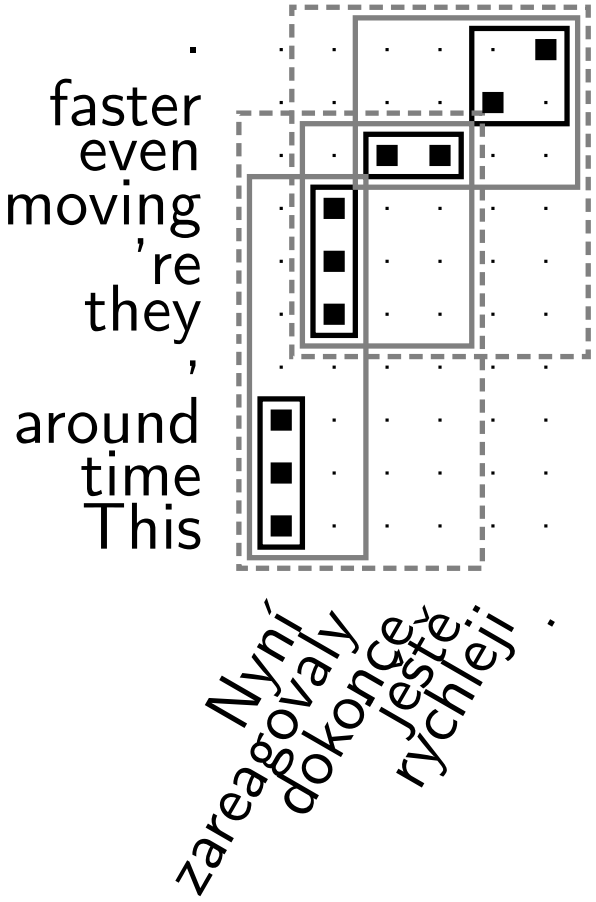
Word Alignment as a Combinatorial Task

Ondřej Bojar
bojar@ufal.mff.cuni.cz
Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics
Charles University, Prague

Outline

- Introduction:
  - When is word alignment used.
  - State of the art.
- Inexact matches $\Rightarrow$ possible alignments.
- Tectogrammatical alignment and LEAF.
- The question for combinatoritians.
- Demotivating examples.
- Summary.

Key Application: Machine Translation



This time around	=	Nyní
they 're moving	=	zareagovaly
even	=	dokonce ještě
...	=	...
This time around, they 're moving	=	Nyní zareagovaly
even faster	=	dokonce ještě rychleji
...	=	...

Phrase-based MT: choose such segmentation of input string and such phrase "replacements" to make the output sequence "coherent" (3-grams most probable).

State of the Art

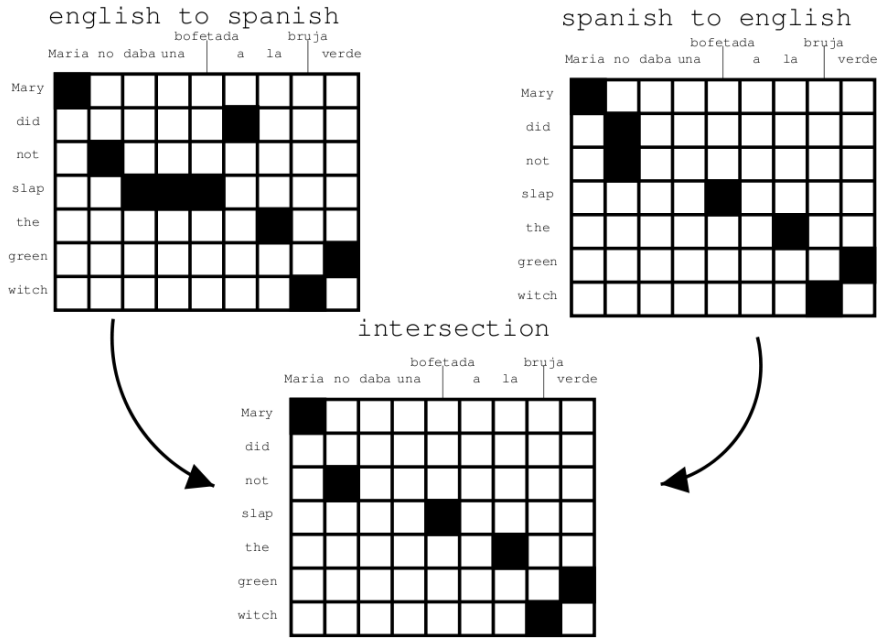
GIZA++ (Och and Ney, 2000):

- Unsupervised, only sentence-parallel texts needed.
- Word alignments formally restricted to a function:
  (Already word-for-word translation is NP-complete; (Knight, 1999).)

  src token $\mapsto$ tgt token or NULL

- A cascade of models refining the probability distribution:
  - IBM1: only lexical probabilities: $P(kočka = cat)$
  - IBM3: adds fertility: 1 word generates several others
  - IBM4/HMM: to account for relative reordering
- Only many-to-one links created $\Rightarrow$ used twice, in both directions.

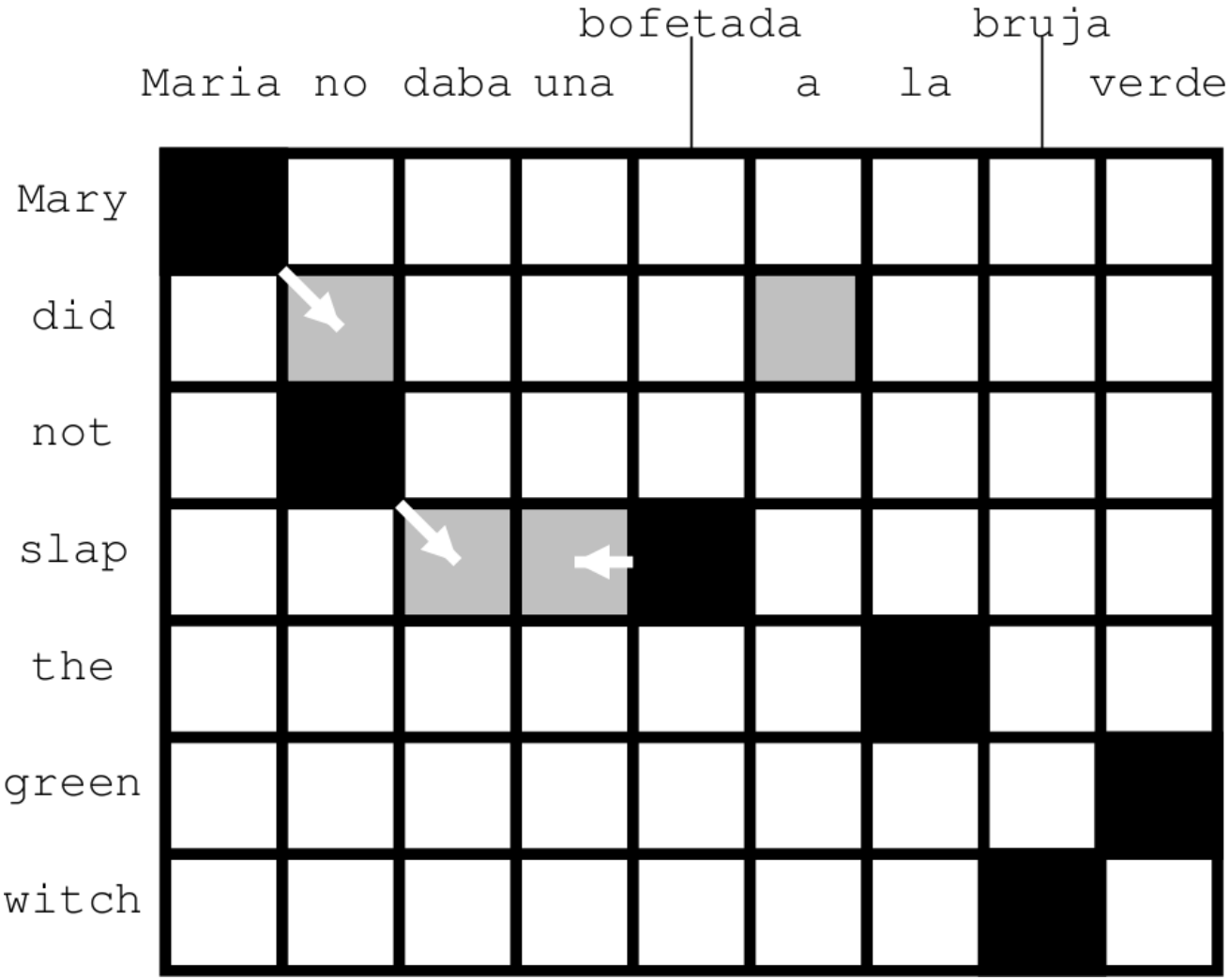
Symmetrization



"Symmetrization" of the two runs:

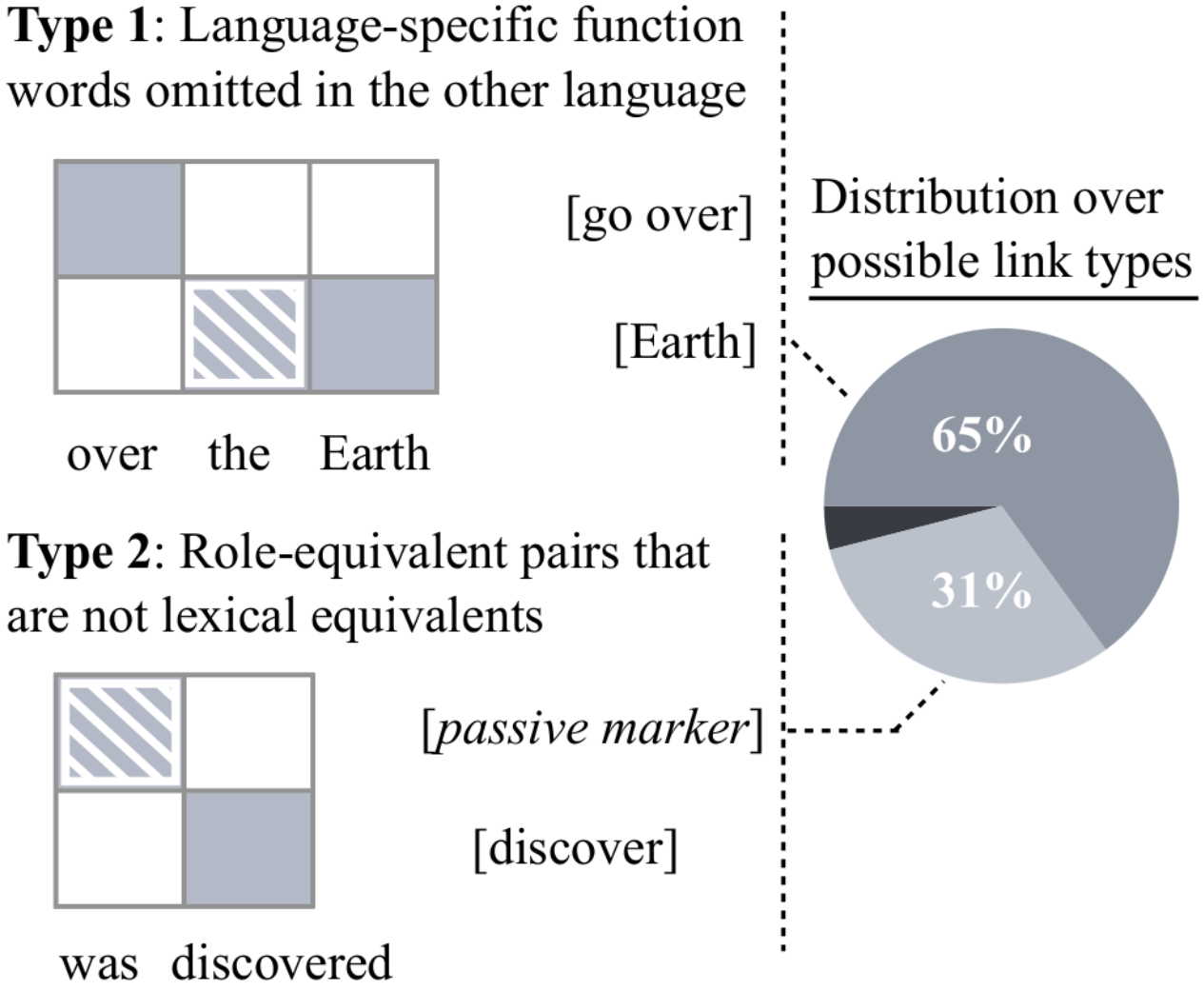
- intersection: high precision, too low recall.
- popular: heuristical (something between intersection and union).
- minimum-weight edge cover (Matusov et al., 2004).

Popular Symmetrization Heuristic



Extend intersection by neighbours of the union (Och and Ney, 2003).

(Inexact) Possible Alignments



Chinese-English from DeNero and Klein (2010).

Human Disagreements

- Humans have troubles aligning word for word.
  - Mismatch in alignments points 9–18%. (Bojar and Prokopová, 2006)

Top Problematic Words

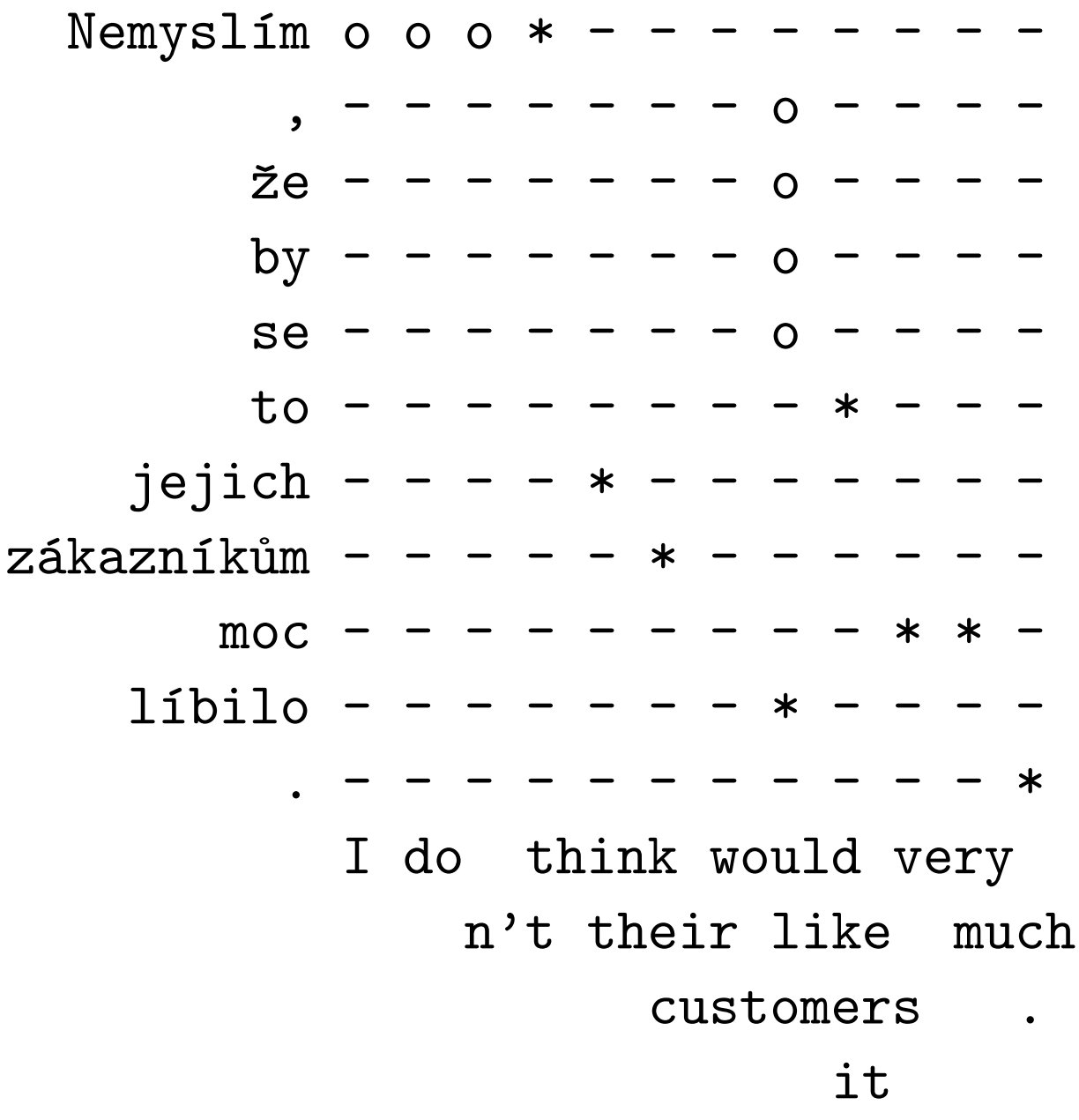
English		Czech	
361	to	319	,
259	the	271	se
159	of	146	v
143	a	112	na
124	,	74	o
107	be	61	že
99	it	55	.
95	that	47	a

Top Problematic Parts of Speech

English		Czech	
679	IN	1348	N
519	DT	1283	V
510	NN	661	R
386	PRP	505	P
361	TO	448	Z
327	VB	398	A
310	JJ	280	D
245	RB	192	J

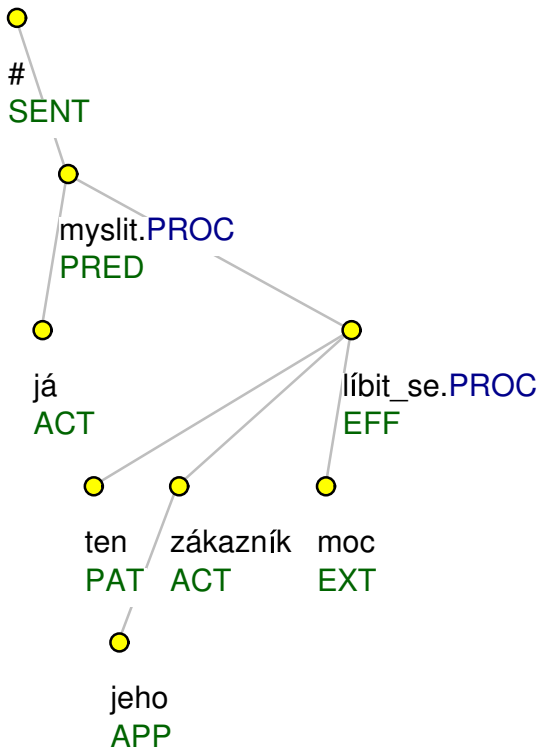
- Where people fail to agree, improving GIZA++ does not help.

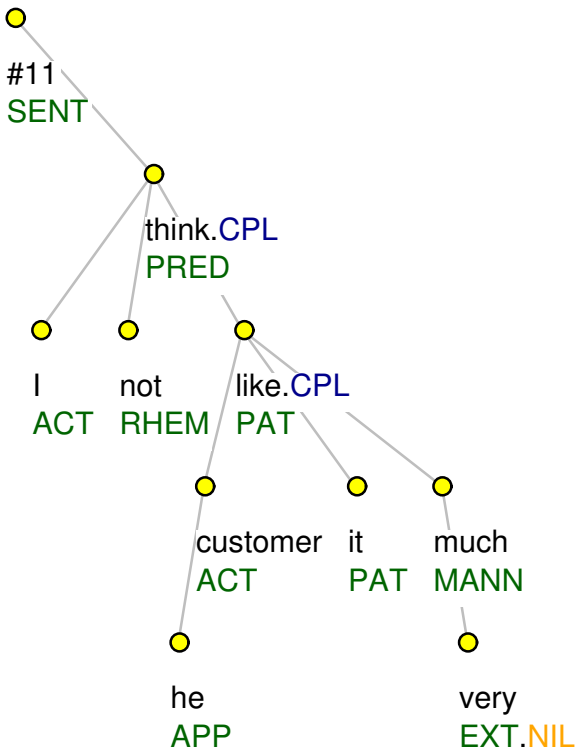
A Czech-English Example



ÚFAL's Family Jewels: T-Layer

- Only content-bearing words have a node.
- Auxiliary words hidden, dropped pronouns added.





(já) Nemyslím **, že by se** to jejich zákazníkům moc líbilo **.**

I **do** n't think their customers **would** like it very much **.**

Tectogrammatical Alignment

- Mareček et al. (2008) align t-nodes, not words.
  $\Rightarrow$ Auxiliary words do not clutter the task.
- Improves human agreement from 91% to 94.7%.
- Application to phrase-based MT: (Mareček, 2009)
  - Improved alignment error rate on content words.
  - Minor improvements in BLEU when combined with GIZA++.

Main disadvantage:

- Language-dependent.
- Heavy use of tools (tagging, parsing, deep parsing).

Related: Fraser and Marcu (2007)

- A generative story called "LEAF" divides:
  - Source words into: head, non-head, deleted.
  - Target words into: head, non-head, spurious.
  - Heads connected across languages, non-heads within languages.

source	absolutely	[comma]	they	do	not	want	to	spend	that	money
word type (1)	DEL.	DEL.	HEAD	non-head	HEAD	HEAD	non-head	HEAD	HEAD	HEAD
linked from (2)			THEY	do	NOT	WANT	to	SPEND	THAT	MONEY
head(3)			ILS		PAS	DESIRENT		DEPENSER	CET	ARGENT
cept size(4)			1		2	1		1	1	1
num spurious(5)	1									
spurious(6)	aujourd'hui									
non-head(7)			ILS	PAS	ne	DESIRENT	DEPENSER	CET	ARGENT	
placement(8)	aujourd'hui		ILS	ne	DESIRENT	PAS	DEPENSER	CET	ARGENT	
spur. placement(9)			ILS	ne	DESIRENT	PAS	DEPENSER	CET	ARGENT	aujourd'hui

- Probabilities in the generative story learnt unsupervised:
  - Starting from GIZA++ outputs.
  - Greedy local updates of alignments to increase the likelihood of the data.

Question for Combinatoritians

Ultimate goal:

Find **minimum translation units** $\sim$ graph partitions:

- such that they are frequent across many sentence pairs.
- without imposing (too hard) constraints on reordering.
- in an unsupervised fashion.

Available data: Word co-occurrence statistics:

- In large monolingual data (usually up to 10^9 words).
- In smaller parallel data (up to 10^7 words per language).
- Optional automatic rich linguistic annotation.

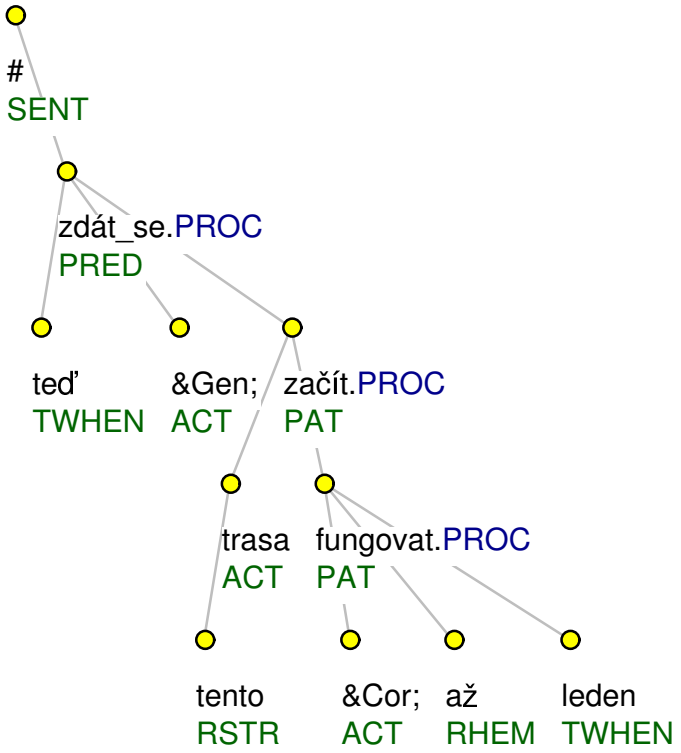
Better Translation ⇝ Uglier Ali. (1)

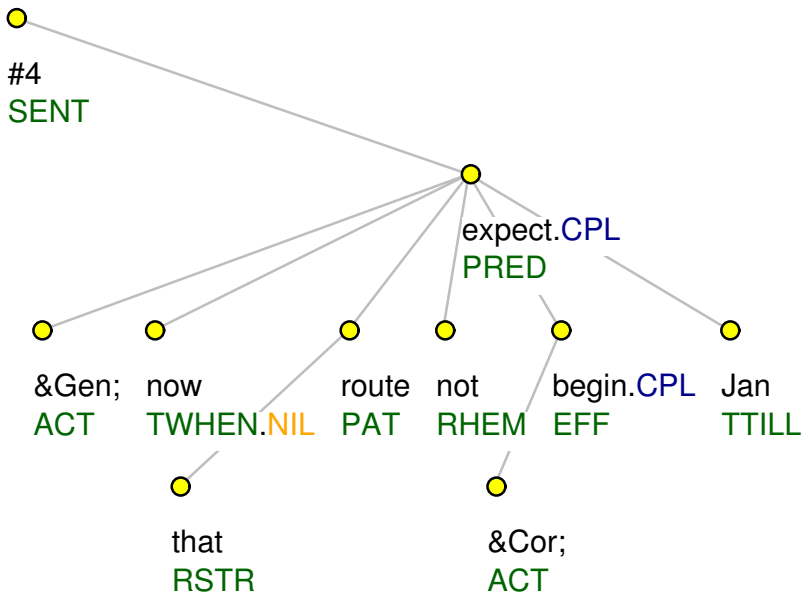
The better (more fluent) translation, the harder to align:

```
      to o * - - - - - - - - - - -
     get - - * - - - - - - - - - -
      in - - - - - - - - @ O O O -
   shape - - - - - - - - O O O @ -
     for - - - * - - - - - - - - -
     the - - - - - - - o - - - - -
    1990s - - - - * * * - - - - - -
       . - - - - - - - - - - - - *
         , aby do  . let co  formě
             vstoupila v   nejlepší
                 90              .
```

Better Translation ⇝ Uglier Ali. (2)

T-layer to no rescue:





Teď se zdá , že tyto trasy začnou fungovat až v lednu .

Now , those routes are n't expected to begin until Jan .

Summary

- Word alignment essential for machine translation, extraction of dictionaries, paraphrasing (synonymy). . .
- Current practice is full of hacks:
  - Word alignments for phrase-based translation.
  - Unrealistic restrictions and heuristical symmetrization.
- More appropriate approaches:
  - too much language-dependent (T-Alignment).
  - suboptimal (LEAF).
- Lots of data available.

$\Rightarrow$ Please come and help us.

References

Ondřej Bojar and Magdalena Prokopová. 2006. Czech-English Word Alignment. In Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006), pages 1236–1239. ELRA, May.

John DeNero and Dan Klein. 2010. Discriminative modeling of extraction sets for machine translation. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 1453–1463, Uppsala, Sweden, July. Association for Computational Linguistics.

Alexander Fraser and Daniel Marcu. 2007. Getting the structure right for word alignment: LEAF. In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), pages 51–60, Prague, Czech Republic, June. Association for Computational Linguistics.

Kevin Knight. 1999. Decoding complexity in word-replacement translation models. Comput. Linguist., 25(4):607–615.

David Mareček, Zdeněk Žabokrtský, and Václav Novák. 2008. Automatic Alignment of Czech and English Deep Syntactic Dependency Trees. In Proceedings of EAMT 2008, Hamburg, Germany.

David Mareček. 2009. Using Tectogrammatical Alignment in Phrase-Based Machine Translation. In Jana Šafránková, editor, WDS'04 Proceedings of Contributed Papers, Prague. Charles University, Matfyzpress.

E. Matusov, R. Zens, and H. Ney. 2004. Symmetric Word Alignments for Statistical Machine Translation. In Proceedings of COLING 2004, pages 219–225, Geneva, Switzerland, August 23–27.

Franz Josef Och and Hermann Ney. 2000. A Comparison of Alignment Models for Statistical Machine Translation. In Proceedings of the 17th conference on Computational linguistics, pages 1086–1090. Association for Computational Linguistics.

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models.

References

Computational Linguistics, 29(1):19–51.

Good Translation ⇝ Ugly Alignment