attributes. Training a C4.5 classifier on this data set produces a decision tree which is rather opaque in terms of interpretation, but the accuracy of such a classifier can be see in Table 5. A derivative data set, which sacrifices accuracy for interpretability, produces

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

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

the C4.5 tree shown in Figure 9. The attributes used in this tree indicate whether there is a tendency in a text to use a modal over a phrasal modal with a similar meaning. Only four attributes were used, derived from a total of eleven modals and phrasal modals. Table 6[1] shows the modals used and the corresponding phrasal modals. The values for these attributes were calculated by subtracting from the number of occurrences of a particular modal the number of occurrences of the corresponding phrasal modals, and dividing the difference by the sum of these two quantities. This yields a real number ranging from -1

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

to 1, with the most negative number indicating all phrasal modals and no modals, zero indicating an even number of each, and 1 indicating all modals and no phrasal modals. For the purposes of display in Figure 9, these numbers are scaled by a factor of 100. The labels given to the attributes are the names of the modal, these being unique for each correspondence.

Table 6: Correspondence Between Modals and Phrasal Modals

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

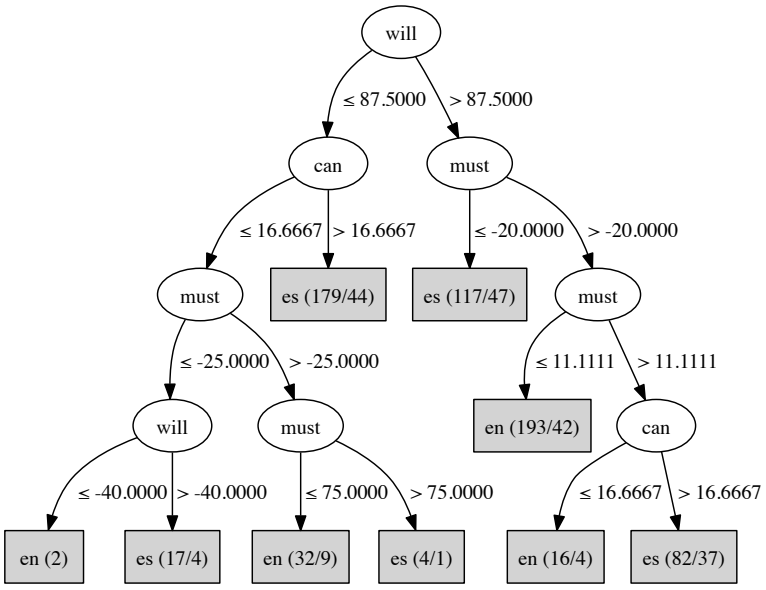

Figure 9: C4.5 Decision Tree Using Modal vs Phrasal Attributes

The resultant tree uses three of the four attributes, with the "should" attribute not being included. The tree indicates that there is a strong tendency for learners to use core

Table 7: Accuracy of C4.5 Classifier Using Modal vs Phrasal Attributes

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

modals more often relative to phrasal modals than native writers. This is despite the similarity between English phrasal modals and Spanish modals. Table 6 also shows common Spanish modal equivalents to the English modals and phrasal modals. As can be seen from this table, each of these English modals can be expressed using a common Spanish modal, including *will*/*be going to*, which has a Spanish modal equivalent that is used alongside Spanish's inflected future [Butt and Benjamin 2004, p. 221]. The best explanation for this is that learners prefer the core modals due to their syntactic simplicity.

The third set of attributes considers various common English verbs which are shown in Table 8. These verbs are identified as common verbs that tend to be overused by

Table 8: High Frequency Verbs

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

English learners in a study by Ringbom [1998]. This study uses an earlier version of the ICLE, examining the French, Spanish, Finnish, Swedish, Dutch, and German subcorpora for usages of these verbs, and comparing the frequency of usage to that found in a native subcorpus of the ICLE[2]. Ringbom gives the breakdown per word, and finds that not all of the verbs show overuse in the Spanish subcorpus, with only *use, believe, feel*, and *come* showing overuse. Ringbom does not attempt to establish statistical significance, however. In the present study, these verbs were used to construct a data set consisting of one attribute per verb with the value of each being equal to the relative frequency of that verb

[2]The version of ICLE used in the present study contains no native subcorpus.

when used as a main verb in a finite clause. The accuracy of a C4.5 classifier trained on this data set is shown in Table 9.

Table 9: Accuracy of C4.5 Classifier Using High Frequency Verb Attributes

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

Finally, to gauge the efficacy of verbal attributes in general, classifiers were trained on a combined data set consisting of all four sets of verbal attributes discussed here. Both a C4.5 classifier and a random forest classifier was tried. The accuracy of these classifiers is shown in Table 10 and Table 11. Combining the various attributes sets results in an

Table 10: Accuracy of C4.5 Classifier Using All Verbal Attributes

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

Table 11: Accuracy of Random Forest Classifier Using All Verbal Attributes

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

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

References

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.

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

RINGBOM, H. 1998. High-Frequency Verbs in the ICLE Corpus. In *Explorations in corpus linguistics*, A. Renouf, Ed. Language and Computers: Studies in Practical Linguistics. Rodopi.

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