Using alignment of tectogrammatical trees in phrase-based machine translation

David Mareček

marecek@ufal.mff.cuni.cz

Statistical Machine Translation seminar

May 11, 2009, Prague

Outline

- Alignment of tectogrammatical trees and its advantages
- GIZA++ word alignment and symmetrization methods
- Combining various alignments
- Testing usability of these alignments on the MOSES toolkit

Alignment of Tectogrammatical Trees

Tectogrammatical tree = rooted dependency tree where only autosemantic (content) words have their own nodes.

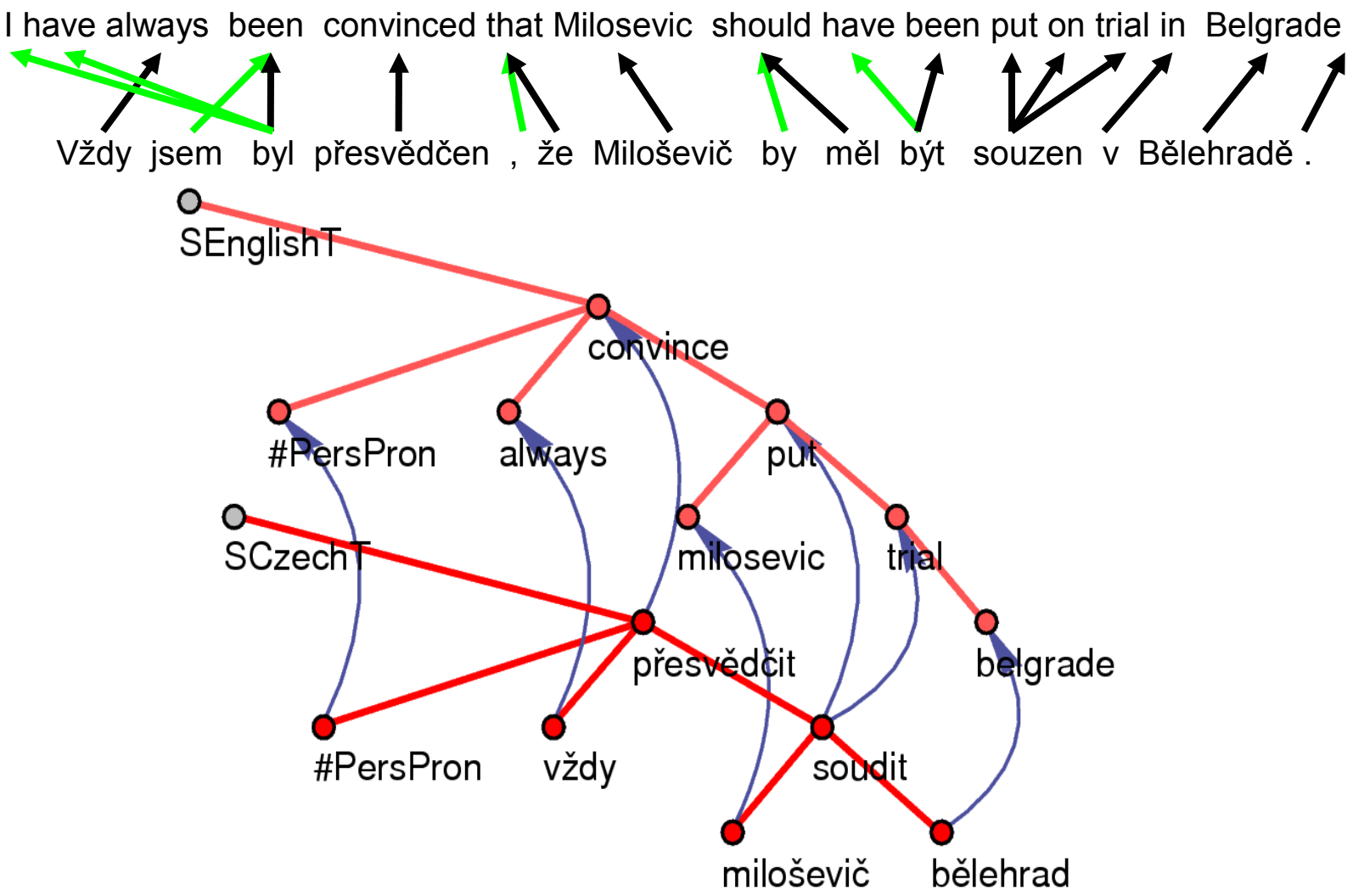
Advantage of tectogrammatical alignment over word alignment:

- Functional words (e.g. articles, prepositions, auxiliary verbs, modal verbs …), that are often problematic to align (they can have different functions in different languages), don't have their own node in the tectogrammatical layer – we needn't align them.
- The tree structure may help.

Disadvantages:

- Errors in tagging and parsing often causes errors in the alignment.

Tecto-alignment x Word-alignment



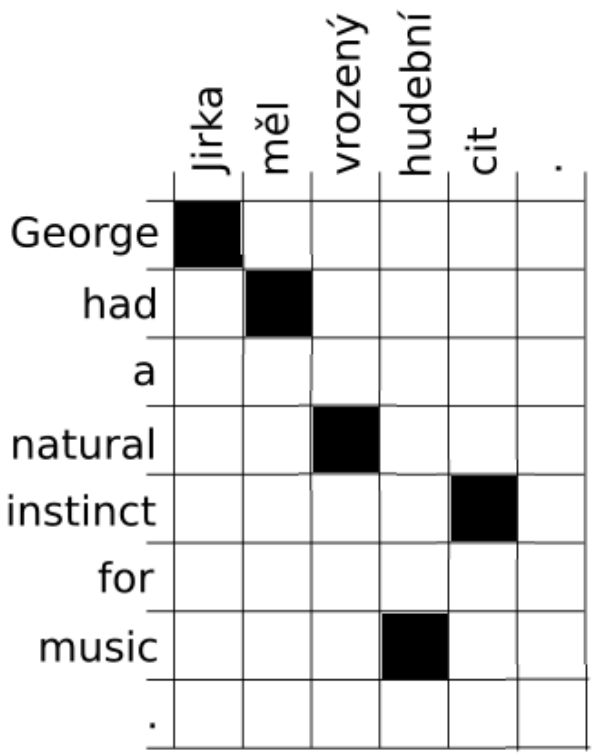
AER and IAA

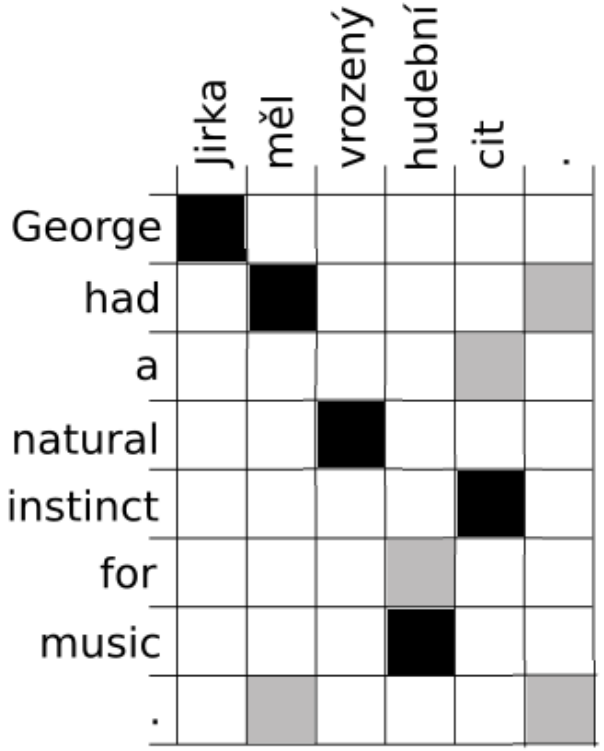
- AER - Alignment error-rate
  - heuristic metric for word-alignment (Och and Ney, 2003)
- IAA - Inter-annotator alignment
- Our testset: 2,500 manually aligned sentences

	IAA	tool	AER
word-alignment	89,6 %	GIZA++	13.2 %
tecto-alignment	94,6 %	T-aligner + GIZA++	7.3 %

Tecto-alignment on the Surface

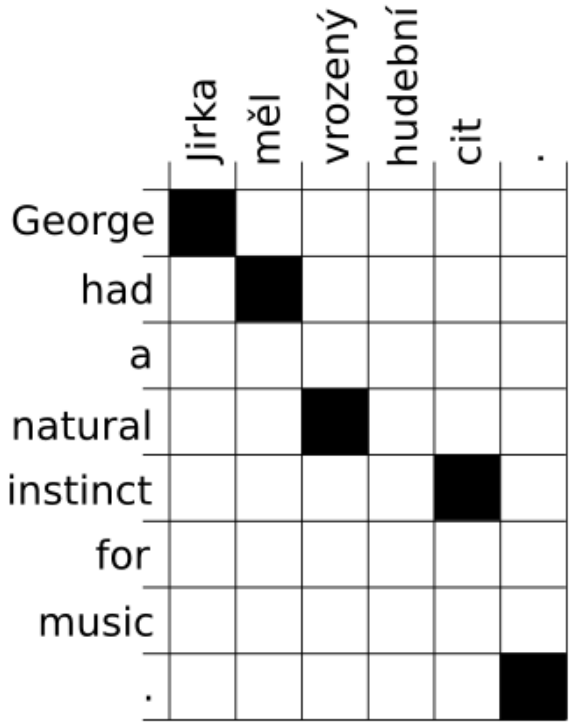
- We link all content words that were linked in tectogrammatical trees.
- We can link also functional words:
  - Two words are linked if the content words they belongs to are also linked.



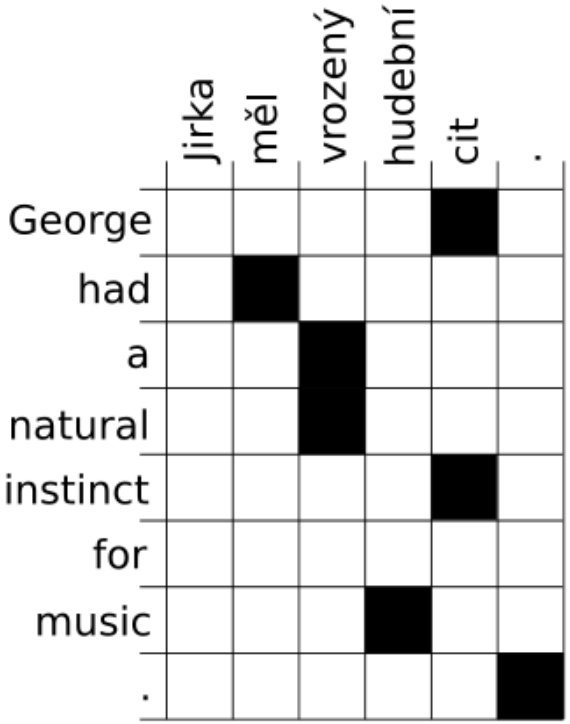


GIZA++ Word alignment

- Sentences are lemmatized before running GIZA++.
- IBM models create a many-to-one mapping – assymetric output.
- We run GIZA++ in both directions and symmetrize the outputs.

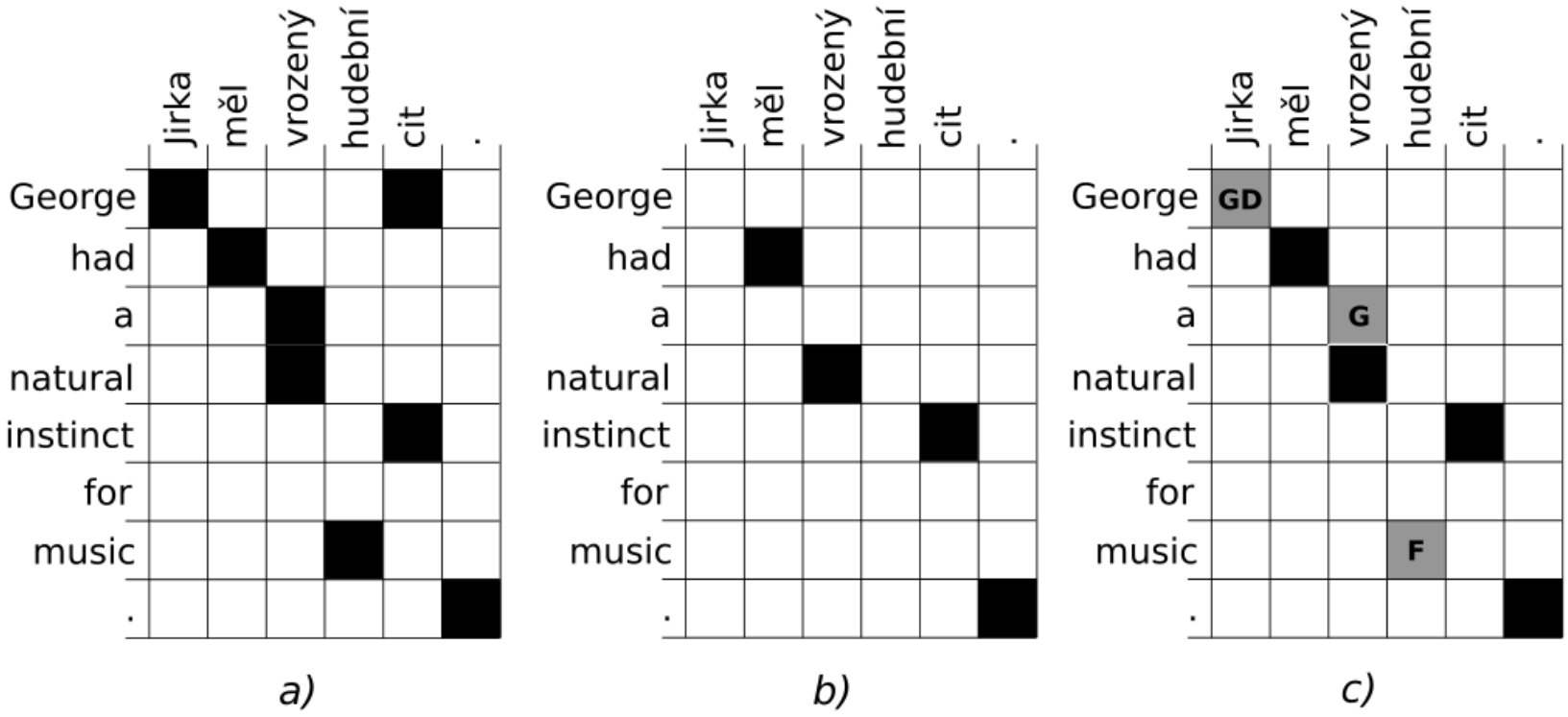


a)



b)

Symmetrization methods



union intersection grow-diag-final

Combining alignments

- We combine word alignments acquired from GIZA++ and tecto-alignment together
- Our goal is to get an alignment with lower AER or higher BLEU when applied to MOSES
- We used this combinations:
  - GIZA_int ∪ tecto_lex
  - GIZA_gdfa ∪ tecto_lex
  - GIZA_gdfa ∩ tecto_lex+aux
  - (GIZA_gdfa ∩ tecto_lex+aux) ∪ GIZA_int

Running MOSES

- Direction EN → CZ
- Training, tuning and testing data from WMT08
- MERT tuning

Results

alignment	density	AER	BLEU
giza_int	0,62	13,2	12,4
giza_gdfa	0,92	20,3	**12,9**
giza_int ∪ tecto_lex	0,74	**10,7**	12,3
giza_gdfa ∪ tecto_lex	1,00	17,8	12,7
tecto_lex+aux	1,43	37,8	8,9
giza_gdfa ∩ tecto_lex+aux	0,60	18,4	12,3
(giza_gdfa ∩ tecto_lex+aux) ∪ giza_int	0,71	19,6	12,8

Future Work

- Try other combinations of alignments
- Use bidirectional tectogrammatical alignment (one-to-many alignment in both directions)
  - First combine the two unidirectional alignments with GIZA++ alignments
  - Then symmetrize
- Extend MOSES phrase table - an information whether a phrase matches with tectogrammatical alignment or not