

Generation in MT



Progress Report #1

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

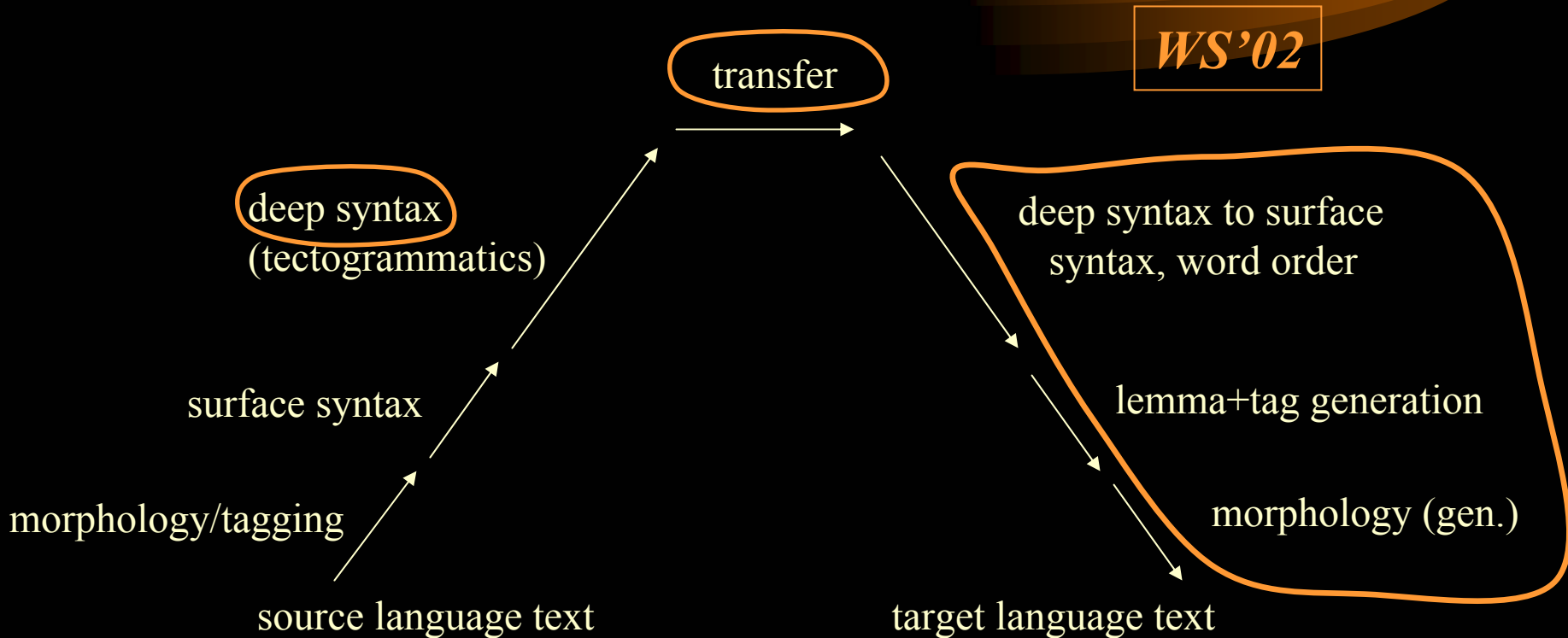
The Motivation

- Tectogrammatical Representation
 - linguistic intuition:
 - TR best represents structure & (linguistic) meaning
 - “best”: using compact description, as abstract as possible (wrt surface syntax, phrase structure)
 - => can be shared to a large extent among languages
 - => needs less data to train statistical models

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 details:

- Window Title: 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, navigation, and editing.
- Text Area: 3/51: #3 According to his opinion UAL's executives were misinformed about the financing of the original transaction.

Two tectogrammatical trees are displayed side-by-side:

- Left Tree:** Root node is a red dot labeled "#3 SENT". It branches to a yellow node labeled "informovat PRED". This node branches to five yellow nodes: "názor CRIT", "&Gen; ACT", "vedení PAT", "financování EFF", and "nesprávně MANN".
 - "názor CRIT" branches to "on APP".
 - "vedení PAT" branches to "UAL RSTR".
 - "financování EFF" branches to "transakce PAT", which further branches to "původní RSTR".
- Right Tree:** Root node is a red dot labeled "#3 SENT". It branches to a yellow node labeled "misinform PRED". This node branches to five yellow nodes: "opinion CRIT", "&Gen; ACT", "executive PAT", "financing EFF", and "original RSTR".
 - "opinion CRIT" branches to "he APP".
 - "executive PAT" branches to "UAL RSTR".
 - "financing EFF" branches to "transaction PAT", which further branches to "original RSTR".

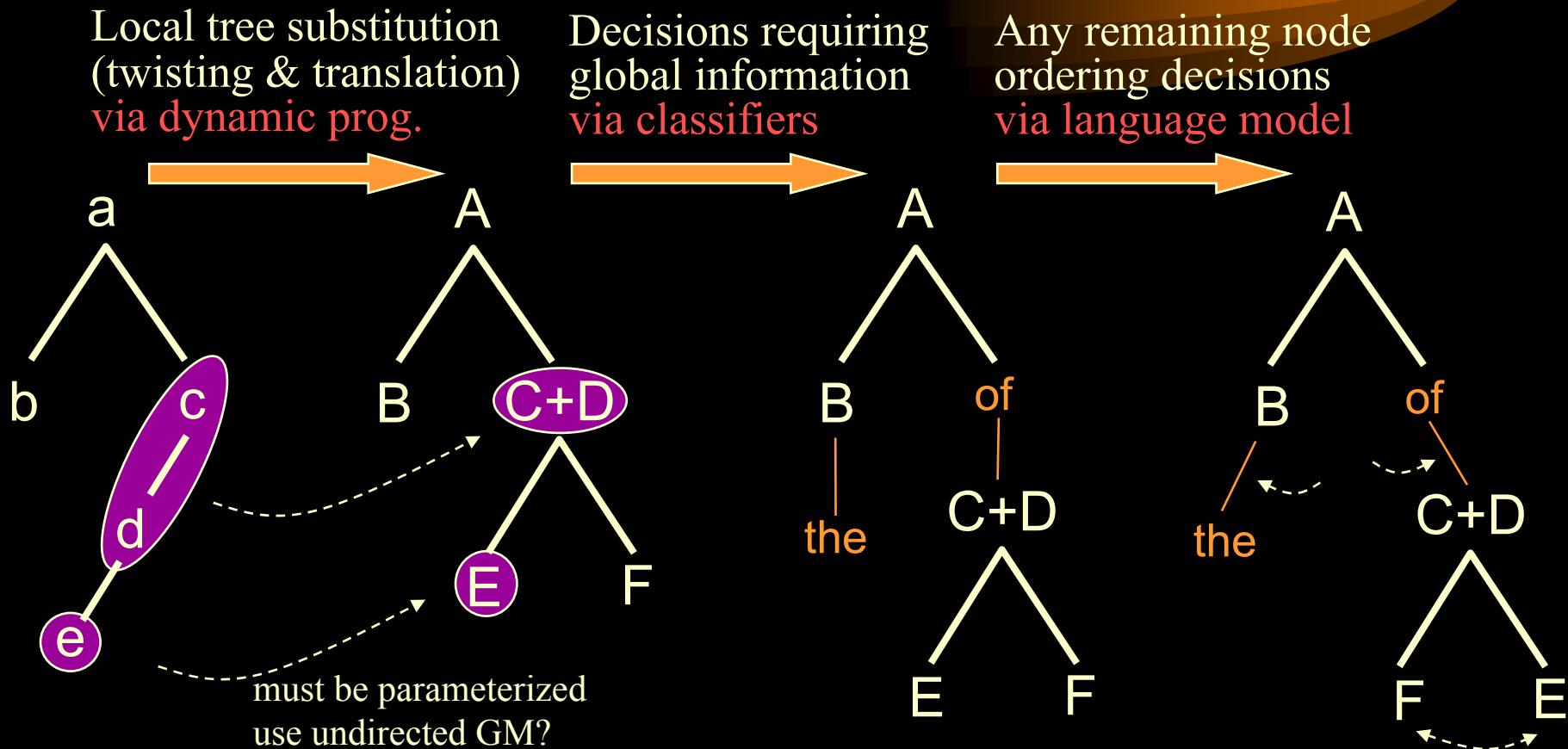
Progress so far (1)

- We have
 - Data and Czech-side tools at CLSP
 - tools working up to transfer (but without it yet)
 - looked at Cze/Eng TR/AR trees
 - spent 120 man-hours discussing how to do it
 - form of the model: source channel vs. classifier
 - features to be used
 - got data & info for conversion of Penn PropBank

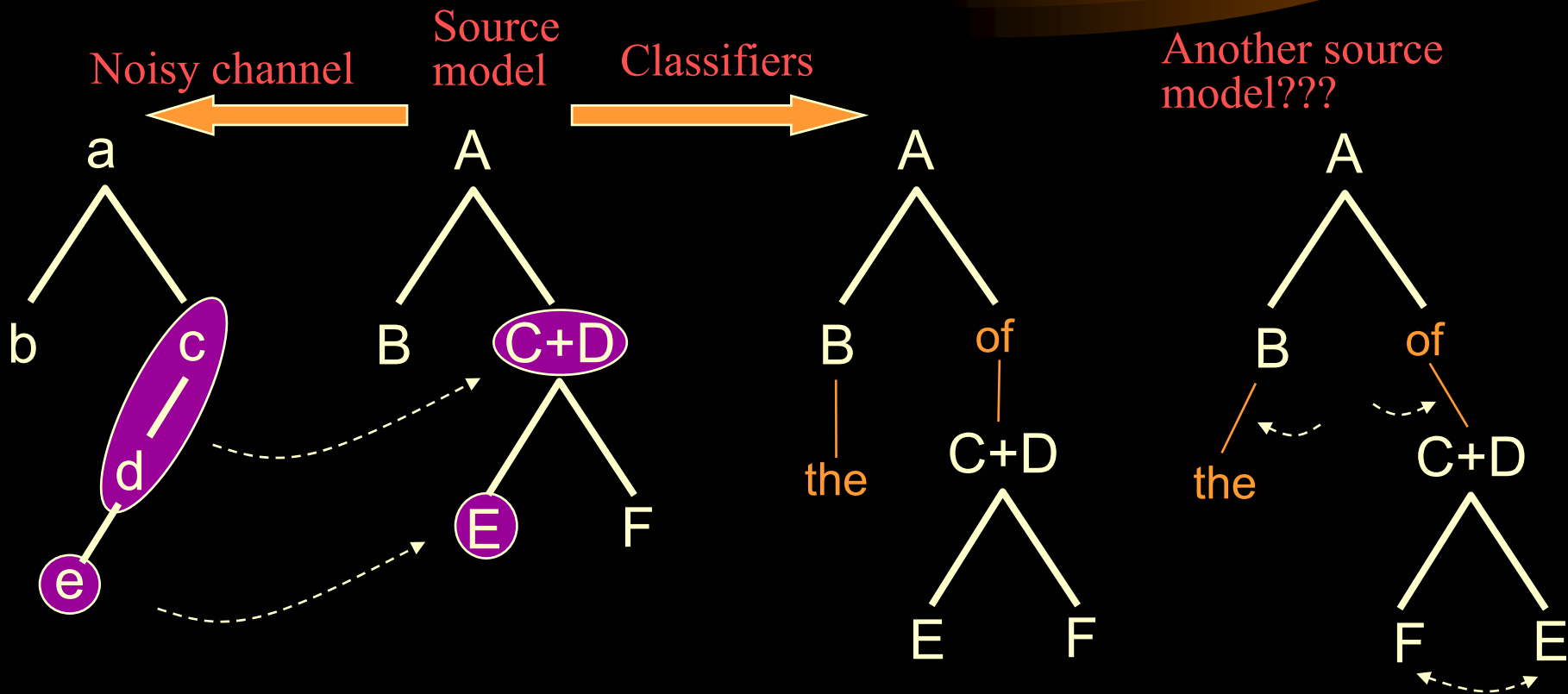
Progress so far(2)

- We have
 - compared the Eng TR trees with what's needed for the symbolic method's input
 - compared Eng TR trees: auto vs. manual
 - experimented with the Eng word-order model
 - working code for Eng morphology
 - working code and scripts for the evaluation experiments proper

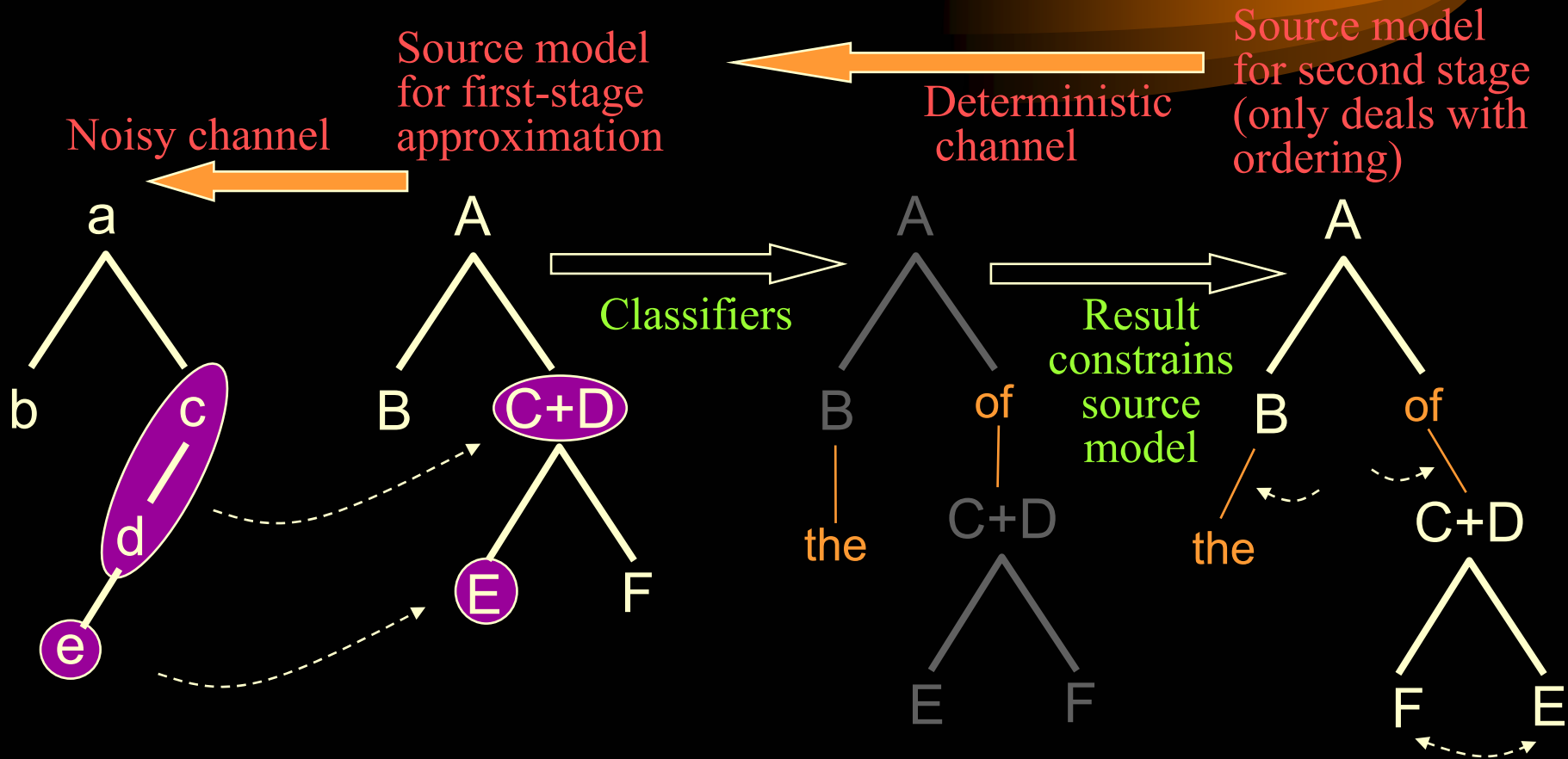
Division of Labor: Pipeline



Model Architecture



Current Compromise



English TR trees: auto vs. manual

- Wrote specific “diff” tools (YD)
- Test: 284 sentences, sect. 17 (5081 nodes)
 - TR lemma match: 93.4%
 - Functor match: 79.4%
 - Dependency (structural) match: 88.3%
 - Swapped dependencies: 36 (< 1%)

PropBank Input

- **PropBank:** Penn project, predicate-argument
- **Goal:** use PropBank to
 - Improve automatic construction of English TRs
 - Allow generation from “generic” pred-arg structures
- **Tasks**
 - Augment PropBank with roleset info ✓
 - Add lexical-conceptual role tags
 - Convert to TG (following Hajičová & Kučerová)

Word order

- Three major experiments (DG):
 - Tree-based models:
 - Collins model on PennTB style (parse) trees
 - 97% words at correct position reconstructed
 - Analytical level surface dependency, tree-based
 - 94% (chance: 68%)
 - levels ≥ 7 nodes ignored: 1.5% of nodes abs.
 - Bigram surface model, PennTB style trees
 - dynamic programming (begin/end words of phrase)
 - 86% (chance: 64%)

English Morphology

- Data (JC)

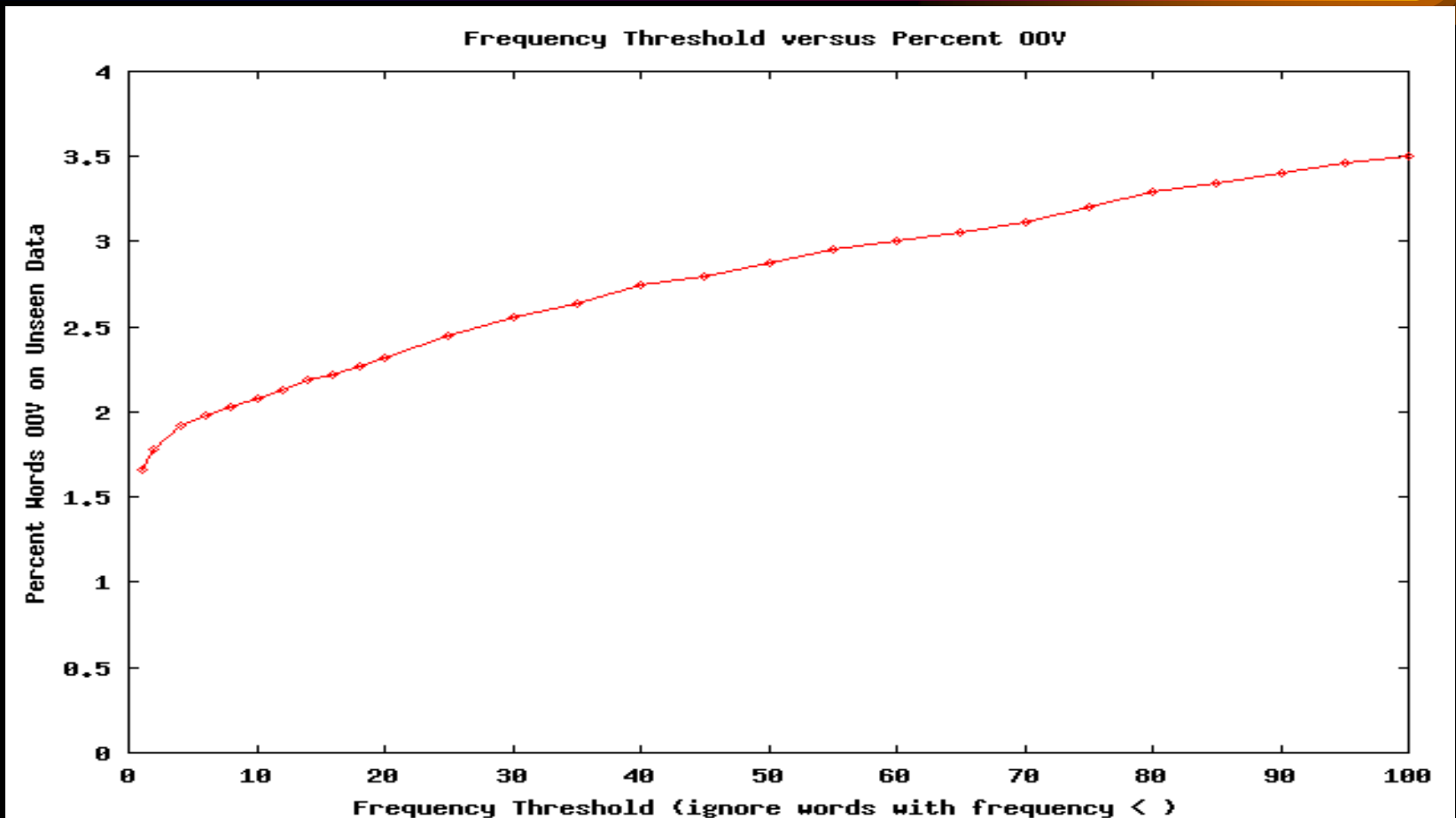
- WSJ (35 MW) analyzed by the “morpha” tool
 - PennTB compatible tagset + lemmatization
- table extracted (659843 entries), some corrected:

VBG	opening	open	14049
NNS	openings	opening	931
NN	opening	opening	12084

- Code (KP)

- table lookup (KP), accepts multiple formats
- currently correcting dictionary, overall error rate computation

Morphology Coverage



Symbolic approaches

- FUF/Surge (Elhadad/Robin): almost everything needs to be specified
- Nitrogen/Halogen (Langkilde-Geary/Knight): less specification is OK, uses statistical reranker

```

(define fd '
  ((cat clause)
  ;; (punctuation ((after "."))
  (common ((mood declarative)))
  (process ((type material)
            (lex "join")
            (tense future)))
  (partic ((agent (
                (complex apposition)
                ;; (punctuation ((after ","))
                (restrictive no)
                (distinct ~((;(cat proper)
                              (lex "Pierre Vinken"))
                              ((cat common)
                              (head ((lex "old")))
                              (determiner none)
                              (classifier ((cat measure)
                                          (quantity ((value 61)))
                                          (unit ((lex "years"))))))))
            (affected ((cat common)
                       (lex "board")
                       (definite yes))))))

(circum ((manner ((cat pp)
                  (prep ((lex "as")))
                  (np ((cat np)
                      (definite no)
                      (describer ((lex "nonexecutive")))
                      (head ((lex "director"))))))))
  (time ((cat adv-p)
         (ralt (((position end)))
         (head ((lex "Nov. 29")))
         ))))

```

Output of FUF/Surge



Pierre Vinken, 61 years old, will step down as nonexecutive director of the board Nov. 26.

Evaluation

- Evaluation scripts (TK)
 - Multiple dimensions of evaluation, presentation
 - Core software: BLEU
 - by and from Kishore Papineni
 - Format conversion
 - from morphology output
 - from translations
 - Status: three reference translations so far

Evaluation - presentation

