# Maximum Spanning Malt: Parser Voting for Indian Languages

Dan Zeman

*Presented by Pavel Straňák*

ÚFAL MFF, Univerzita Karlova, Praha

Grant MSM 0021620838

# System of Three Parsers

- MST Parser (McDonald et al., 2005)
  - http://sourceforge.net/projects/mstparser/
    - Starts with a fully connected graph, finds spanning tree with maximal weights. Weights computed from features (machine learning).
- Malt Parser (Nivre et al., 2007)
  - http://maltparser.org/
    - Transitions *shift, left/right-arc, swap*. Sequence of transitions constructed by oracle, trained on features.
- DZ Parser (Zeman, 2004)
  - http://ufal.mff.cuni.cz/~zeman/projekty/parser/
    - Looks at whole graph, bigrams of m-tags, orig. for Czech.

# The Baseline

- We were primarily interested in optimizing the *unlabeled attachment score (UAS).*
  - Dependency labels can be assigned later by an independent classifier.
  - If UAS goes higher, labeling accuracy should improve, too.
  - All scores presented are from dev data

| | MST | Malt | DZ |
|---|---|---|---|
| hi | 80.32 | **81.84** | 62.00 |
| bn | 82.00 | **84.71** | 71.02 |
| te | 77.63 | **80.89** | 70.52 |

# Weighted Voting

- A parser's vote weighs the more the higher UAS
  - Select the parent node with highest vote-weight, unless it creates a cycle.
  - If necessary, break a cycle by discarding a parent candidate that would otherwise win.

| | MST | Malt | DZ | Vote |
|---|---|---|---|---|
| hi | 80.32 | 81.84 | 62.00 | **82.48** |
| bn | 82.00 | **84.71** | 71.02 | 83.11 |
| te | 77.63 | **80.89** | 70.52 | 80.59 |

# Morphology

- So far, only word forms *chunk tags* were used.
- What is available:
  - Form = उसके usake (his)
  - Lemma = वह vaha (he)
  - Chunk = NP (noun phrase)
  - POS = PRP (personal pronoun)
  - g-m … gender = masculine
  - n-s … number = singular
  - p-a … person = ?
  - c-1 … case = oblique
  - v-kA_sAmanA … vibhakti (postposition) = का सामना kA sAmanA (opposite of)
  - t-$ … tense / aspect / modality = unknown

# Morphology

- We change POS that the parsers learn:
  - Form = उसके usake (his)
  - Lemma = वह vaha (he)
  - Chunk = NP (noun phrase)
  - POS = PRP|c-1|v-kA_sAmanA
  - g-m … gender = masculine
  - n-s … number = singular
  - p-a … person = ?
  - c-1 … case = oblique
  - v-kA_sAmanA … vibhakti (postposition) = का सामना kA sAmanA (opposite of)
  - t-$ … tense / aspect / modality = unknown

# Results with Morphology

- Helps for MST/all, Malt/hi, DZ/hi
  - DZ is worse on bn, te
  - But even though it improves voting!
    - Measured when the other parsers were fixed.

| | MST | Malt | DZ | Vote |
|---|---|---|---|---|
| hi | 86.16 | 85.84 | 75.12 | **87.12** |
| bn | 85.70 | 77.31 | 54.38 | **85.82** |
| te | **79.85** | 77.78 | 45.78 | 79.70 |

# Results with Morphology

- Malt uses morphology for Hindi only
- MST and DZ use it for all languages

| | MST | Malt | DZ | Vote |
|---|---|---|---|---|
| hi | 86.16 | 85.84 | 75.12 | **87.12** |
| bn | 85.70 | 84.71 | 54.38 | **86.19** |
| te | 79.85 | 80.89 | 45.78 | **82.37** |

# Different Malt Algorithms

- Not all differences are statistically significant

| | hi | bn | te |
|---|---|---|---|
| nivreeager | 85.84 | 84.71 | 80.89 |
| nivrestandard | 86.56 | 85.45 | 80.15 |
| covproj | 86.32 | **85.57** | 80.30 |
| covnonproj | 86.96 | 84.09 | 80.30 |
| stackproj | 86.56 | 85.08 | 80.89 |
| stackeager | 81.76 | 83.85 | **81.04** |
| stacklazy | **87.60** | 84.71 | 80.59 |

# Nonprojectivity

- Hindi: 1.83 % edges, 13.93 % sentences
- Bangla: 0.96 % edges, 5.49 % sentences
- Telugu: 0.45 % edges, 1.31 % sentences

- Malt: covnonproj, stackeager, stacklazy
  can create non-projective trees

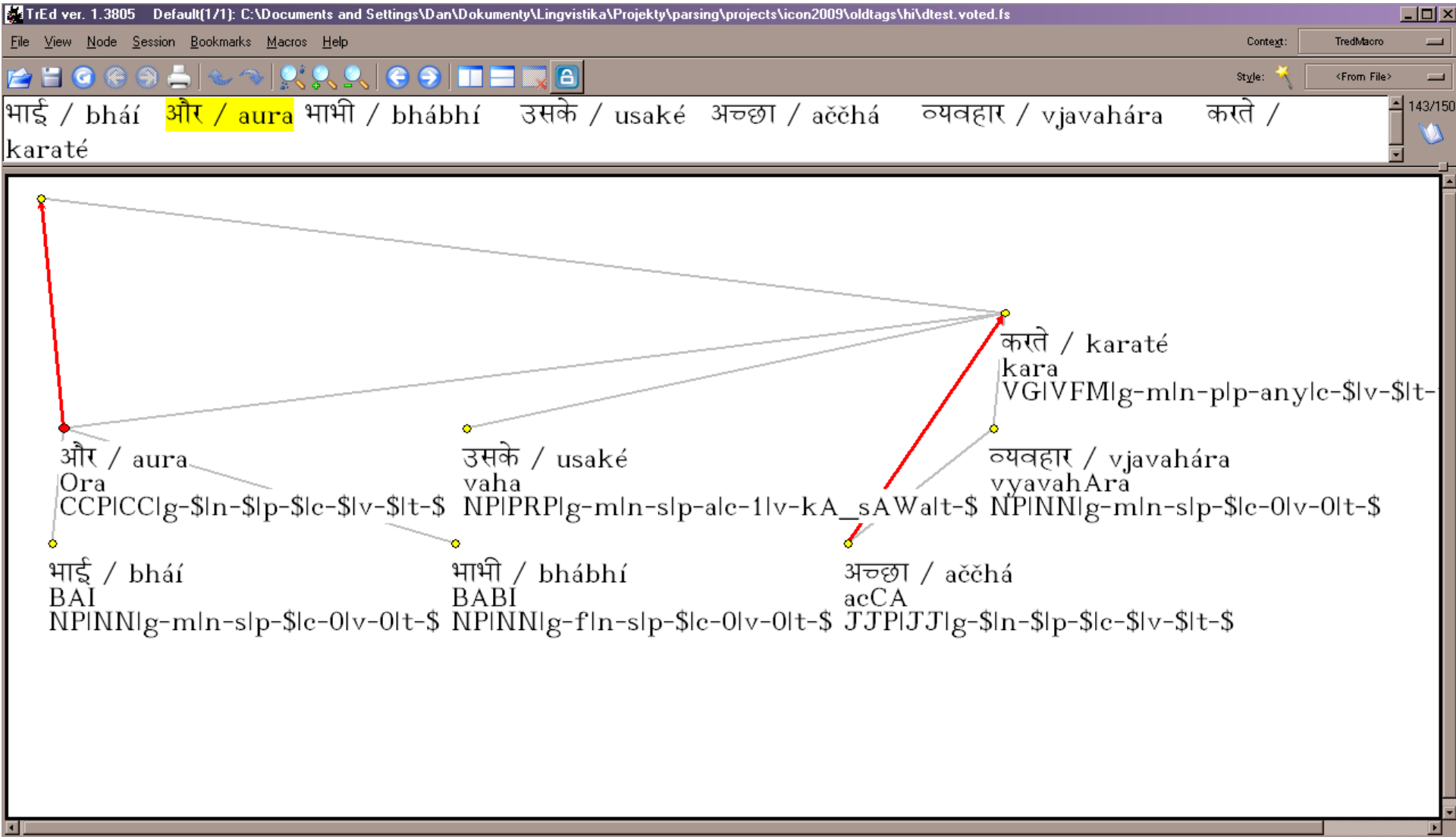
- MST: We were not able to try the non-projective mode (no time)

# Results with Best Malt Algorithm

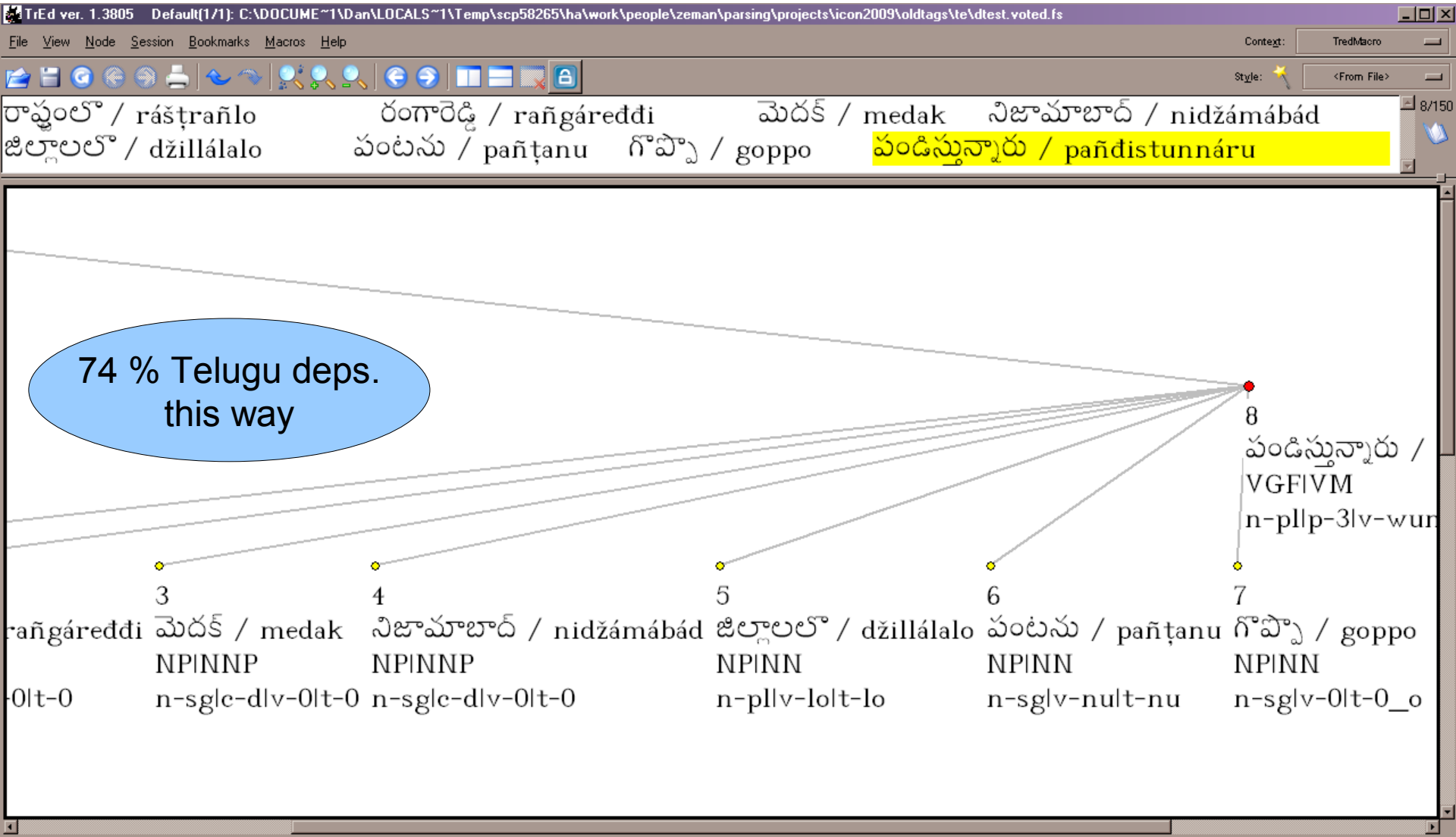
- Malt uses a different algorithm for each language
  - Malt uses morphology for Hindi only
  - MST and DZ use it for all languages

| | MST | Malt | DZ | Vote |
|---|---|---|---|---|
| hi | 86.16 | 87.60 | 75.12 | **88.00** |
| bn | 85.70 | 85.57 | 54.38 | **86.56** |
| te | 79.85 | 81.04 | 45.78 | **82.52** |

# Some Error Patterns: Coordination



# “Naïve Telugu” Structure



# References


- McDonald R., Pereira F., Ribarov K., Hajič J.: *Non-projective dependency parsing using spanning tree algorithms.* HLT-EMNLP 2005
- Nivre J., Hall J., Nilsson J., Chanev A., Eryiğit G., Kübler S., Marinov S., Marsi E.: *Maltparser: A language-independent system for data-driven dependency parsing.* Natural Language Engineering, 13(2):95-135, 2007
- Zeman D.: *Parsing with a statistical dependency model* (Ph.D. Thesis), 2004
- Zeman D., Žabokrtský, Z.: *Improving parsing accuracy by combining diverse dependency parsers.* IWPT 2005