

Grammatical Error Detection

Øistein E. Andersen



UNIVERSITY OF
CAMBRIDGE
Computer Laboratory

1. Problem

Identify not only spelling mistakes, but also erroneous combinations of individually correct words.

The Cambridge Learner Corpus contains over 11 million words of manually corrected text, which can be used both to learn the difference between correct and incorrect constructions and to verify that a system actually catches common mistakes.

Potentially useful as a tool for writers, in exam evaluation and for linguists who want to analyse genuine errors.

2. Tagging and parsing

Parsing with RASP provides further information about the words and the relations between them.

Fig. 1 shows the parse tree for the sentence 'Then a thought occurred to me.' Each word is assigned a part-of-speech tag indicating the word class to which it belongs, and the branches makes the sentence structure explicit.

Part-of-speech tags may provide useful generalisations, and the parse tree gives access to grammatical closeness as opposed to mere juxtaposition of words.

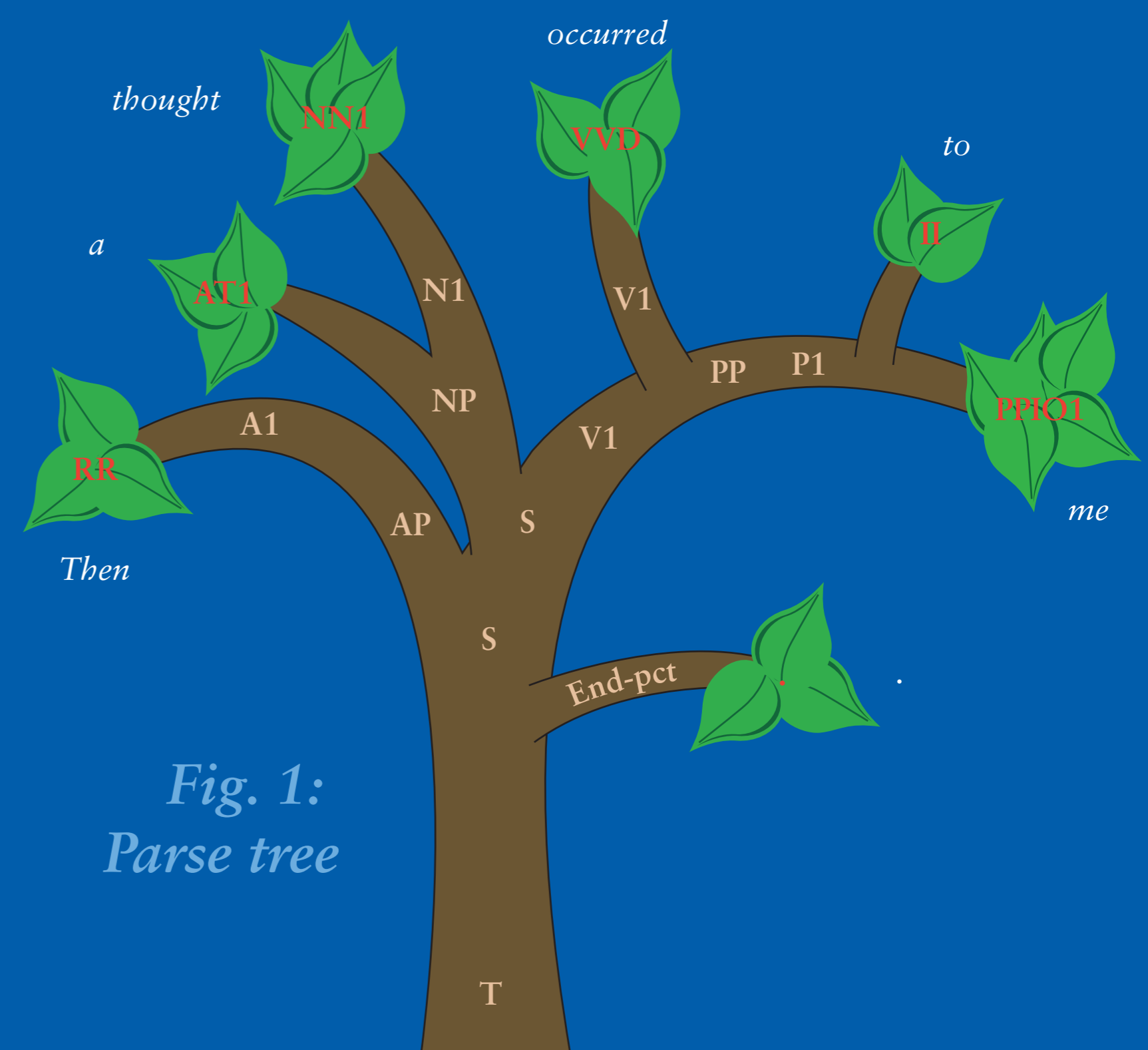


Fig. 1:
Parse tree

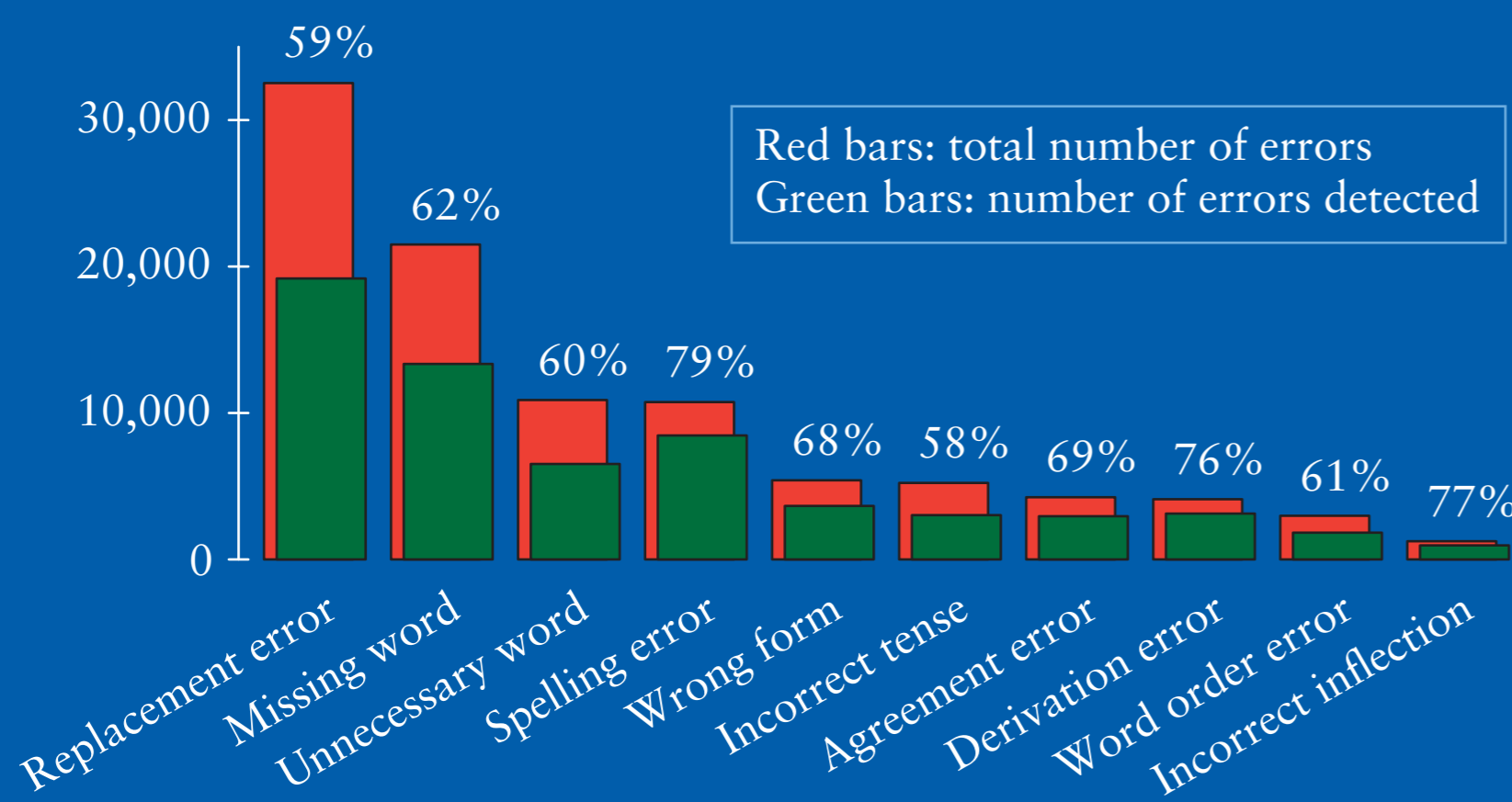


Fig. 2: Results per error type

3. Baseline system

Correct and incorrect sentences extracted from the CLC can be used to train a binary classifier to identify incorrect sentences.

Using mainly word tokens and part-of-speech tags, on their own as well as in combination with adjacent and grammatically related ones, to train a naïve Bayesian classifier, we obtained an overall accuracy of ca. 70%.

More details can be found in Fig. 2 and 3.

4. Model inadequacy

The performance reported in Fig. 3 suggests that sentences with few errors are unlikely to be detected due to overwhelming counterevidence.

Simple machine-learning experiments involving the form of the indefinite article ('a'/'an') shows that this is a real problem: Sentence-level classification gives only 55% recall with 80% precision, whereas a word-level approach gives >95% recall with >90% precision.

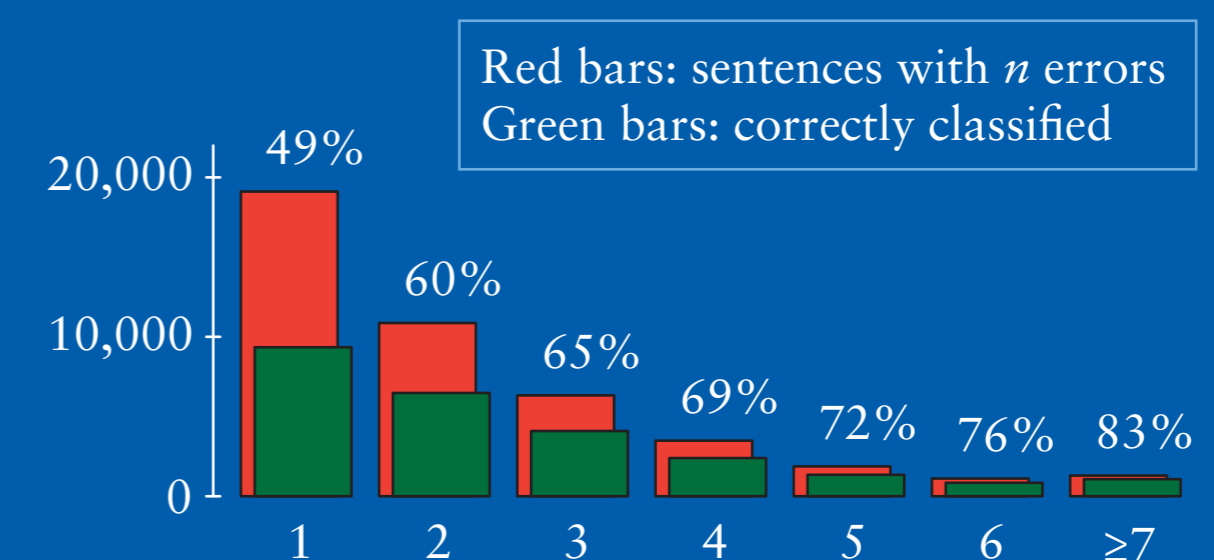


Fig. 3: Results per number of errors

5. Parser adaptations

RASP performs better on correct English; e.g., its grammatical rules does not allow a singular determiner like 'this' to be used in conjunction with a plural noun, which may lead to a parse tree that is difficult to interpret.

Simple adaptations to allow ungrammatical constructions to be parsed may be useful to detect them efficiently. This approach notably allows 92% of the incorrect occurrences of 'this' instead of 'these' to be identified.

6. Further work

Further work includes developing specialised classifiers for more complex errors and finding a way of combining the evidence from each.

We would like to thank Cambridge Assessment and Cambridge University Press for having granted us access to the Cambridge Learner Corpus as part of the English Profile Project. This poster reports on research supported by the University of Cambridge ESOL Examinations.