

Generation in MT

Jan Hajič

The Team

– Senior members & affiliate members

- Jan Hajič, Charles Univ., Prague
- Drago Radev, Univ. of Michigan
- Gerald Penn, Univ. of Toronto
- Jason Eisner, Johns Hopkins Univ.
- Owen Rambow, Univ. of Pennsylvania
- Dan Gildea, Univ. of Pennsylvania
- Bonnie Dorr, Univ. of Maryland

– Students:

- Yuan Ding, Univ. of Pennsylvania
- Martin Čmejrek, Charles Univ., Prague
- Terry Koo, MIT
- Kristen Parton, Stanford Univ.

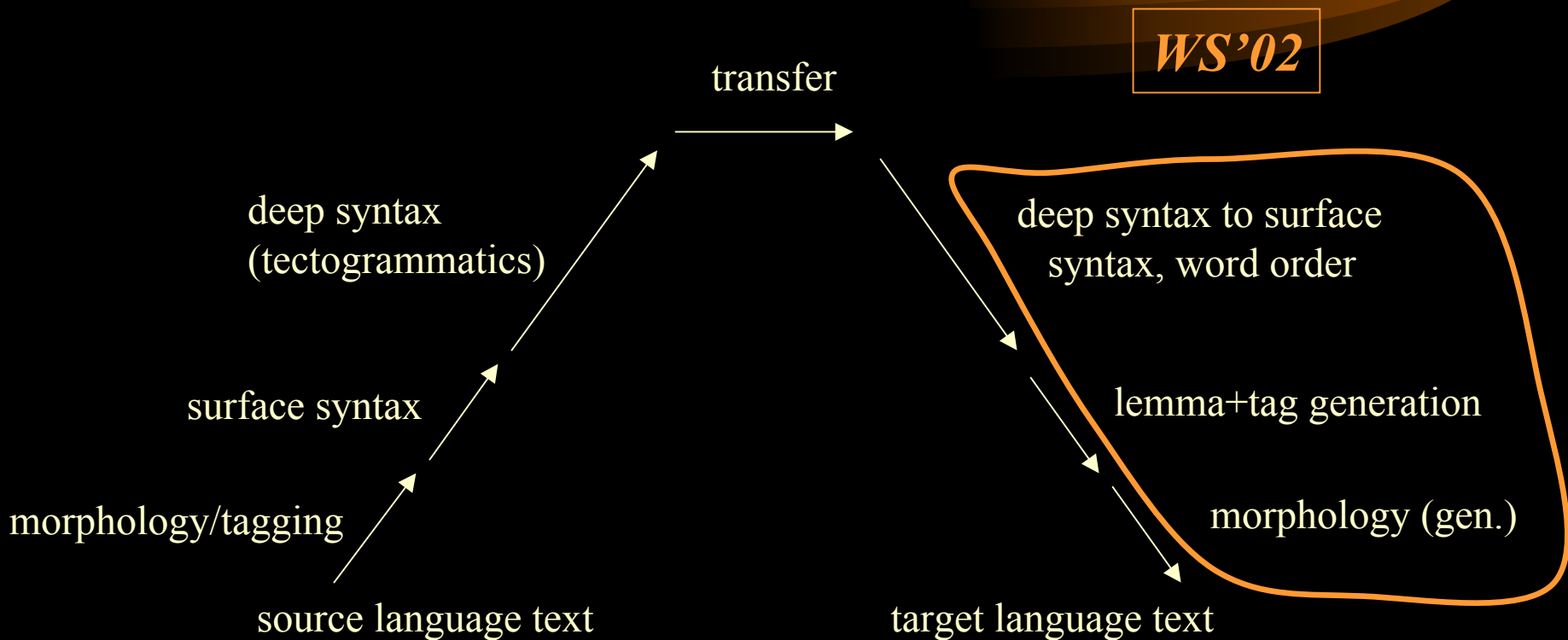
The Goal

- Generate English (plain surface form)
 - from syntactic-semantic sentence representation (so-called “tectogrammatical”, or TR)
- Possible application setting:
 - machine translation
 - other uses:
 - part of front-end for QA systems, full generation
- Evaluate under various circumstances

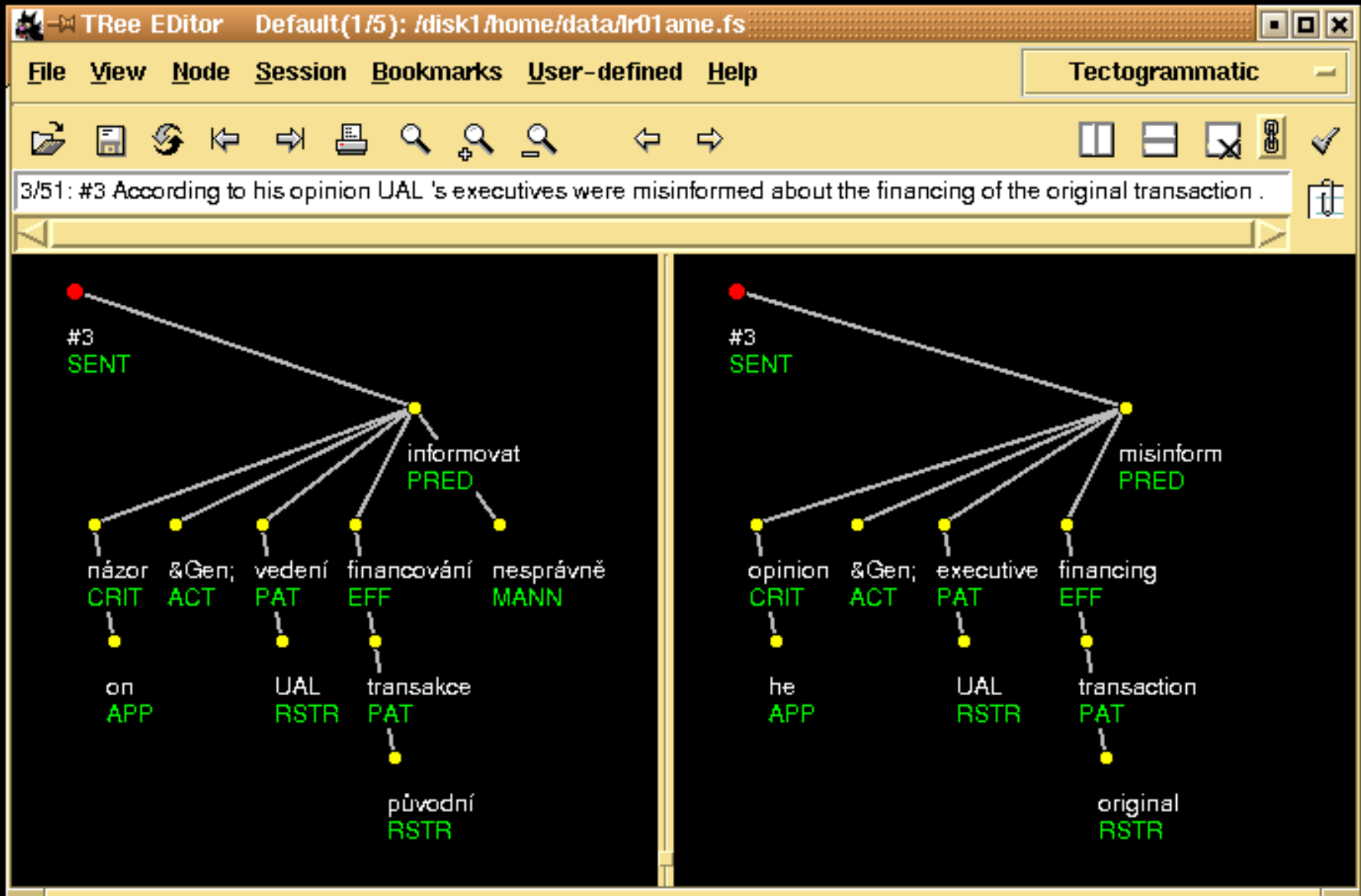
The Framework

- “Classic” MT design assumed
 - Analysis - Transfer - Synthesis
- Tectogrammatical level at transfer stage
 - Dependency syntactic-semantic representation
- Language pair:
 - from Czech to English

The Framework



Tectogrammatical representation



The screenshot shows the TTree Editor interface with the following components:

- Title Bar:** TTree Editor Default(1/5): /disk1/home/data/lr01 ame.fs
- Menu Bar:** File View Node Session Bookmarks User-defined Help
- Toolbar:** Includes icons for file operations (open, save, refresh), navigation (back, forward), search (find, find next, find previous), and window management (split, close, maximize, check).
- Status Bar:** 3/51: #3 According to his opinion UAL 's executives were misinformed about the financing of the original transaction .
- Main Area:** Two side-by-side tectogrammatical trees.

Left Tree (Czech):

- Root: #3 SENT
- Child: informovat PRED
- Children of informovat:
 - názor CRIT → on APP
 - &Gen; ACT
 - vedení PAT → UAL RSTR
 - financování EFF → transakce PAT → původní RSTR
 - nesprávně MANN

Right Tree (English):

- Root: #3 SENT
- Child: misinform PRED
- Children of misinform:
 - opinion CRIT → he APP
 - &Gen; ACT
 - executive PAT → UAL RSTR
 - financing EFF → transaction PAT → original RSTR

For comparison: Surface dependency

TRE Editor Default(1/6): /disk1/home/data/r01amea.fs

File View Node Session Bookmarks User-defined Help Analytic

3/51: #3 According to his opinion UAL 's executives were misinformed about the financing of the original transaction .

The image displays two dependency trees side-by-side, illustrating surface dependencies for the sentence: "#3 According to his opinion UAL 's executives were misinformed about the financing of the original transaction .".

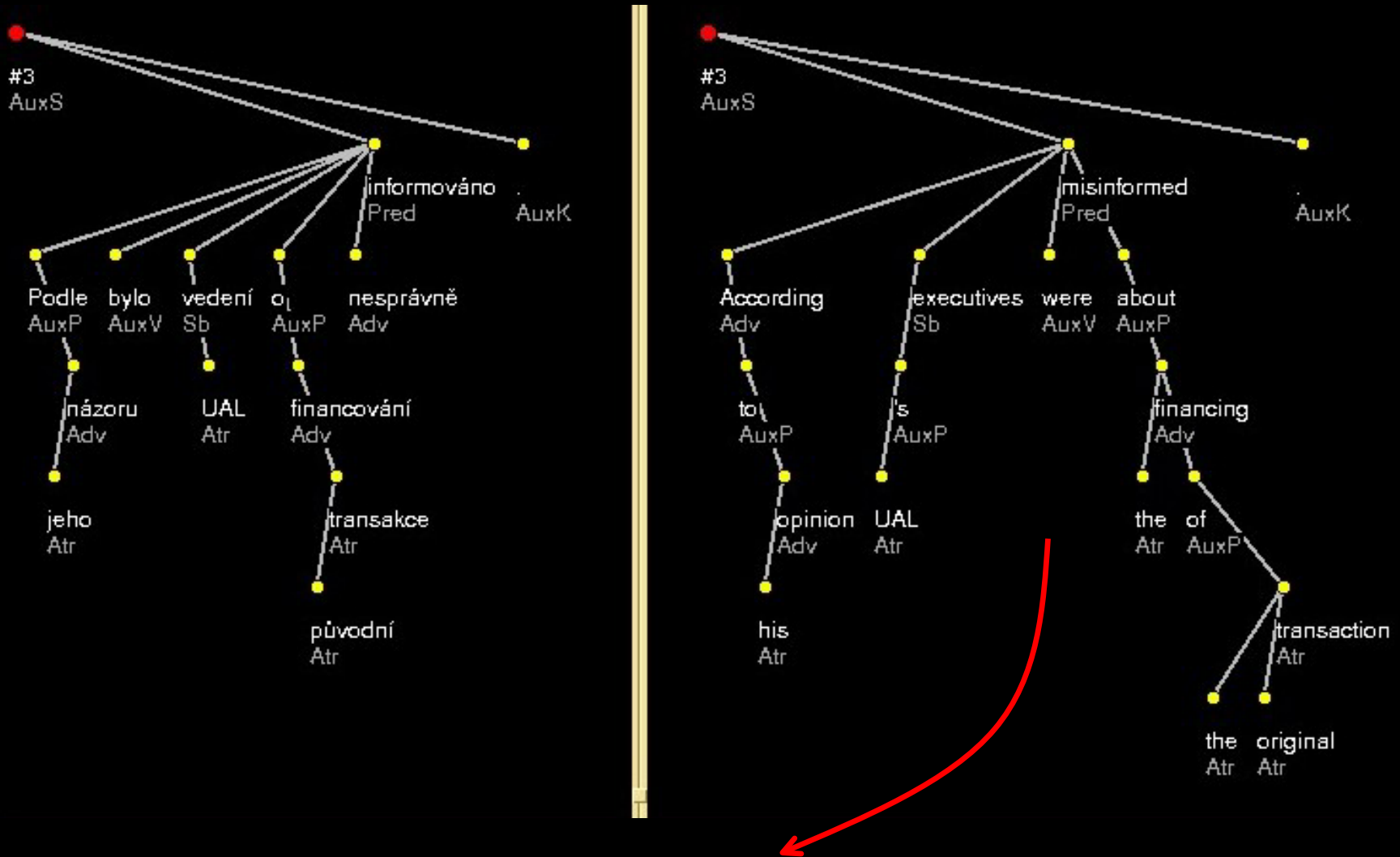
Left Tree (Czech):

- Root: #3 AuxS
- Children: Podlé AuxP, bylo AuxV, vedení Sb, o_i AuxP, nesprávně Adv, informováno Pred, AuxK
- Podlé AuxP → názoru Adv → jeho Atr
- bylo AuxV → UAL Atr
- vedení Sb → UAL Atr
- o_i AuxP → financování Adv → transakce Atr → původní Atr
- nesprávně Adv → transakce Atr
- informováno Pred → AuxK

Right Tree (English):

- Root: #3 AuxS
- Children: According Adv, executives Sb, were AuxV, about AuxP, misinformed Pred, AuxK
- According Adv → to AuxP → opinion Adv → his Atr
- executives Sb → 's AuxP → UAL Atr
- about AuxP → financing Adv → the Atr → of AuxP → the Atr → transaction Atr → original Atr
- misinformed Pred → AuxK

For comparison: Surface dependencies



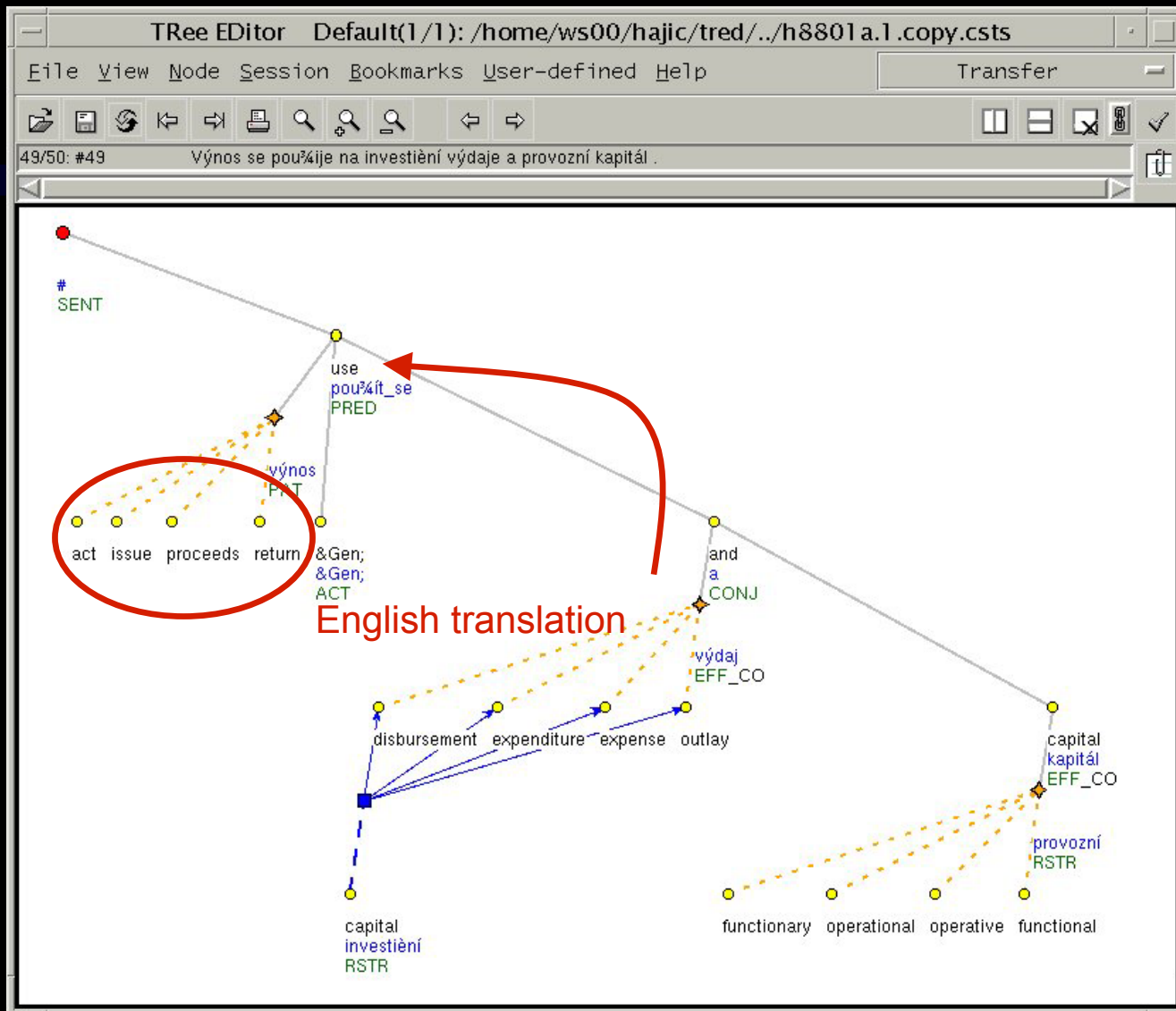
According to his opinion UAL's executives were misinformed about the financing of the original transaction.

- Generation [in an MT framework]
 - from: deep tectogrammatical representation
 - to: surface plain text form
 - via:
 - surface syntax representation (dependency/parse tree)
 - lemmas+tags (plus morphological synthesis)
- Key issues
 - lexical choice; “auxiliaries”; word order; morphology

WS02 Data + Tools Available

- Translation of the Penn Treebank to Czech (250k)
- Czech analysis (text → tectogrammatical representation)
 - tagger
 - parser (surface syntax, from WS'98)
 - deep parser (rule-based)
- Translation dictionary (multiple translations, non-prob.)
- Tectogrammatic annotation of English (Penn TB data)
- English morphology (morpha: adapted, cleaned)
- Evaluation script

Generation: Complex Input



Evaluation



- Simple and measurable
 - BLEU (IBM): 5 variants, 250+250 sentences
- Two (three?) “tracks”:
 - baseline from TR English (manually annotated)
 - complete translation from Czech
 - ?from Proposition Bank (comparison to TR)
- Generation: rule based vs. statistical
- Does it work?
 - Compare also to word-based statistical MT (WS’99)