Parser Evaluation

“easy” measures:

- **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 (2)**

- **precision**: number of correct constituents (yield) in parser output divided by number of constituents in the parser output
- **recall**: number of constituents from the gold standard (yield) that can be found in the parser output divided by the number of constituents in the gold standard

- **labeled precision**: percentage of correct constituents (yield + label) in parser output
**PARSEVAL Measures (2)**

- **precision**: number of correct constituents (yield) in parser output divided by number of constituents in the parser output.
- **recall**: number of constituents from the gold standard (yield) that can be found in the parser output divided by the number of constituents in the gold standard.
- **labeled precision**: percentage of correct constituents (yield + label) in parser output.
- **labeled recall**: percentage of constituents from the gold standard (yield + label) that can be found in the parser output.

**F-score**:

\[ F_\beta = \frac{(\beta^2 + 1) \times \text{precision} \times \text{recall}}{\beta^2 \times \text{precision} + \text{recall}} \]

**PARSEVAL Measures – Example**

Gold: Parse:  

\[
\begin{align*}
\text{gold:} & \quad PX \quad APPR \quad NX \quad ART \quad einem \quad NN \quad oder \quad NN \\
& \quad SAMstag \quad Sonntag \\
\text{parse:} & \quad PX \quad APPR \quad NX \quad ART \quad einem \quad NN \quad oder \quad NN \\
& \quad vielseit \quad ADV \quad vielseit \\
& \quad PX \quad APPR \quad NX \quad ART \quad einem \quad NN \quad oder \quad NN \\
& \quad SAMstag \quad Sonntag
\end{align*}
\]

Precision: \( \frac{3}{6} = 0.5 \)
Recall: \( \frac{3}{7} = 0.42 \)
F-score: \( \frac{2 \times 0.5 \times 0.42}{0.5 + 0.42} = 0.457 \)