Some “easy” measures include:

- **accuracy**: how many sentences are correctly parsed?
- **coverage**: how many sentences can be parsed by the grammar?
- **n-best accuracy**: for how many sentences is the correct parse among the n best parses?

PARSEVAL Measures

- **PARSEVAL**: Workshop in 1991 to decide how to evaluate parsers
- more “exact” measures, looking at single constituents
  - correct constituent: has the correct yield
  - gold standard: correctly annotated data

PARSEVAL Measures – Example

Precision: 3/6 = 0.5
Recall: 3/7 = 0.42
F-score: $F_{\beta} = \frac{(\beta^2 + 1) \times \text{precision} \times \text{recall}}{\beta^2 \times \text{precision} + \text{recall}}$

F-score: $F_{0.5} = \frac{(0.5^2 + 1) \times 0.5 \times 0.42}{0.5^2 \times 0.5 + 0.42} = 0.457$
Other metrics

  - Obtain each word's path to the root S
  - Compare to the gold standard, using edit distance
- Dependency evaluation (Lin 1995 and others):
  - Convert phrases to dependencies and evaluate percentage of correct dependencies
  - Does a better job of evaluating on “important” attachments and not penalizing parsers for propagating errors