Listening. Learning. Leading.®

Measuring Feature Diversity in Native Language Identification

Shervin Malmasi
Macquarie University
Australia

Aoife Cahill
Educational Testing Service
USA

ML for NLI

- Predicting the native language of a writer based on a piece of English writing
- Typically solved using supervised-ML: multi-class classification
- Previous Work has investigated the predictive power of individual feature classes
- No systematic analysis of feature interaction

Beyond NLI System Performance

- Context: language teaching and learning
- Goal: identify L1-specific usage patterns and errors
- Improve teaching methods, instructions and learner feedback
- Previous work shows that the features capture different pieces of information
- How diverse are the features? How can we measure the diversity?

Feature Types for NLI

Lexical

- character n-grams
- word n-grams
- lemma n-grams
- function words

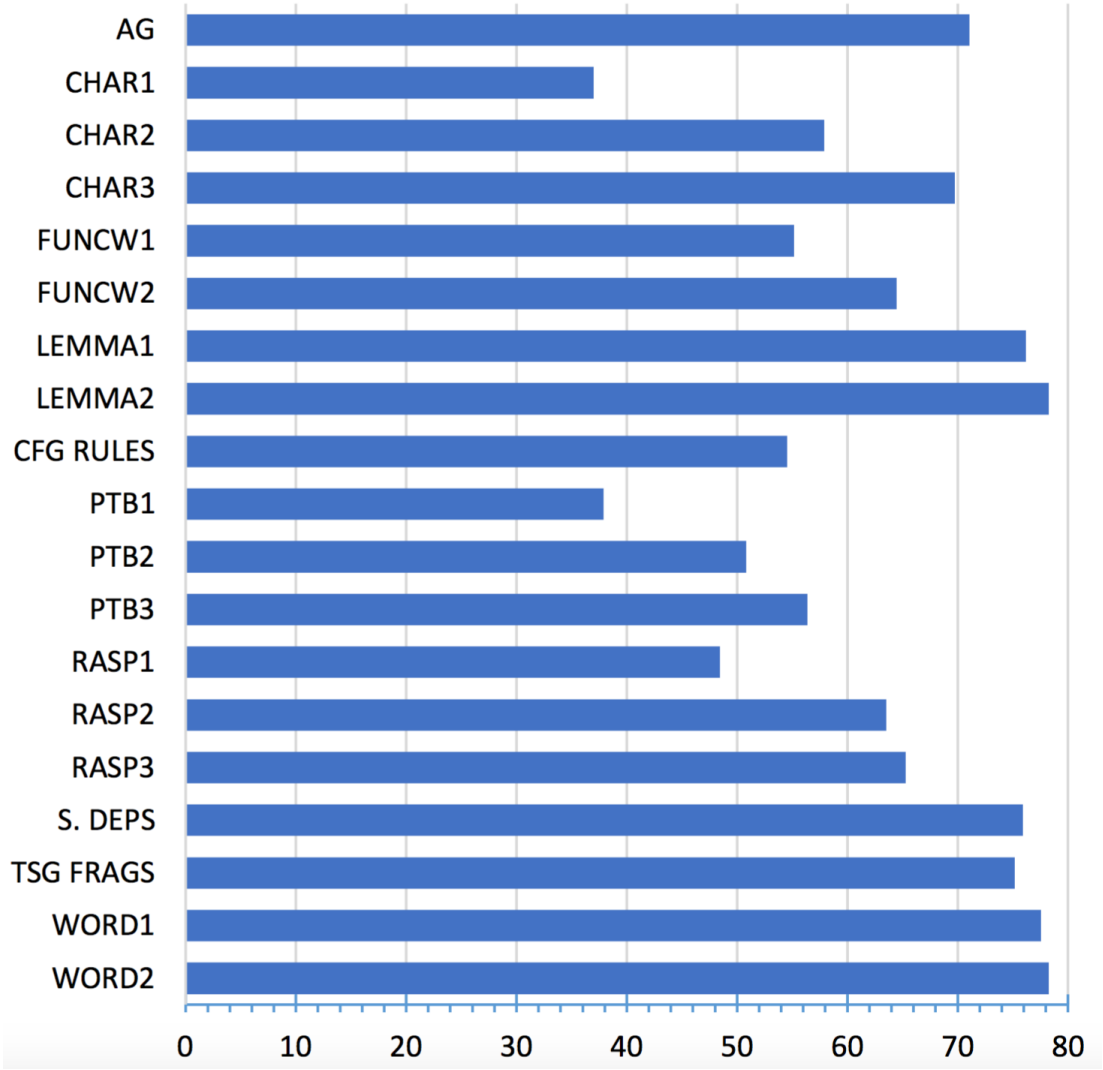
Syntactic

- POS n-grams
- syntactic dependencies
- TSG fragments
- CFG rules
- Adaptor grammars

Data

- ETS Corpus of Non-Native English Writing (TOEFL 11)
- 11 L1s: Arabic, Chinese, French, German, Hindi, Italian, Japanese, Korean, Spanish, Telugu, Turkish
- 1100 essays per L1, 900 train, 100 dev/test
- 8 prompts
- Train on train+dev, Evaluate on test

Accuracy of Individual Features



Measuring Feature Diversity

- Measure agreement between each pair of features for predicting labels on the same dataset
- Idea: the higher the agreement, the lower the diversity of those two features
- Yule's Q-coefficient statistic

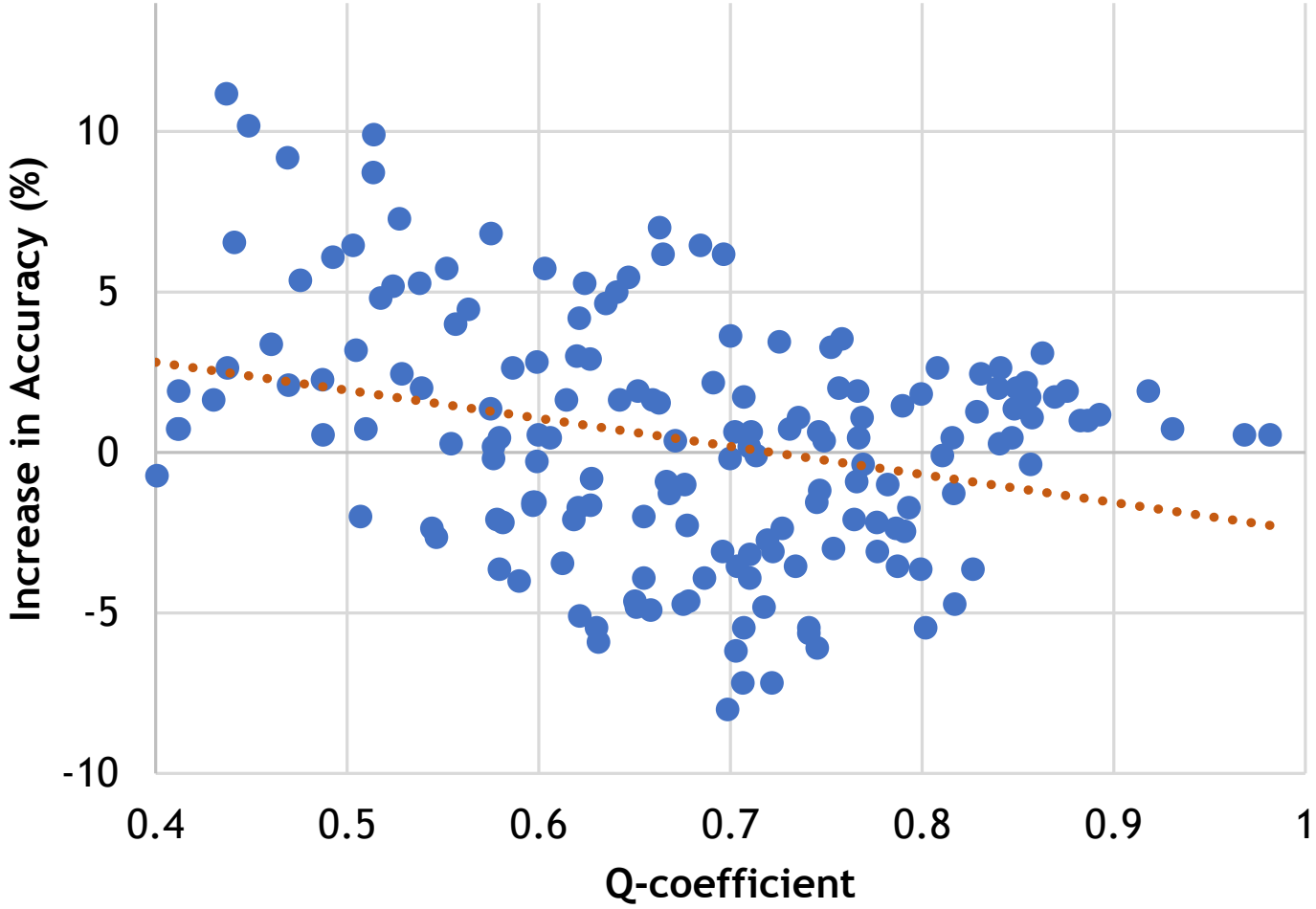
Yule's Q-coefficient

- Correlation coefficient for binary measurements
- Range from -1 to +1

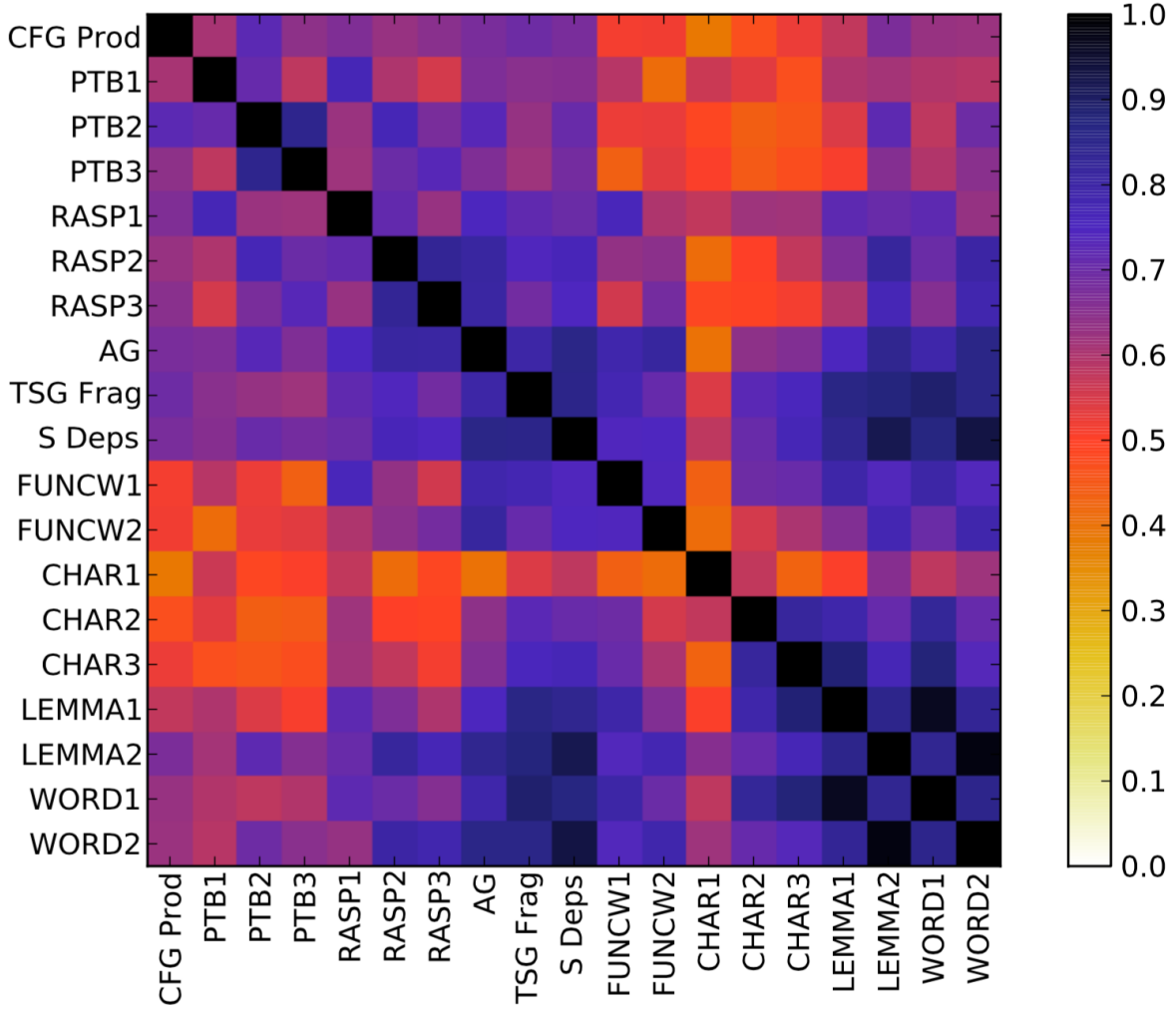
	C_k Correct	C_k Incorrect
C_j Correct	N^{11}	N^{10}
C_j Incorrect	N^{01}	N^{00}

$$Q_{j,k} = \frac{N^{11}N^{00} - N^{01}N^{10}}{N^{11}N^{00} + N^{01}N^{10}}$$

Q-coefficients (171 pairs)

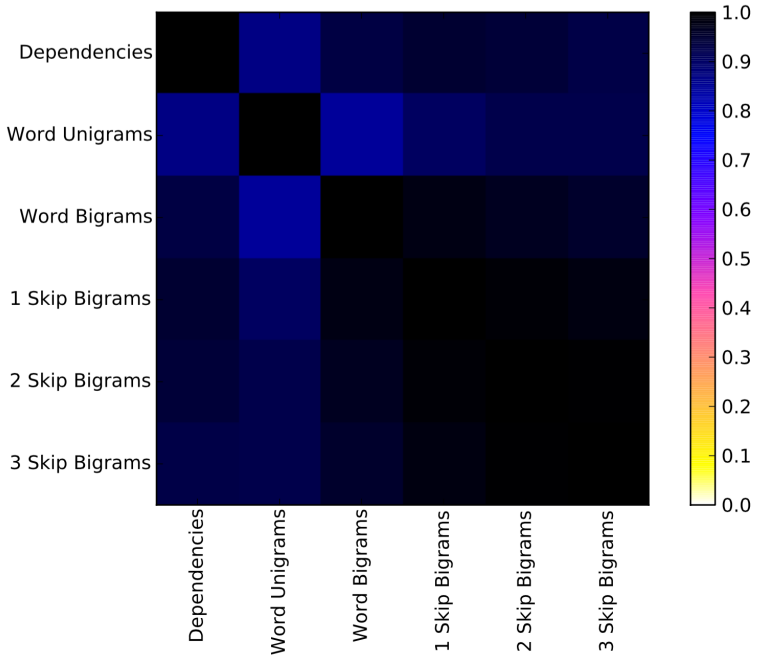


Q-coefficient Matrix

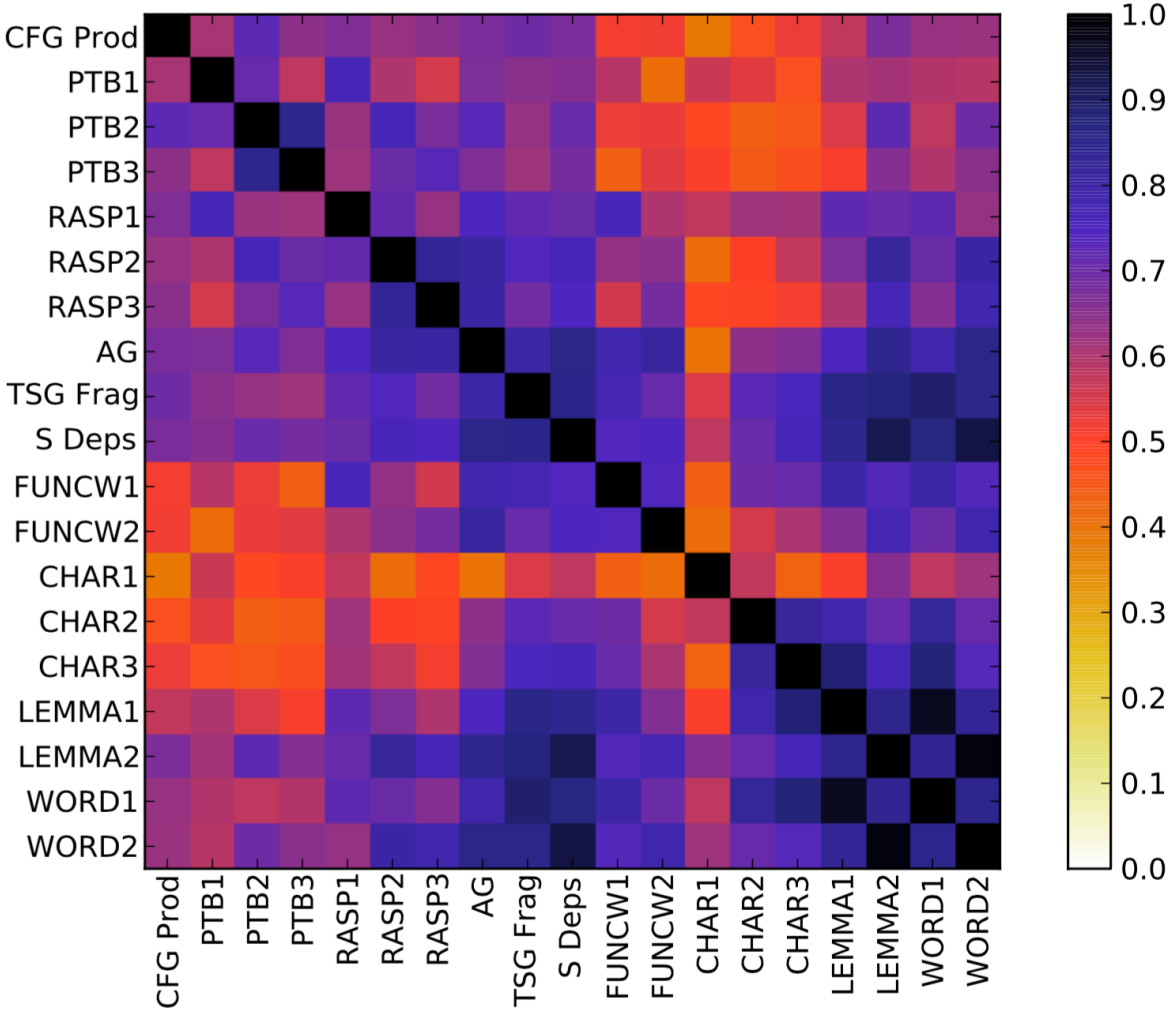


Words and Dependencies

- Naively not thought to be strongly related
- Liu (2008) reports 51% of deps are adjacent
- How does this relate to k-skip word bigrams?



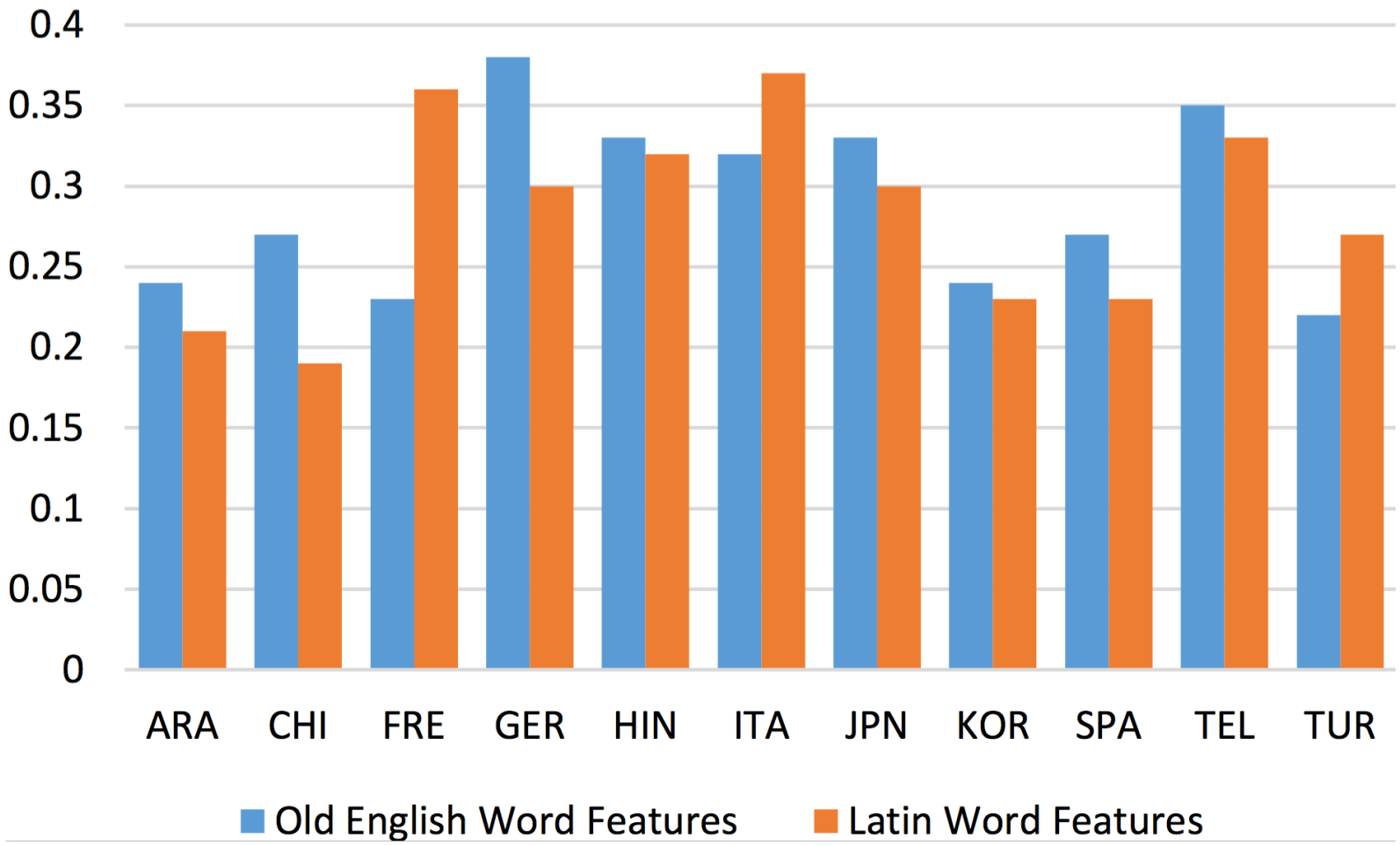
Q-coefficient Matrix



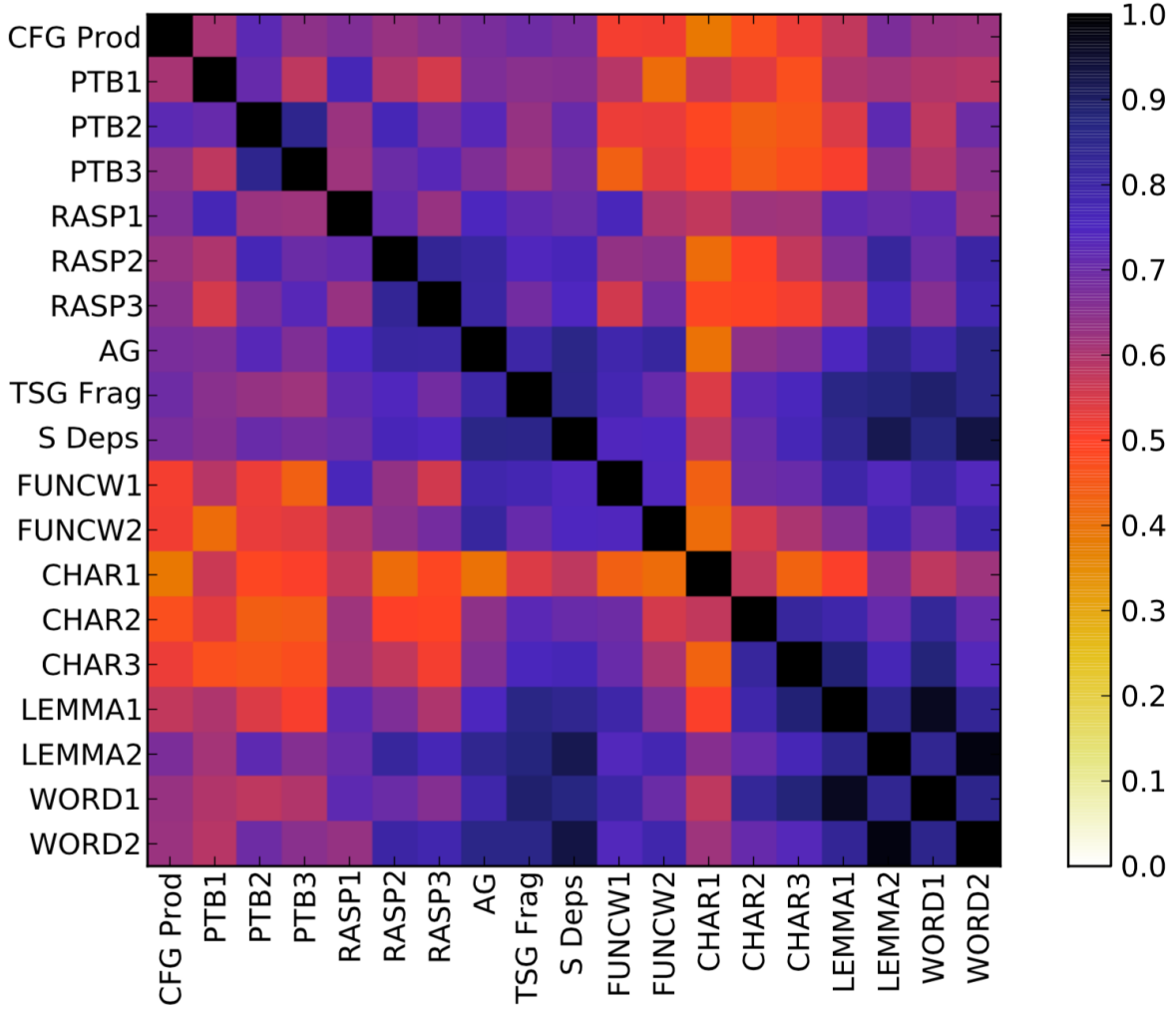
L1 and Word Usage

- Hypothesis: learners tend to use words similar in form and meaning to words in their L1
- Test: Extract English words from Etymological WordNet
  - Germanic roots
  - Latin roots
- Train 2 classifiers with just word unigrams
  - 2 SVMs each trained on different features

L1 and Word Usage Results

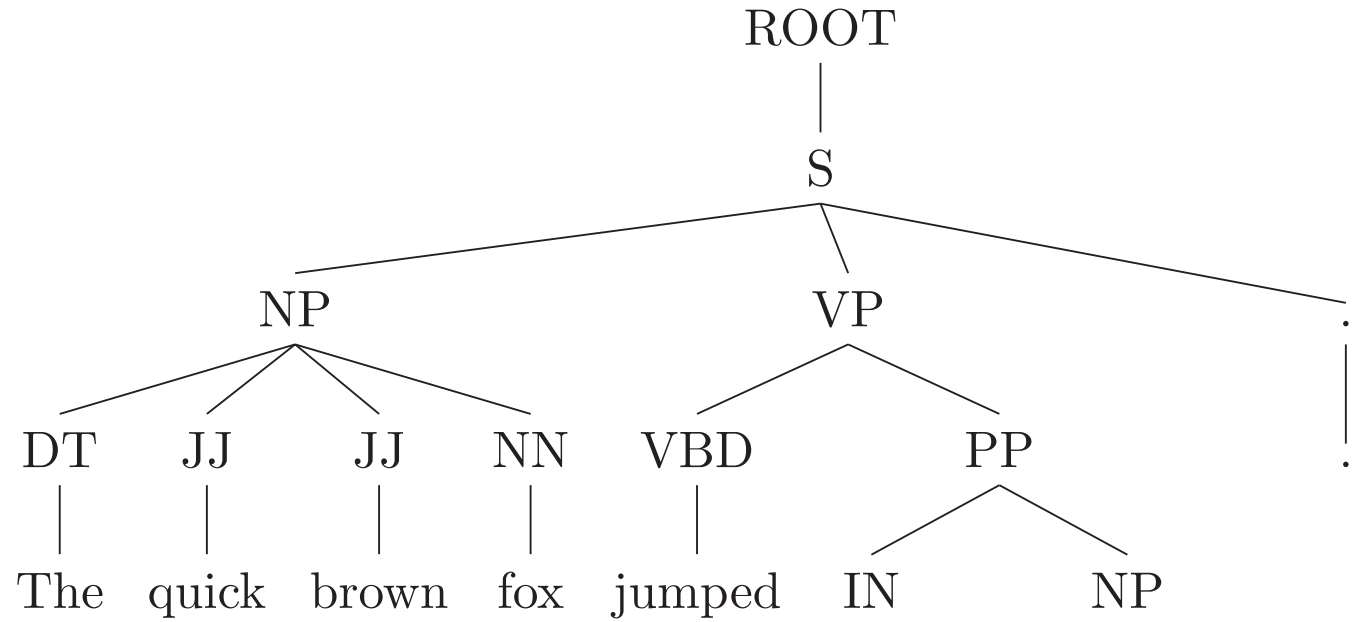


Q-coefficient Matrix



Extending CFG Rules

- Parent Annotations (Johnson, 1998)



Production Rules Extracted from Tree:

ROOT → S	S → NP VP .
NP → DT JJ JJ NN	PP → IN NP
VP → VBD PP	NP → DT JJ NN

Parent-Annotated CFG Rules

```
ROOT          → S^<ROOT>
S^<ROOT>      → NP^<S> VP^<S> .
NP^<S>        → DT JJ JJ NN
VP^<S>        → VBD PP^<VP>
PP^<VP>       → IN NP^<PP>
NP^<PP>       → DT JJ NN
```

Building an NLI system with these features yields accuracy of 55.6%, a +1.3% increase over the standard CFG rules feature.

Conclusions

- Q-coefficient provides a method for measuring feature diversity for high-dimensional feature spaces
- Experiments with NLI on TOEFL data show interesting feature correlations
- Analysis of feature diversity can help suggest new features