A System to Distinguish Native and Nonnative Written English

By Philip White

Goals

- Develop classifiers to distinguish native written English from that written by L1-Spanish speakers.
- Focus on advanced learner English (i.e., don't rely on detecting errors).
- Design the classifier such that it could be used as the basis of a learning tool.

Resources Used

- Corpora of native and nonnative English
- Stanford Parser
- Weka Machine Learning Tools

Corpora

Corpus	Tokens Native	Tokens Nonnative
BROWN	57,809	0
ICE-HK	59,674	0
MICUSP	163,218	29,897
MSUELI	0	538
OANC	84,052	0
SPICLE	0	216,879
SULEC	0	39,254
WRICLE	0	96,247
ICE-CAN	25,225	2,070
Total	389,978	384,885

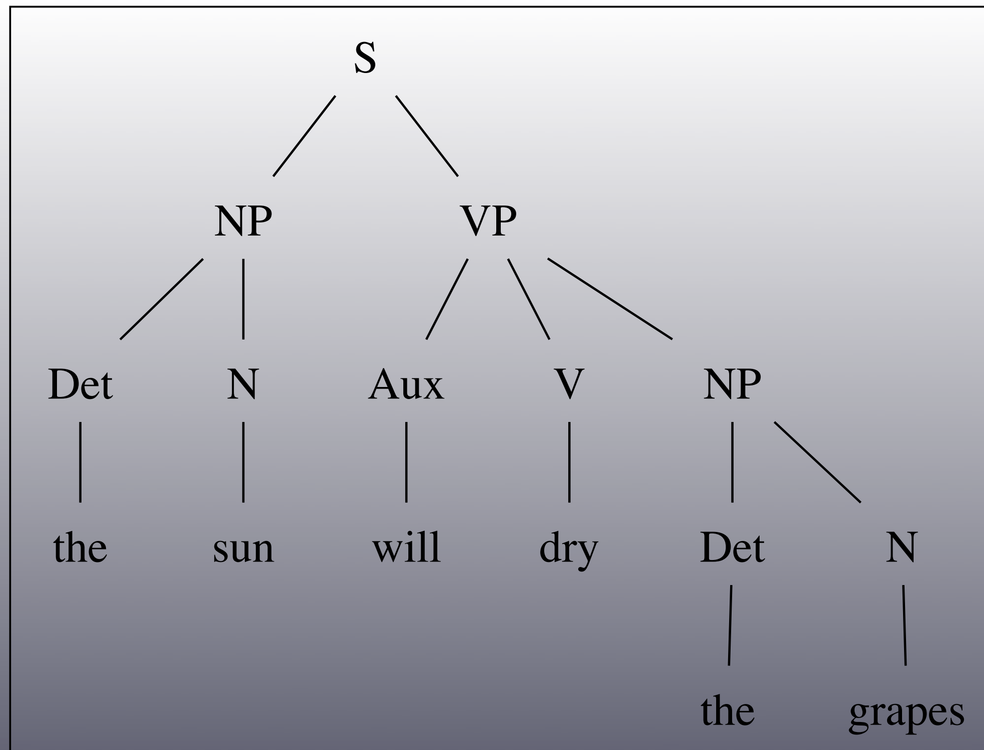
321 Native Samples and 321 Nonnative Samples

Stanford Parser

- Probabilistic parser
- Developed by Stanford
- Generates parse trees and dependency relationship.
- Parse trees can contain up to about 40 different node types.
- Parser generates 58 different dependencies.

Parse Trees

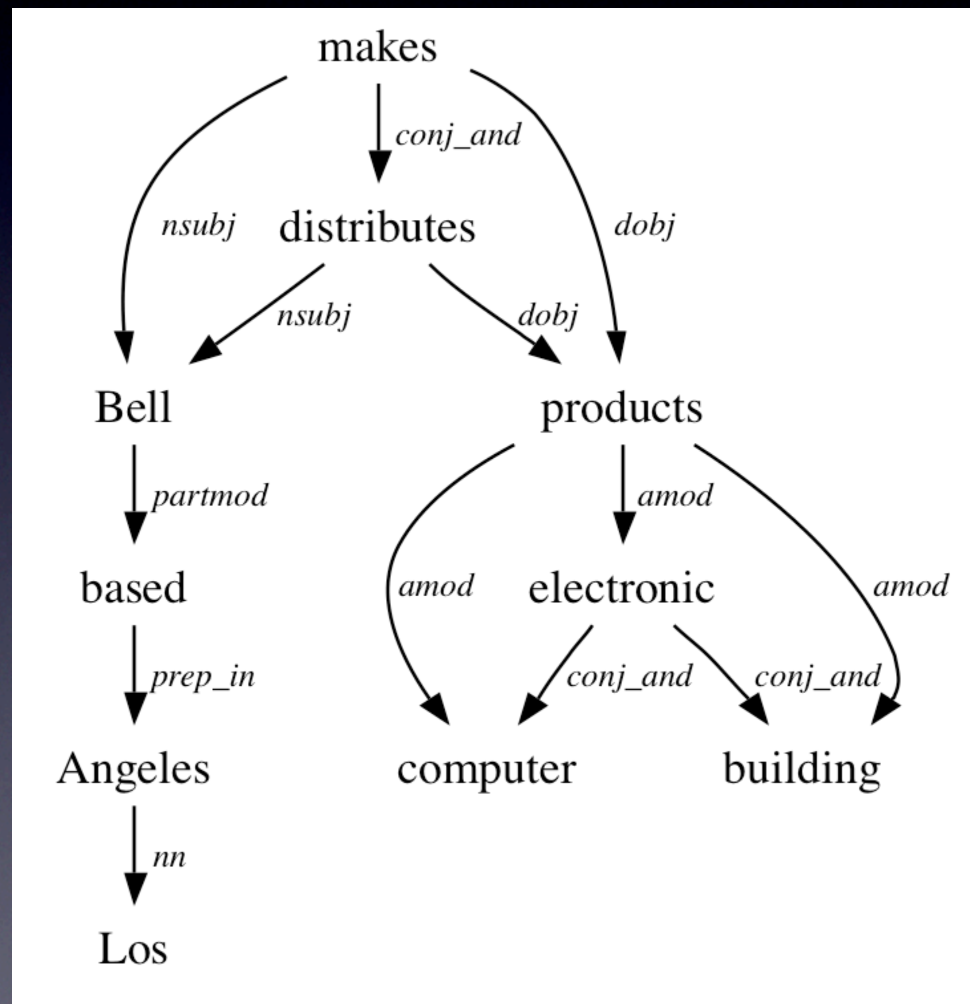
$$
\begin{array}{lll}
\text{a.} & \text{S} & \rightarrow \text{NP VP} \\
\text{b.} & \text{NP} & \rightarrow \text{(Det) N (PP)} \\
\text{c.} & \text{VP} & \rightarrow \text{(Aux) V (NP) (AdvP)}^{n} \\
\text{d.} & \text{PP} & \rightarrow \text{P NP} \\
\text{e.} & \text{AdvP} & \rightarrow \left\{ \begin{array}{c} \text{Adv} \\ \text{PP} \end{array} \right\}
\end{array}
$$



Dependencies

- Dependencies are 3-tuples: a relation, a governor, and a dependent.

  nn(Angeles,Los)
  prep_in(based,Angeles)
  partmod(Bell,based)
  nsubj(makes,Bell)
  nsubj(distributes,Bell)
  etc.



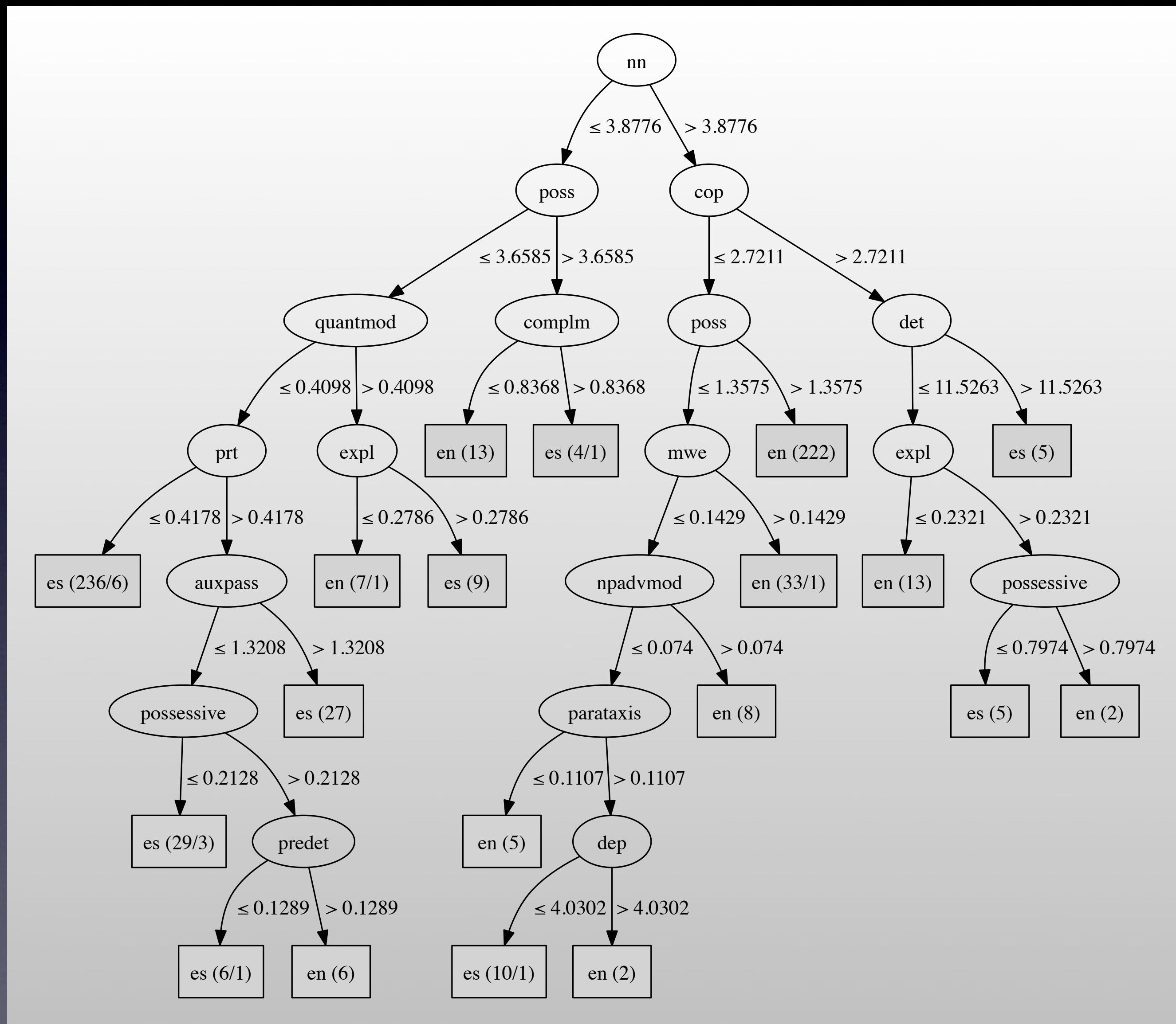
Weka

- Machine Learning
  - Classify instances into classes.
  - Instances are represented as lists of labeled values.
  - Each instance has a class value, which may be unknown.
  - Training and testing instances have known classes.

Classifiers

- C4.5 Decision Tree (J48)
  - Building Stage: constructs tree to divide training sets into subsets of a single class.
  - Pruning Stage: simplifies and generalizes tree.
- Random Forest
  - Uses multiple trees each using a subset of the total number of attributes.
  - Trees vote on best class.

Decision Tree



The Experiments

- Performed three experiments.
- Each experiment consisted of:
  - Extracting grammatical features from parse trees or dependency relations
  - Generating C4.5 classifier trees
  - Analyzing the features used in the trees to determine the linguistic significance of the features used
  - Measuring accuracy of classifiers (C4.5 and R.F.)

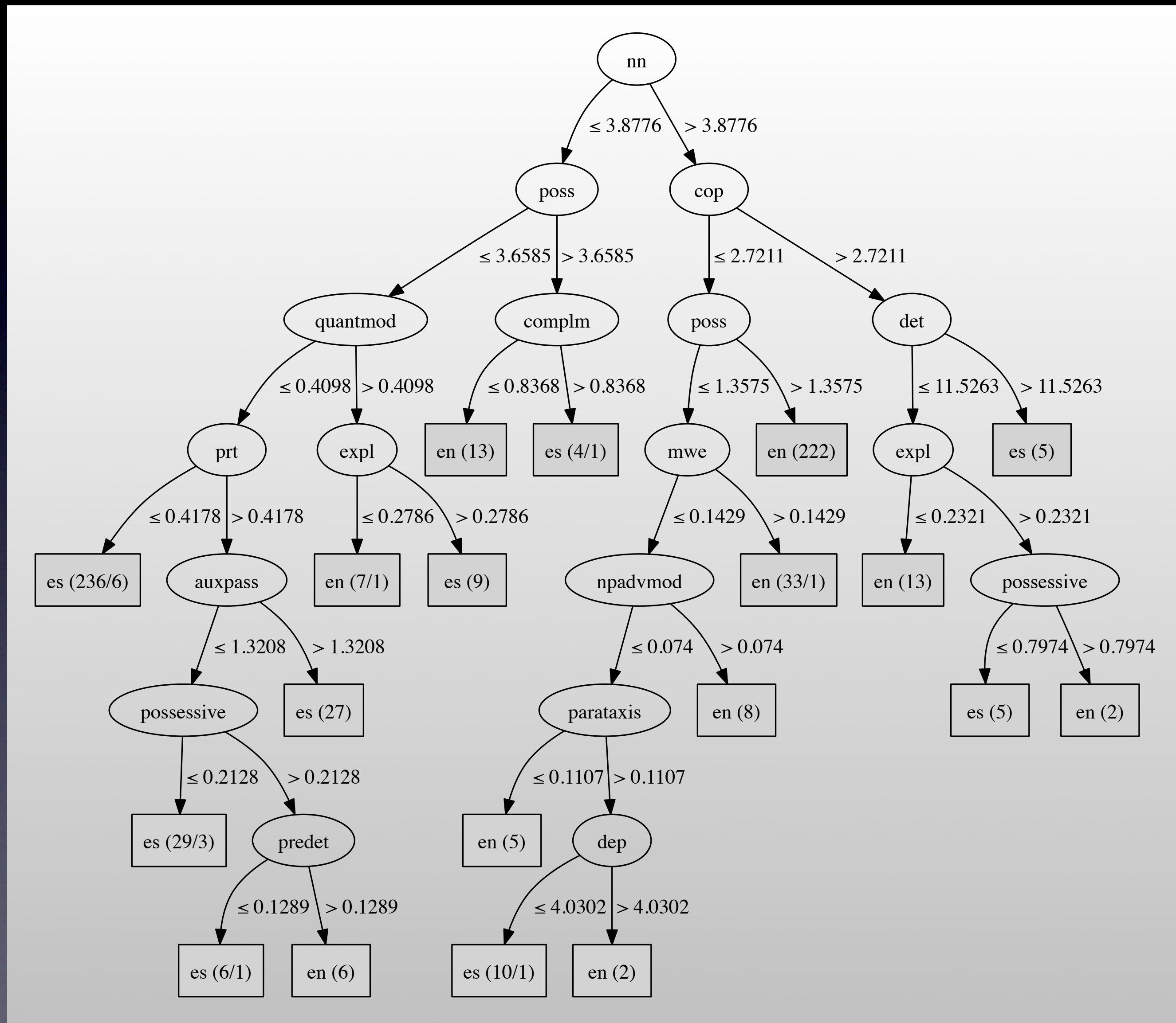
Dependency Experiment

- One attribute per dependency relation
- Value associated with each dependency is the number of occurrences of that relation divided by the number of occurrences of all relation.

Dependency Experiment

Label	Dependency
auxpass	passive auxiliary
complm	complementizer
cop	copula
dep	generic dependency
det	determiner
expl	expletive
mwe	multi-word express.
nn	noun compound
npadvmod	np as adv modifier
parataxis	parataxis
poss	possession modifier
possessive	possessive modifier
predet	predeterminer
prt	phrasal verb particle
quantmod	quantifier

Decision Tree



Dependency Experiment

C4.5 Tree Accuracy	
Nonnative	89.7%
Native	87.9%
Overall	88.8%
95% C.I.	86.3% – 91.2%

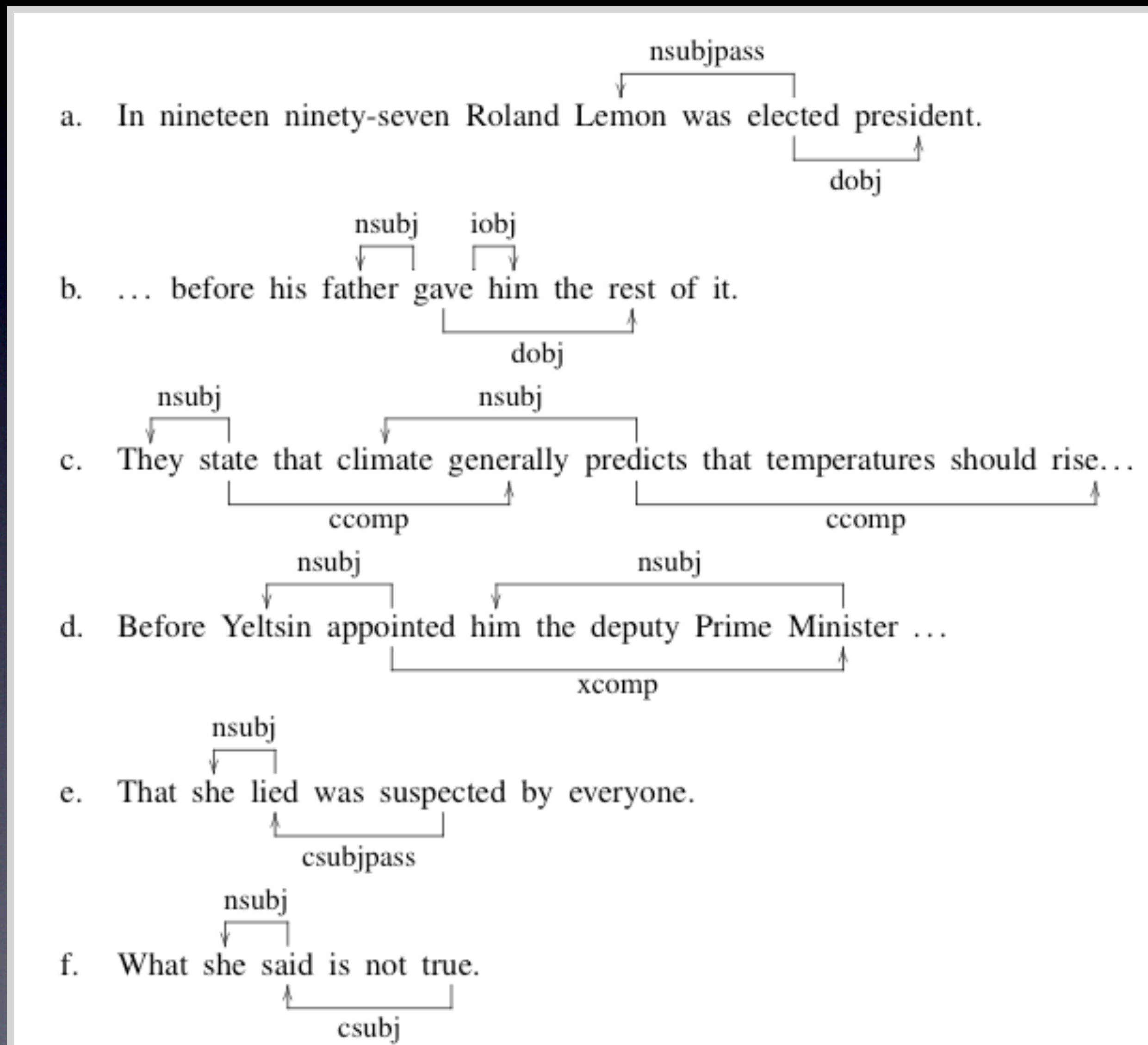
Random Forest Accuracy	
Nonnative	96.0%
Native	91.9%
Overall	93.9%
95% C.I.	92.1% – 95.8%

Verbal Argument Experiment

- Verbal argument roles:
  - Subject, object, indirect object, subject complement, patient, etc.
  - Derivable from dependency graphs
- An argument can either be a pronoun or a lexical noun.
- From this a number of features can be derived.

Du Bois, J. W. 2003. *Discourse and Grammar*.

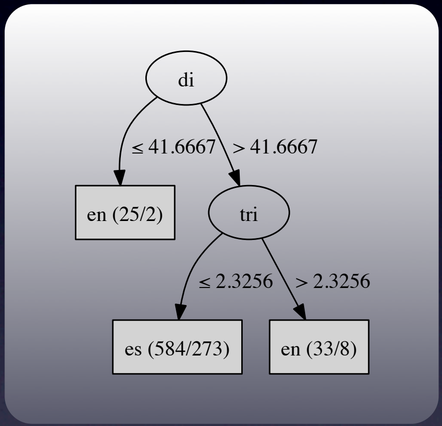
Verbal Arguments

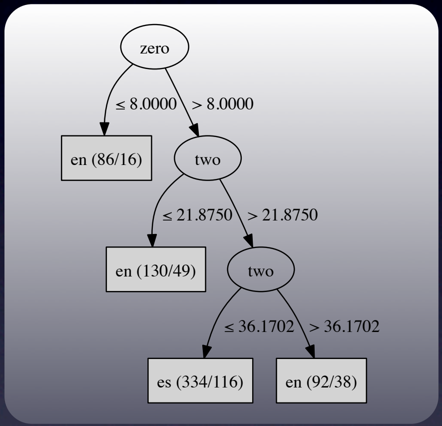


Verbal Arguments

Pronominal Forms					
other	another	else	same	one	
this	that	these	those	what	
myself	yourself	herself	himself	itself	
ourselves	yourselves	themselves	oneself		
mine	yours	hers	his	ours	theirs
me	you	her	him	it	us
them	I	she	he	we	they

Verbal Argument Experiment





C4.5 Tree Accuracy Verb Valency	
Nonnative	84.4%
Native	17.4%
Overall	50.9%
95% C.I.	47.1% – 54.8%

C4.5 Tree Accuracy Lexical Argument Density	
Nonnative	70.7%
Native	48.6%
Overall	58.7%
95% C.I.	55.9% – 63.5%

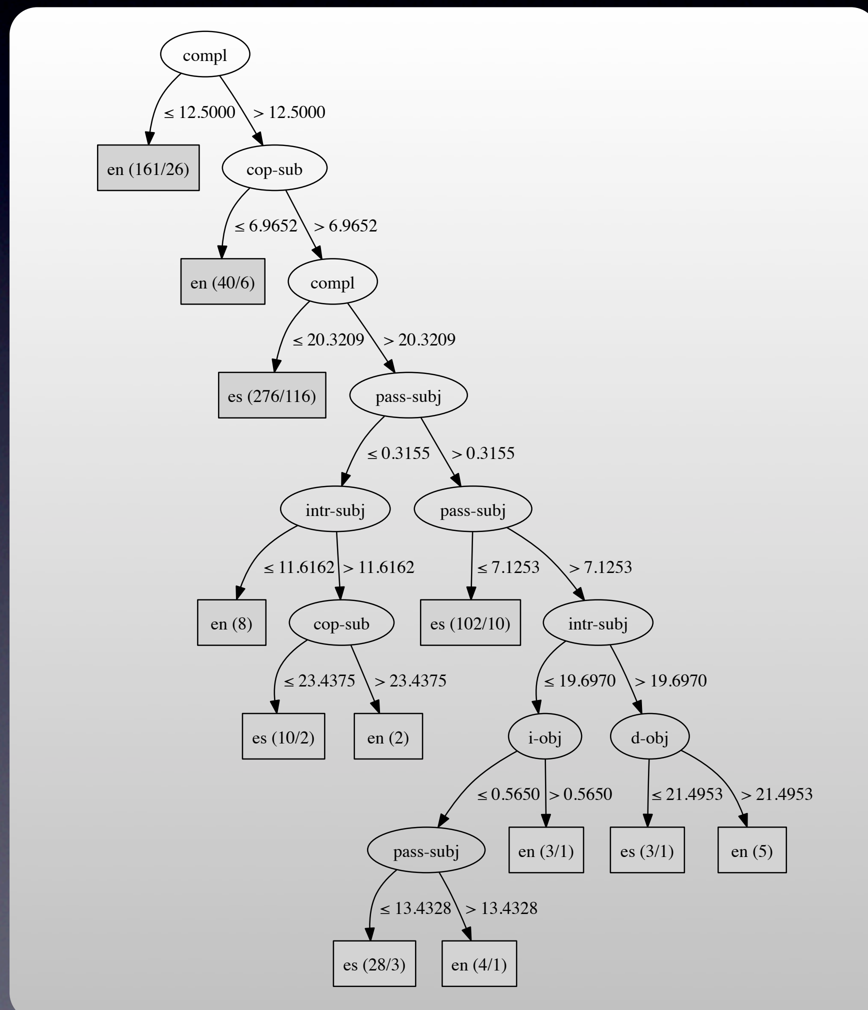
Verbal Argument Experiment

C4.5 Tree Accuracy
Lexical Argument Role

Nonnative	78.5%
Native	54.8%
Overall	66.7%
95% C.I.	63.0% – 70.3%

Random Forest Accuracy
Lexical Argument Role

Nonnative	76.6%
Native	63.2%
Overall	70.0%
95% C.I.	66.4% – 73.5%



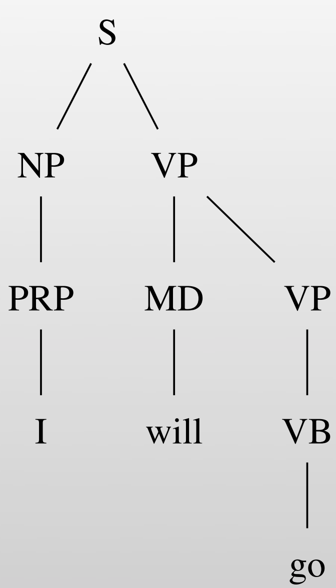
Verbal Argument Experiment

Random Forest Accuracy Combined Attributes	
Nonnative	76.6%
Native	63.9%
Overall	70.2%
95% C.I.	66.7% – 73.8%

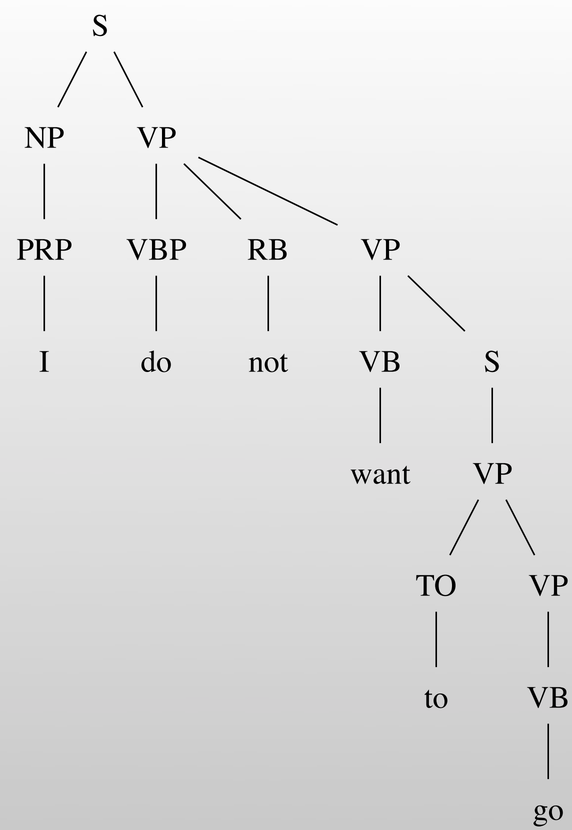
Verb Form Experiment

- Extract verb details from parse trees
- Identifies the following features:
  - Tense, aspect, voice, core modals, phrasal modals, the helping verbs *be*, *get*, *have*, *do*, and the negative particle *not*
- Also considers common verbs

Verb Form Experiment

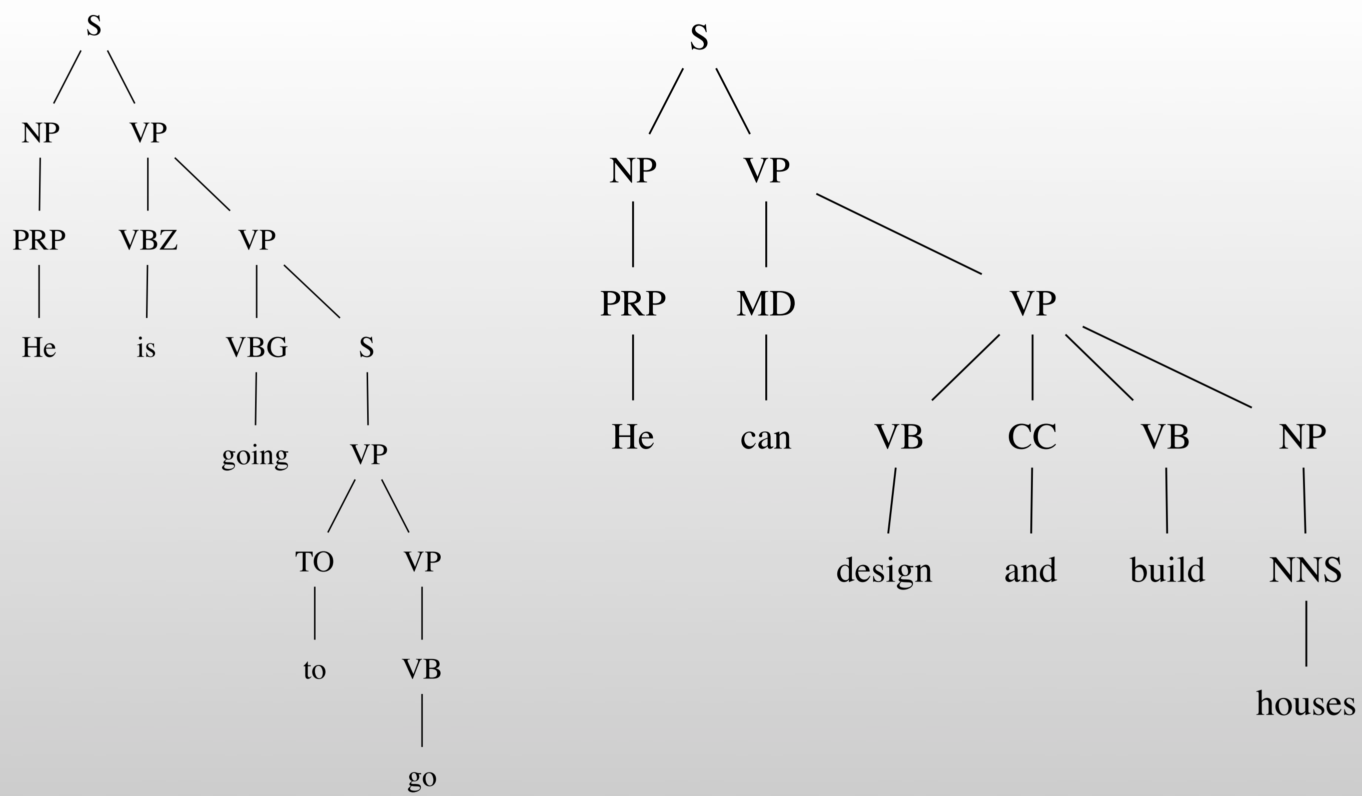


Core Modal



Verb with *not* and *do*

Verb Form Experiment

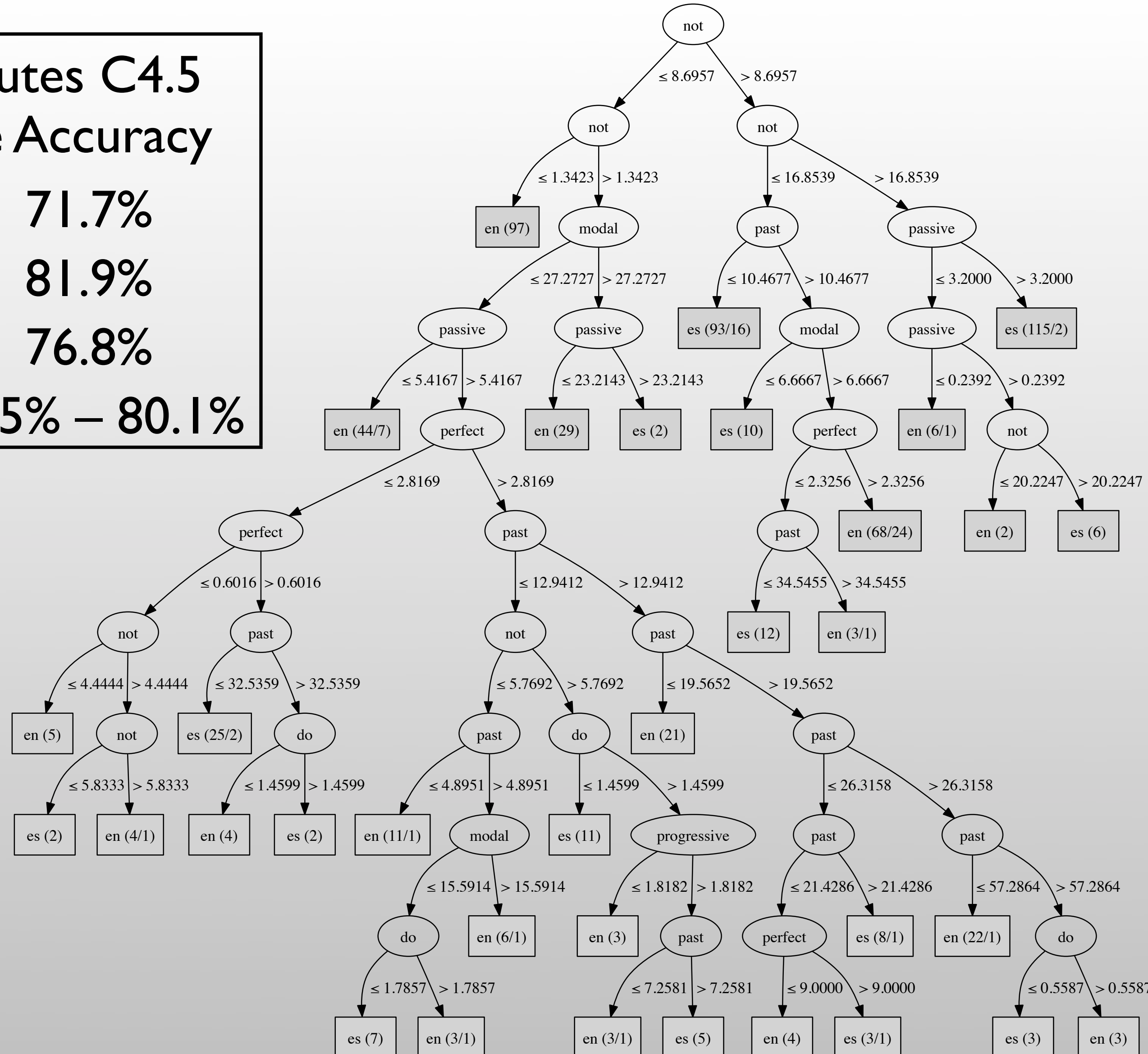


Phrasal Modal

Core Modal with Two Main Verbs

Verb Form Experiment

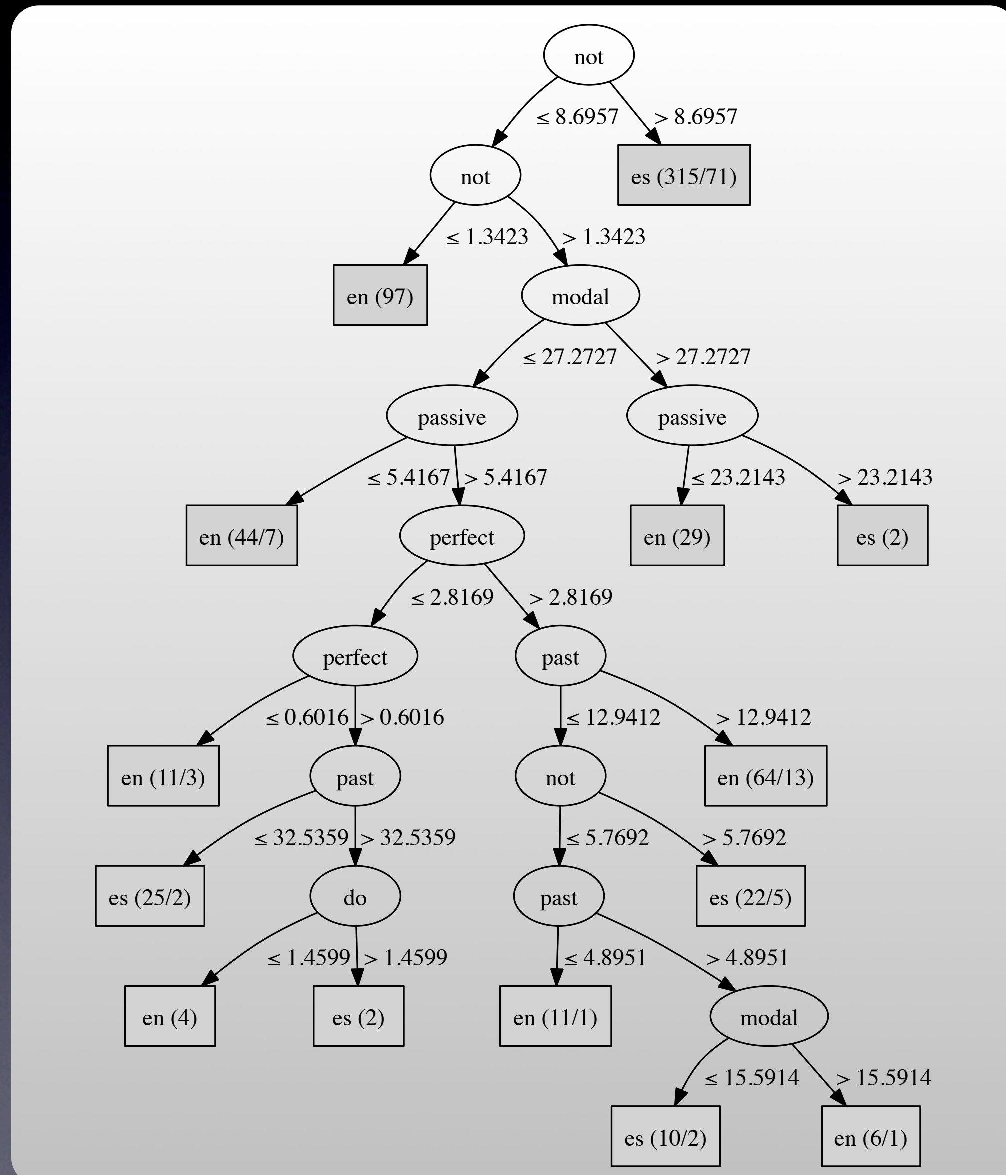
Verbal Attributes C4.5 Decision Tree Accuracy	
Nonnative	71.7%
Native	81.9%
Overall	76.8%
95% C.I.	73.5% – 80.1%



Verb Form Experiment

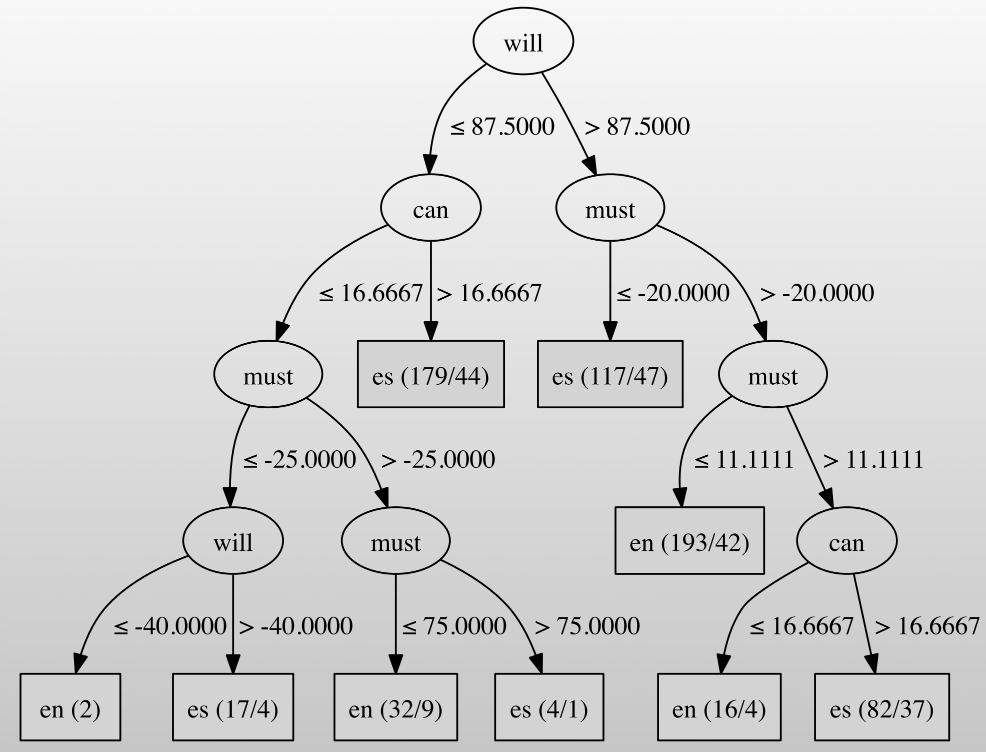
Verbal Attributes Simplified
C4.5 Decision Tree Accuracy

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



Verb Form Experiment

Core vs Phrasal Modal C4.5 Decision Tree Accuracy	
Nonnative	70.7%
Native	61.4%
Overall	66.0%
95% C.I.	62.3% – 69.7%



Verb Form Experiment

This experiment uses the relative frequency of finite usages of certain common verbs as classifier attributes.

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

Ringbom, H. 1998. *High-Frequency Verbs in the ICLE Corpus.*

High Frequency Verb C4.5 Decision Tree Accuracy	
Nonnative	70.7%
Native	61.4%
Overall	66.0%
95% C.I.	62.3% – 69.7%

Verb Form Experiment

Combined Verb Attributes Random Forest Accuracy	
Nonnative	80.4%
Native	77.3%
Overall	78.8%
95% C.I.	75.7% – 82.0%

Combined Verb Attributes C4.5 Decision Tree Accuracy	
Nonnative	90.3%
Native	85.4%
Overall	87.9%
95% C.I.	85.3% – 90.4%

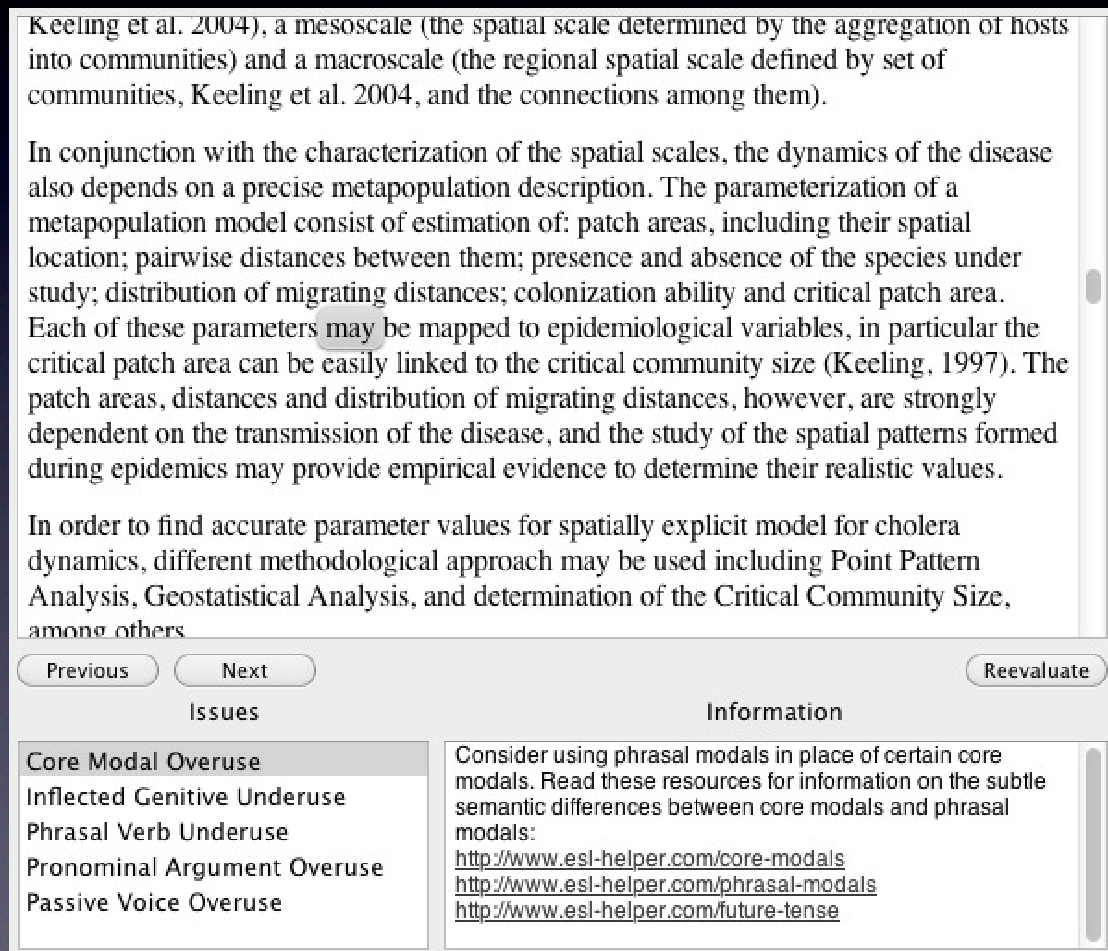
Other Possible Attributes

- Syntactic complexity (parse tree depth, phrase nesting, etc.)
- Types of phrases
- Vocabulary (of structural words in particular)
- etc.

Learning Tool

- Use classifiers to determine if text is identifiably nonnative.
- Inform user which features were responsible for classification as nonnative.
- Show user which parts of the text exhibit these features.
- User edits and reevaluates until text is classified as native.

Learner Tool Mock-up



Other Possible Applications

- Identifying plagiarism among ESL student.
- Determining first language of the writer of a text sample (forensics?).