Annotation Quality Checking and Its Implications for Design of a Treebank

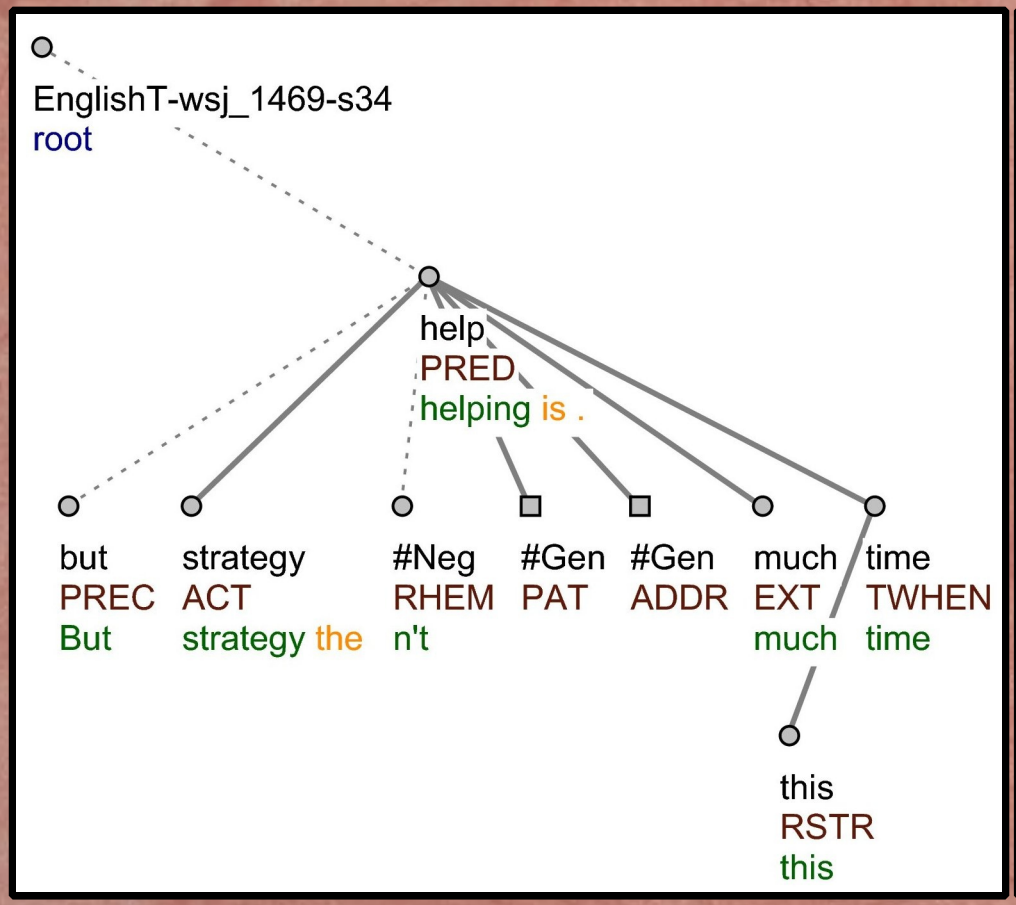
(in Building the Prague Czech-English Dependency Treebank)

Marie Mikulová and Jan Štěpánek
Charles University in Prague
ÚFAL

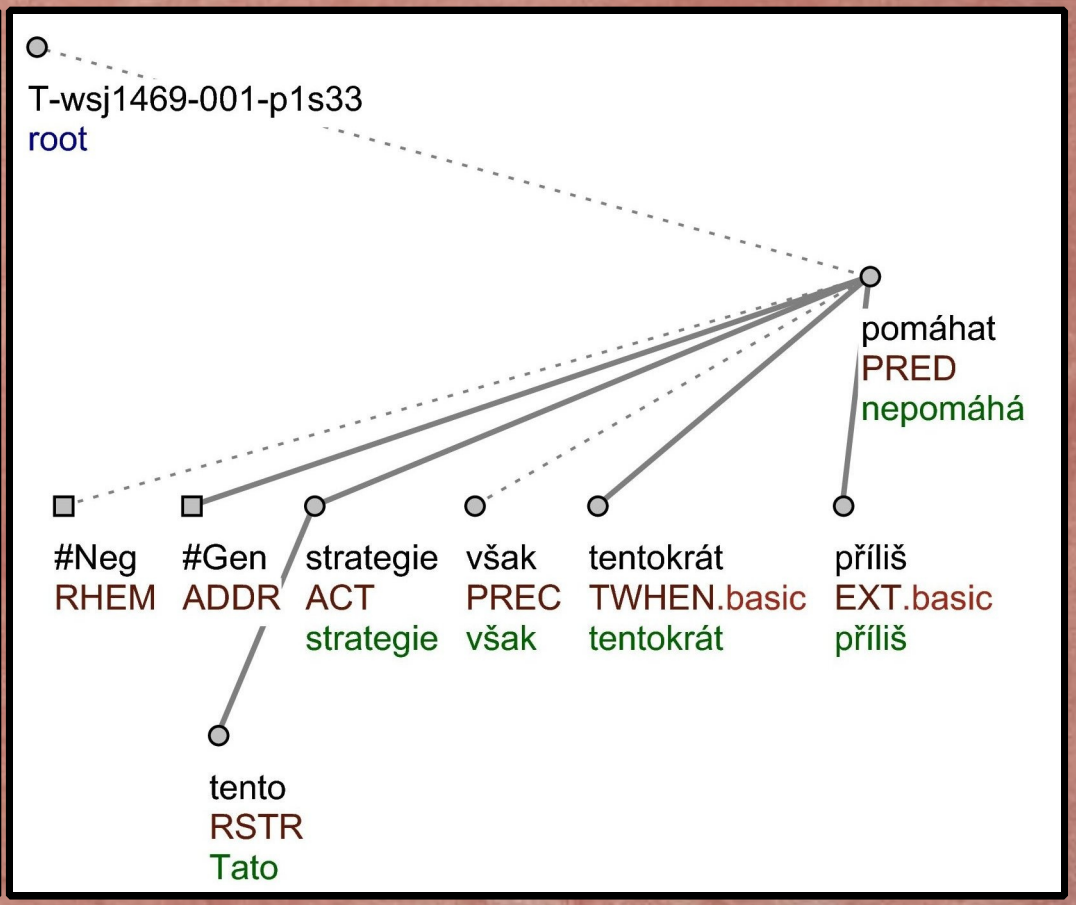
Prague Czech-English Dependency Treebank

- Deep syntactic (tectogrammatical) parallel treebank
- Similar to Prague Dependency Treebank 2.0
  - Stand-off annotation
  - 4 layers (word-form, morphological, analytical, tectogrammatical) – differences
- Wall Street Journal part of the Penn Treebank (49,000 sentences)

PCEDT – Example



But the strategy isn't helping much this time.



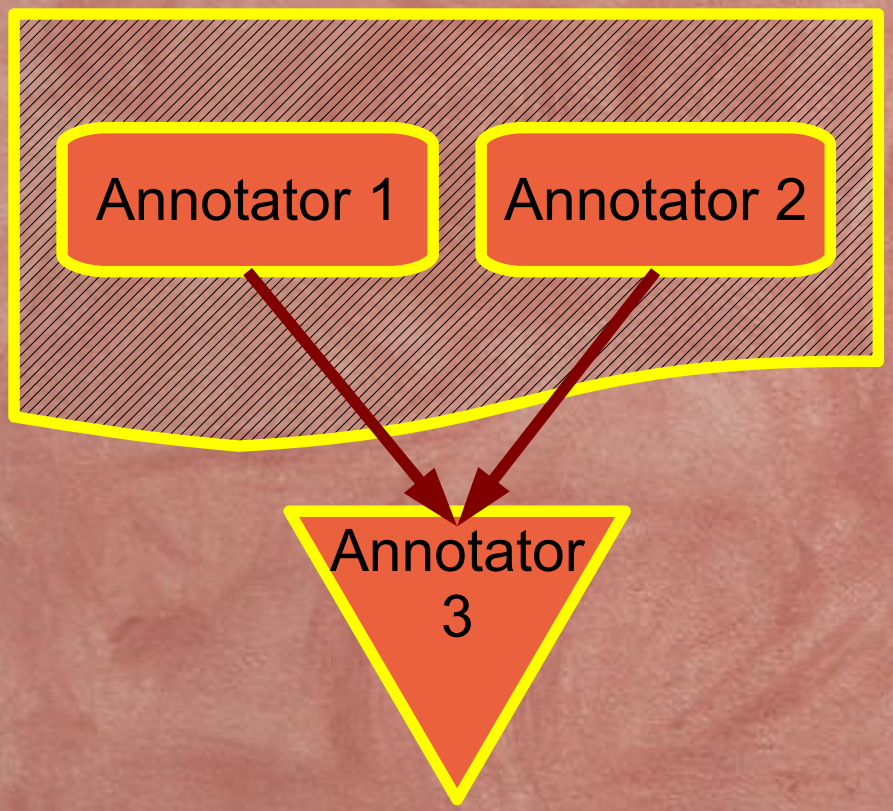
Tato strategie však tentokrát příliš nepomáhá.

Annotation Procedure

- Tectogrammatical layer only
  - 39 attributes (8.42 per node in PDT 2.0)
  - pre-built tree as an input
- Division into several phases
- Periodic measurement of inter-annotator agreement
- Periodic checking of correctness of the annotation

Annotation Quality Checking

Usual approach:



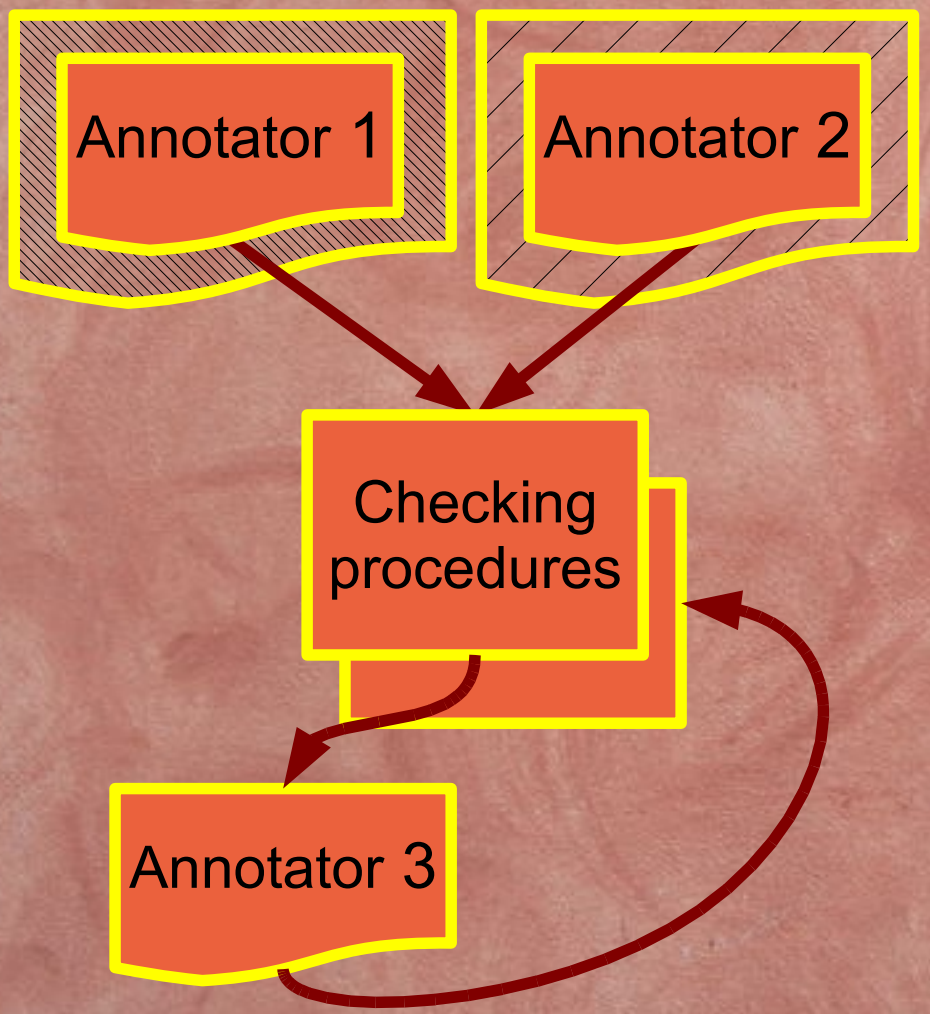
9.2 sentences per hour

5 years at a half-time job
€: 3 x 5 = 15

Too slow and too expensive :-(

Annotation Quality Checking (2)

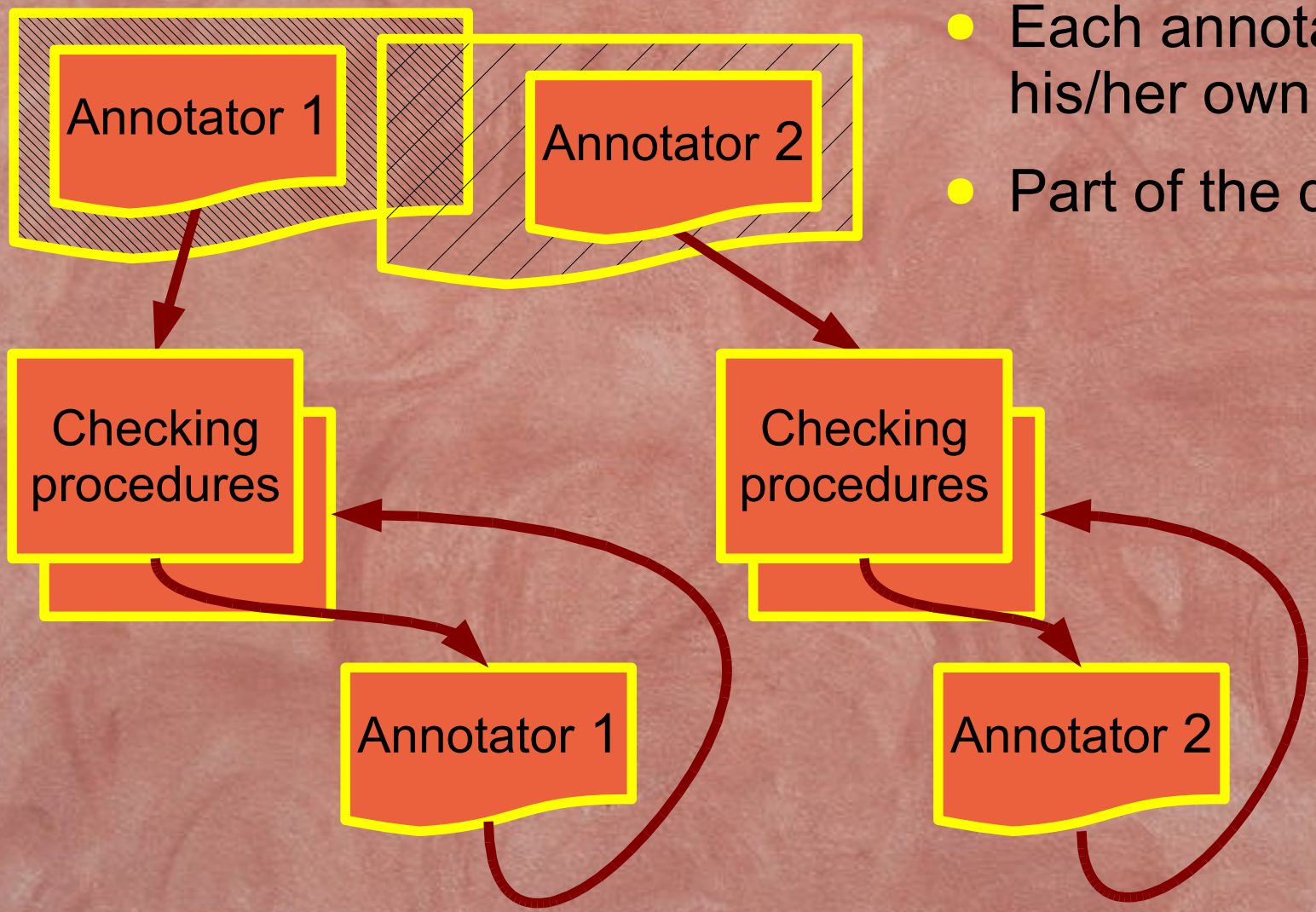
PDT 2.0 approach:



- Checking of finished data.
- No parallel data at all.

Annotation Quality Checking (3)

PCEDT approach:



- Each annotator checks his/her own data.
- Part of the data parallel.

Checking Procedures

- Invariants, impossible or necessary combinations of the nodes and their attributes
- Source:
  - annotation rules
  - annotators' feedback
  - generalization of the output of an automatic checking procedure: searching for the same surface coverage with different annotation

Checking Procedures (2)

- Implemented in TrEd (based on Perl)
- Output table columns:
  - procedure name
  - type of violation
  - last column: position
- Only accurate procedures (exceptions)
- 50 procedures, 103 possible violations
- 5 categories

Checking Procedures – Attribute

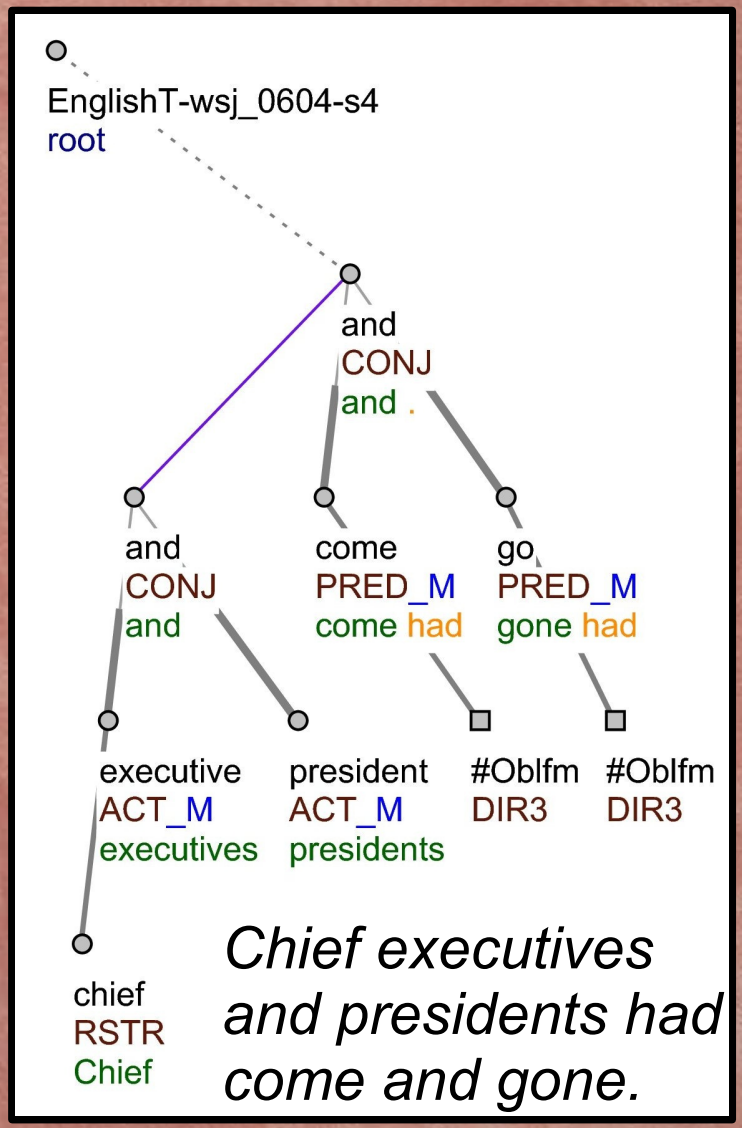
- Only a single attribute is tested, the structure is ignored.
  - Currently, only t_lemma (no other non-structural attribute being annotated)
- Example:
  - Reasons are given for every change in pre-generated tectogrammatical lemma.

Checking Procedures – Structure

- Relation between the governing and dependent node and their attributes
- Examples:
  - The root's functor must be PRED, DENOM, PARTL, or VOCAT.
  - PRED and DENOM are possible only for a root.
  - The adnominal attribute (RSTR) can never depend on a verb.
  - Every negated verb has a *#Neg* child.
  - *#EmpVerb* and *#EmpNoun* are never leaves.

Checking Procedures – Coordination

- “Effective” dependencies
- Examples:
  - Every coordination has at least two members.
  - Some functors cannot be coordinated together (inner participant (argument) only with an argument of the same sort).



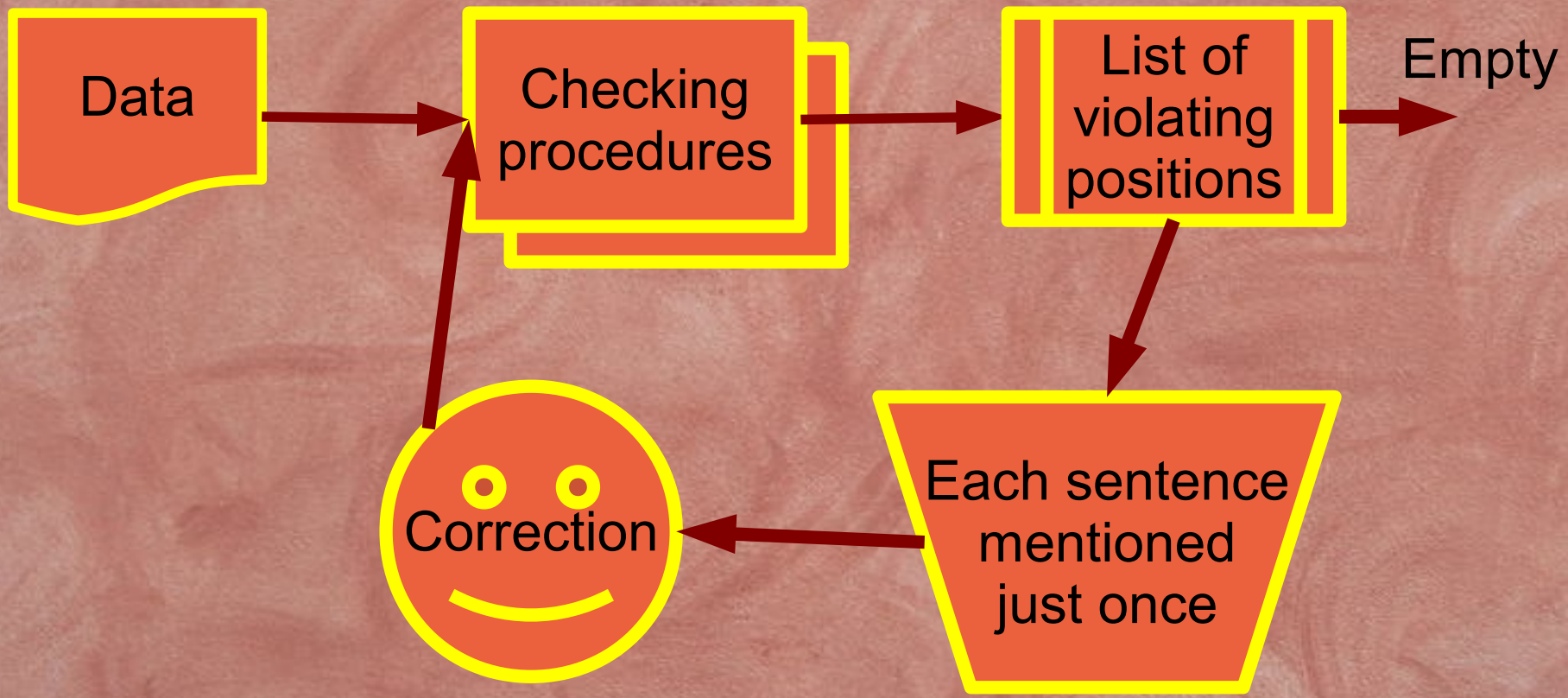
Checking Procedures – Links

- Links from the t-layer to the a-layer
- Examples:
  - For every a-node representing a word (i.e. not punctuation) there must be a link from a t-tree.
  - The same a-node can be linked as auxiliary to several t-nodes only if the t-nodes are coordinated, or they or their parents have the same t-lemma, or...
  - No links to prepositions from DENOM and VOCAT.

Checking Procedures – Valency

- Each verb and deverbative noun is assigned a valency frame.
- Obligatory modifications omitted on the surface must be added to the t-tree.
- Examples:
  - Valency frame is assigned where required.
  - No obligatory modification is missing, no actant is superfluous.
  - “Copied” node has the same valency frame as its original.

Correction Workflow



Impact on the Treebank Design

- Checking procedures
  - Find errors
  - Reveal vague annotation rules
  - Appreciation of the annotators

Evaluation of Annotators

- Average error rate per sentence for each annotator
- Ranks remain the same in long-term monitoring

Annotator	Errors / Sentences	Errors per Sentence
ma	3 271 / 6 026	0.54
al	1 214 / 3 213	0.38
iv	2 648 / 8 125	0.33
ji	301 / 1 064	0.28
mi	430 / 1 786	0.24
ka	1 834 / **8 132**	0.23
le	373 / 1 903	0.20
ol	1 177 / 6 828	0.17
ALL	12 139 / 39 609	0.31
ORIG	119 090 / 34 862	3.42

Refining the Annotation Rules

- Example: “Copied” verb has the same valency frame as its original.
  Peter gave Mary flowers and [he gave] Jane sweets.
  - Metaphoric or phraseological usage:
    For a conflict, he does not have enough attention nor [he has] stomach.
  - One meaning split into several valency frames:
    Company A’s stock closed mixed and company B’s [stock closed] down modestly.

Most Common Errors

Checking Procedure	Occurences	Percentage
valency003_2_PAT_missing	883	7.27
links001_6.1_same_aux	700	5.77
valency003_2_ACT_missing	623	5.13
links001_1.1_no_tnode	438	3.61
valency001_1_no_frame	405	3.34
valency003_4_wrong_aux	387	3.19
structure016_1_no_neg	378	3.11
attribute001_1_t-lemma	352	2.90
structure003_1_fphr_lemma	348	2.87
valency003_1_invalid_lemma	345	2.84

Thank you.