A System to Distinguish Between Native and Nonnative Written English

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 classifiers such that they can be used as the basis of a learning tool.

Basic Process

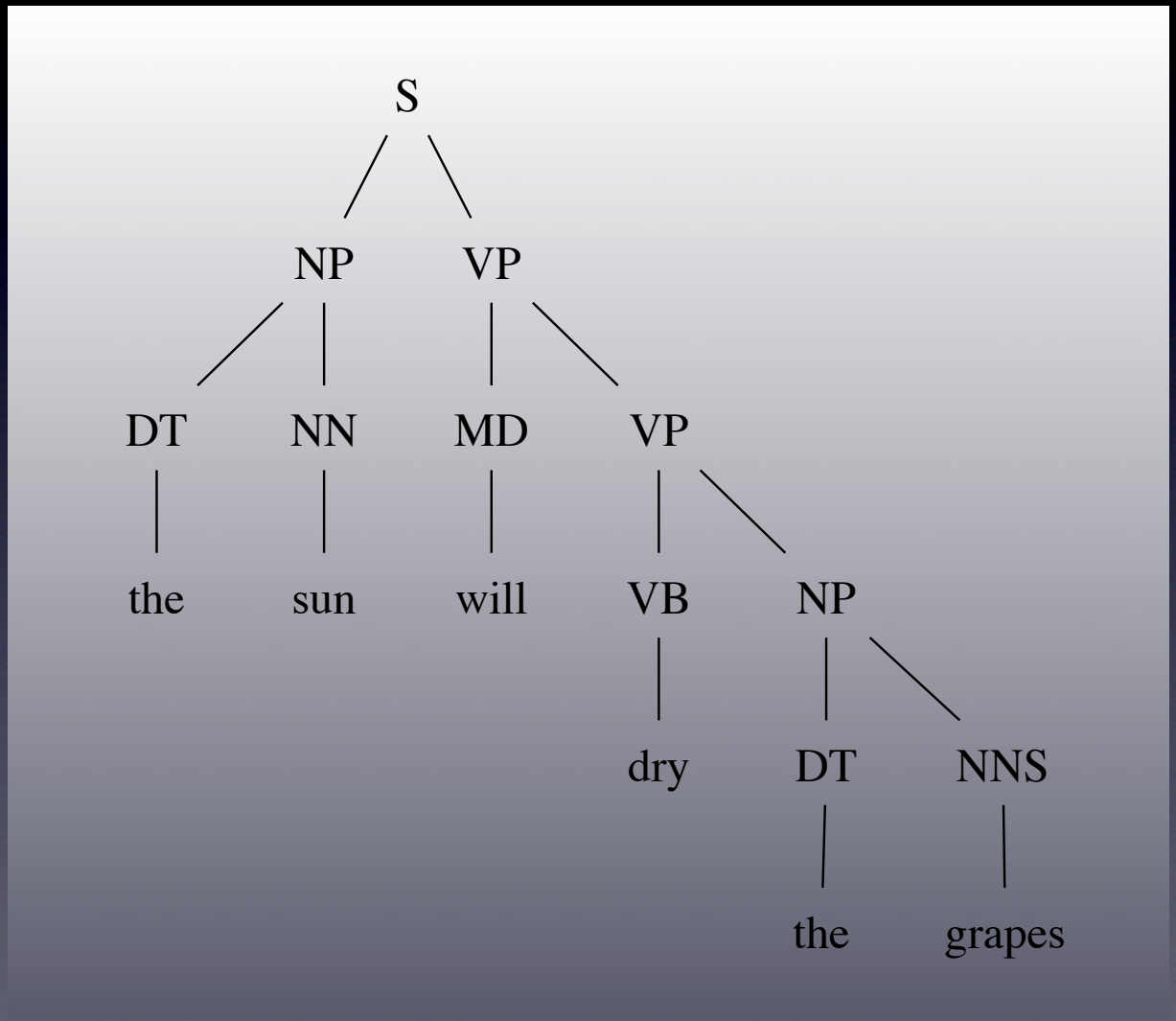
1. Use the Stanford Parser to process native and nonnative text.
2. Convert output of parser into a form that can be used as input for a classifier.
3. Use a large number of texts to train classifiers.
4. Measure accuracy of classifiers and identify the linguistic reasons behind their success.

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

Parse Trees

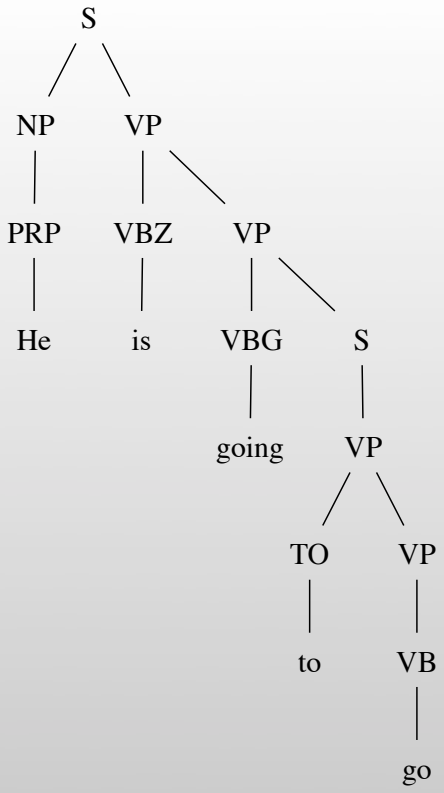


Verb Form Experiment

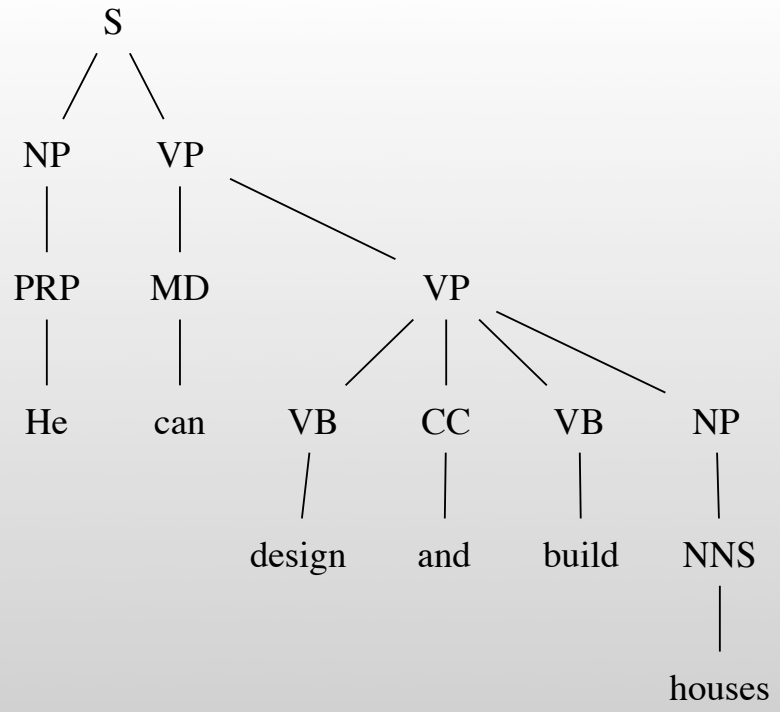
- Attempt to classify text based on verb usage.
- Feature sets based on the following features:
  - Tense, aspect, voice, core modals, phrasal modals, the helping verbs *be, get, have, do*, and the negative particle *not*

Verb Form Experiment

- Determine verbal forms from parse trees.

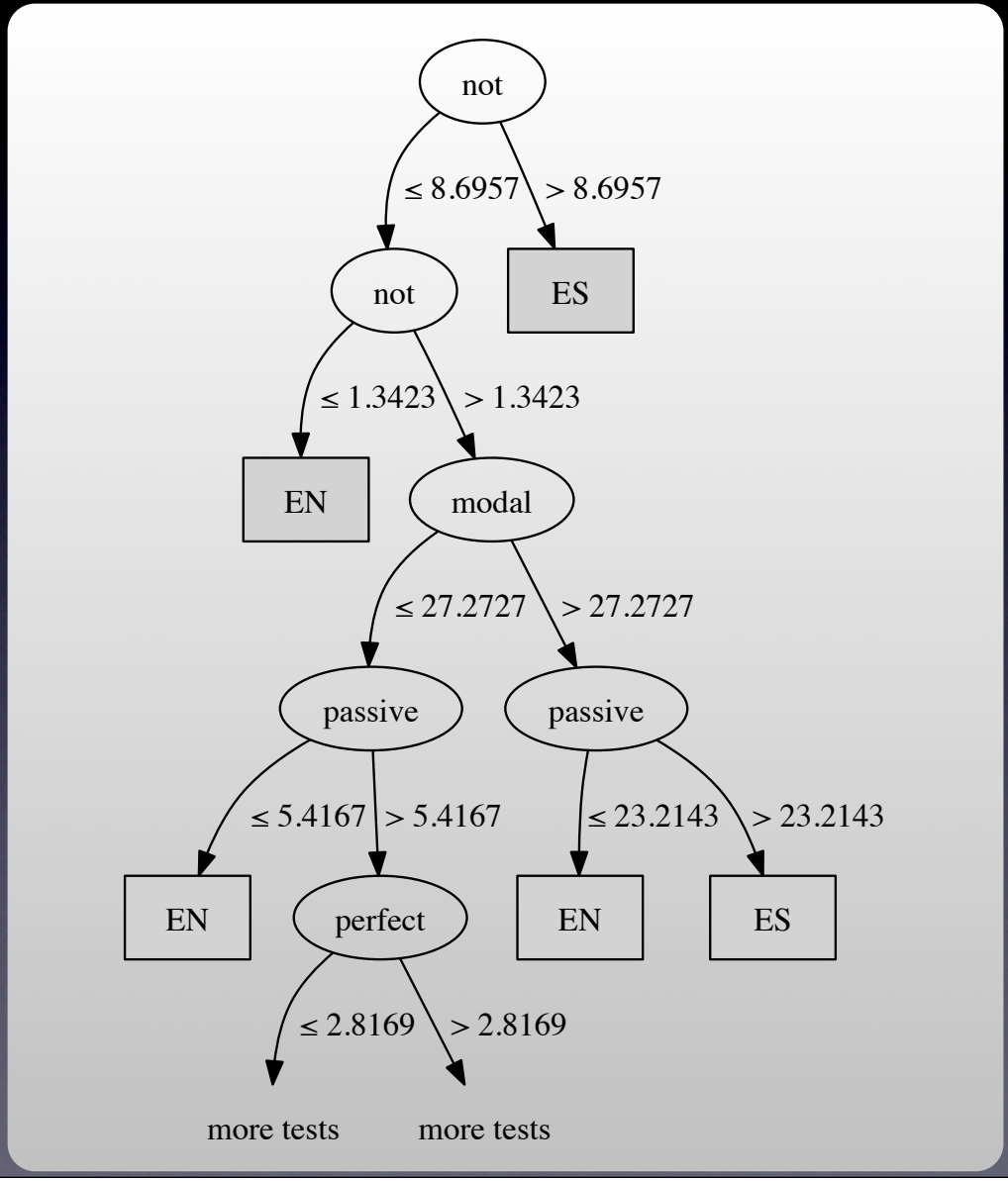


Phrasal Modal



Core Modal with Two Main Verbs

Simplified Decision Tree



Verb Form Experiment

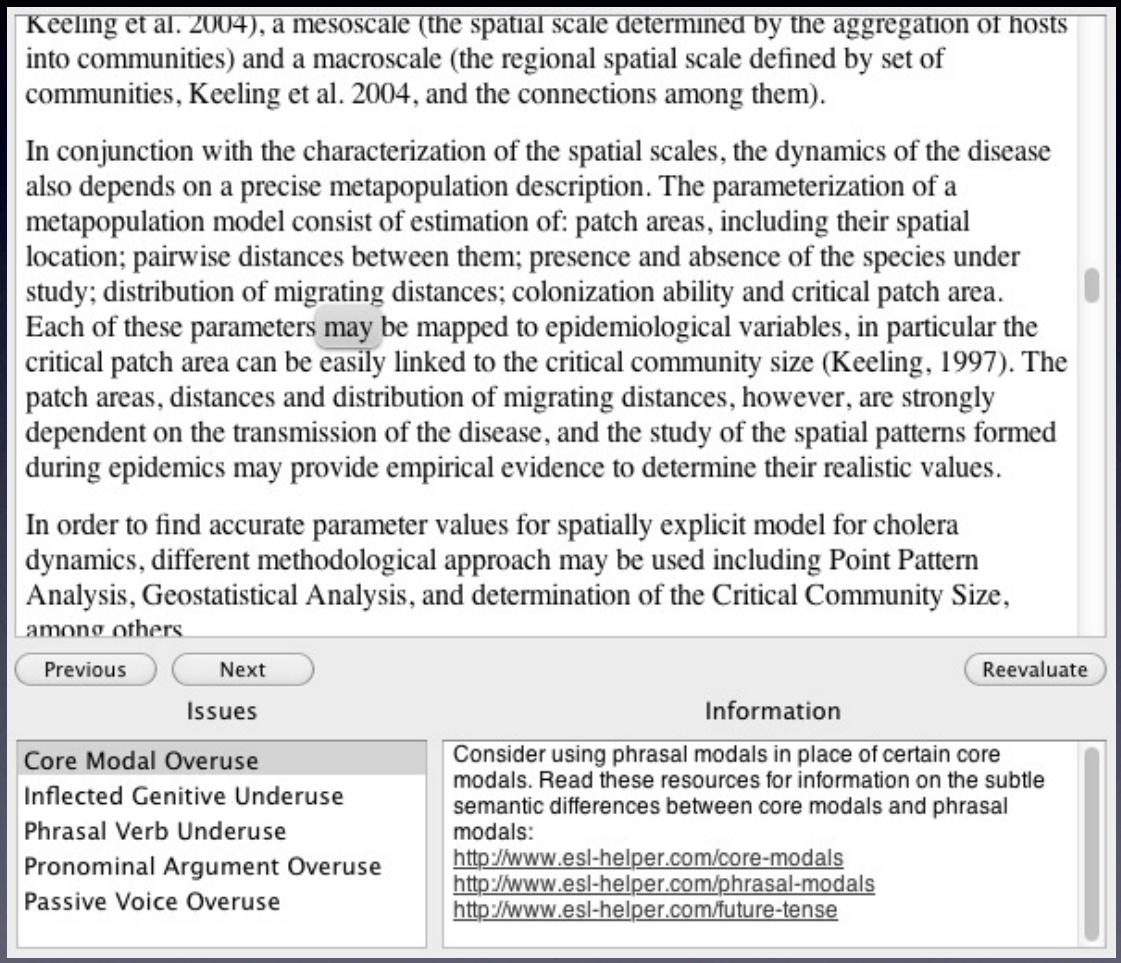
Results

- Overall accuracy between 85% and 90%.
- Errors tend to come from classifying native as nonnative.

Learner's 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?).