A Tookit for Visualizing the Coherence of Tree-based Reordering with Word Alignments

Authors:

Gideon Maillette de Buy Wenniger,
Maxim Khalilov, Khalil Sima'an

Statistical Language Processing and Learning Lab at the
Institute for Logic Language and Computation (ILLC),
University of Amsterdam, the Netherlands

Big Picture: Motivation

- Improving SMT:
  - Better models
  - Better learning methods and decoding

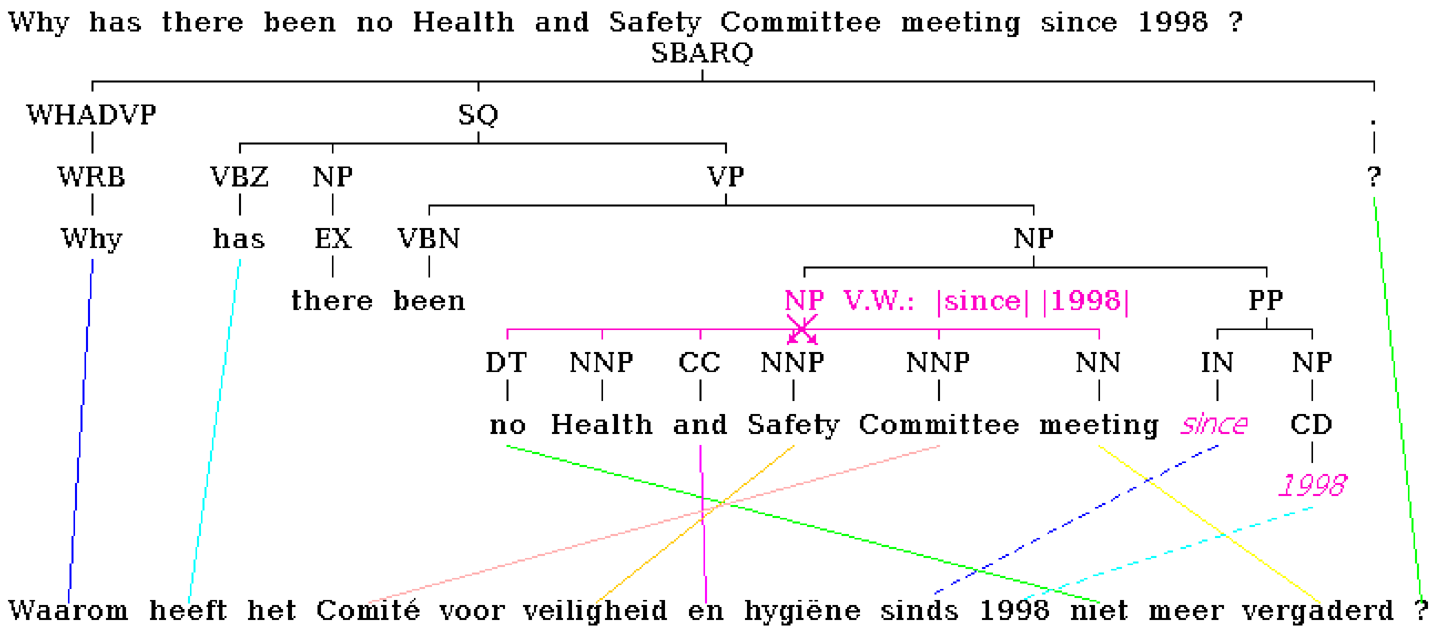
- Better models:
  - More sensible alignments
  - Explicit language-specific reordering models
  - Adding all sorts of extra information

- This work: support search better models

- Data visualization facilitates SMT research

Motivating Example

- Tree Structures: basis **syntactic SMT**
- ***Un-cohesiveness*** resulting from negation

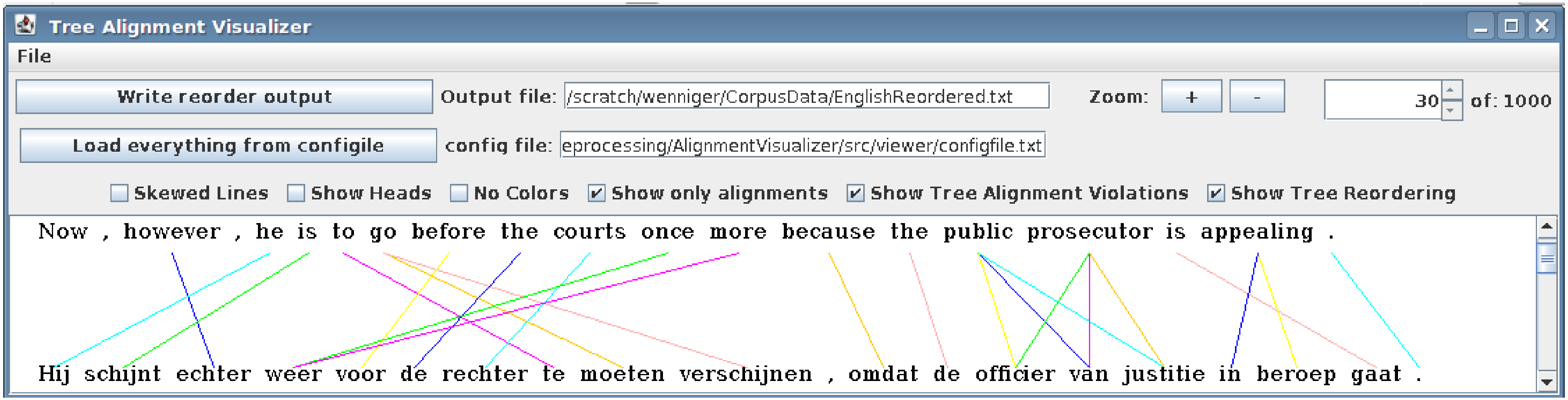


Why has there been no Health and Safety Committee meeting since 1998 ?
SBARQ
WHADVP
SQ
.
WRB
VBZ
NP
VP
?
Why
has
EX
VBN
NP
there
been
NP V.W.: |since| |1998|
PP
DT
NNP
CC
NNP
NNP
NN
IN
NP
no
Health
and
Safety
Committee
meeting
since
CD
1998
Waarom heeft het Comité voor veiligheid en hygiëne sinds 1998 niet meer vergaderd ?

- Is this big NP subtree appropriate?
  - For translation?
  - For reordering?
- Insight in **coherence** trees and alignments

Basic Alignment Visualization

- Alignment between source and target sentence
  - General: m-to-n mappings between words



Output file: /scratch/wenniger/CorpusData/EnglishReordered.txt

config file: eprocessing/AlignmentVisualizer/src/viewer/configfile.txt

30 of: 1000

□ Skewed Lines □ Show Heads □ No Colors ☑ Show only alignments ☑ Show Tree Alignment Violations ☑ Show Tree Reordering

Now , however , he is to go before the courts once more because the public prosecutor is appealing .

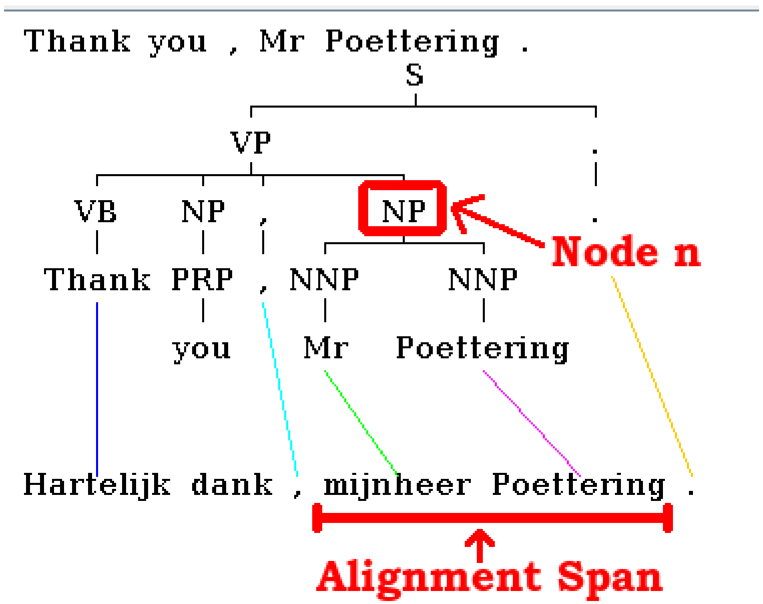
Hij schijnt echter weer voor de rechter te moeten verschijnen , omdat de officier van justitie in beroep gaat .

Alignment Span of Nodes

- Alignment mapping function

$$A(n) \rightarrow \{1, \cdots, m\}^*$$

- Span of target positions covered by subtree rooted at node **n**


Thank you , Mr Poettering .
S
VP
.
VB
NP
,
NP
Node n
Thank
PRP
,
NNP
NNP
you
Mr
Poettering
Hartelijk dank , mijnheer Poettering .
Alignment Span

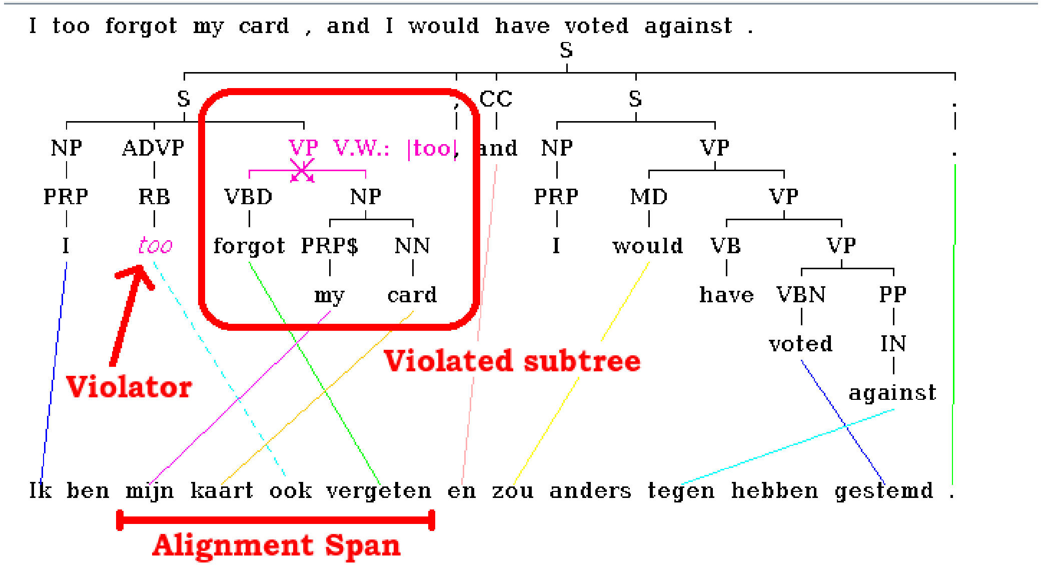
Definition 2.1 (Alignment Span)
AlignmentSpan(n) :=

$$[a_{n_{min}}, a_{n_{max}}] = \left[\min_{x \in \mathrm{LeafNodes}(n)} \left(\min_{a_{x'} \in A(x)} a_{x'} \right), \max_{y \in \mathrm{LeafNodes}(n)} \left(\max_{a_{y'} \in A(y)} a_{y'} \right) \right]$$

Cohesiveness

- Alignment ***Cohesive*** nodes: source side syntactic phrase pair

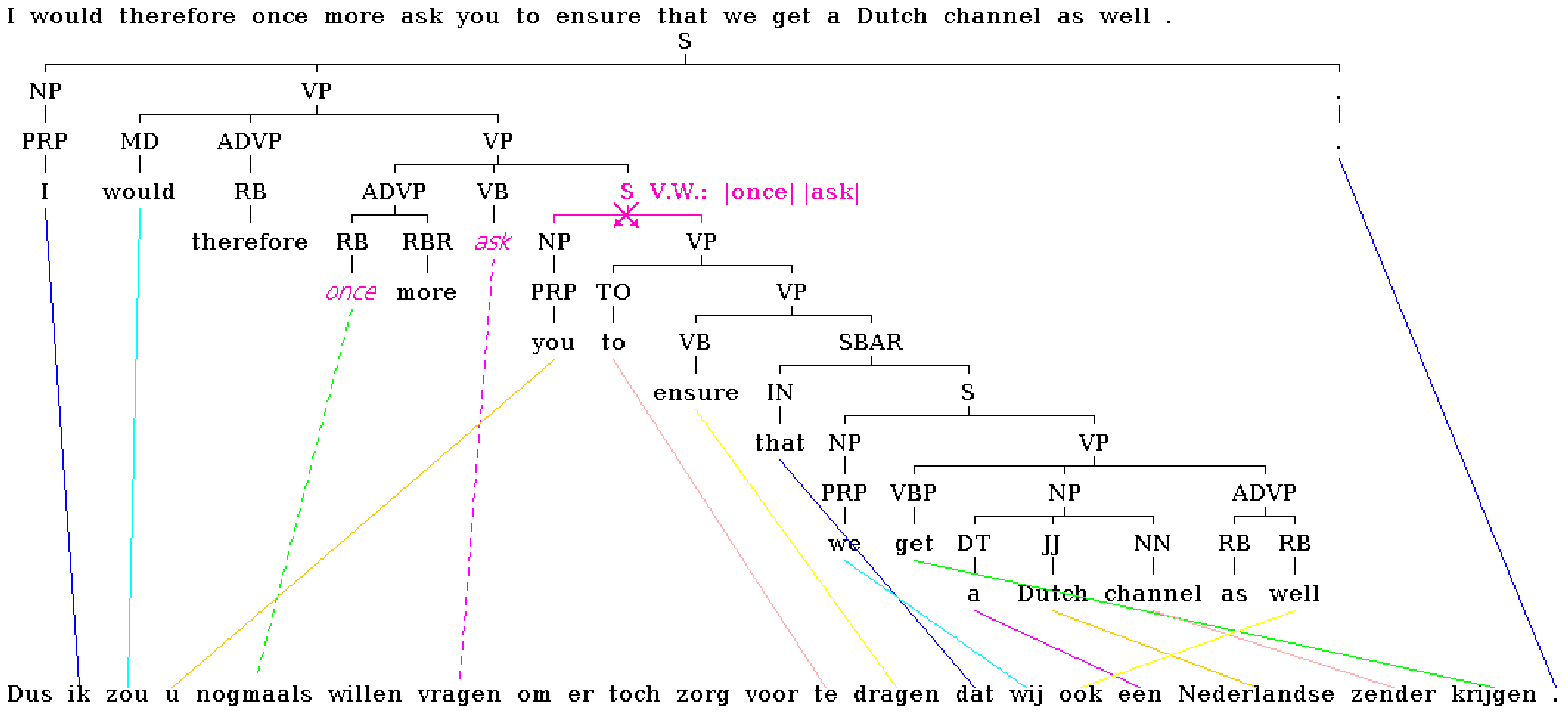
- Un-cohesive / **Alignment Violation**: Two distinct subtrees align within same target range



Definition 2.2 (Alignment Violation)

$$violates(n', n) := terminal(n') \land n' \notin \text{descendants}(n) \land$$
$$(\text{AlignmentSpan}(n) = [a_{n_{min}}, a_{n_{max}}]) \land$$
$$(a_{n_{min}} \leq A(n') \leq a_{n_{max}})$$

(Un)-Cohesiveness Example


I would therefore once more ask you to ensure that we get a Dutch channel as well .
S
NP
VP
.
PRP
MD
ADVP
VP
.
I
would
RB
ADVP
VB
S V.W.: |once| |ask|
therefore
RB
RBR
ask
NP
VP
once
more
PRP
TO
VP
you
to
VB
SBAR
ensure
IN
S
that
NP
VP
PRP
VBP
NP
ADVP
we
get
DT
JJ
NN
RB
RB
a
Dutch
channel
as
well
Dus ik zou u nogmaals willen vragen om er toch zorg voor te dragen dat wij ook een Nederlandse zender krijgen .

- ***Ask*** and ***once*** violate the alignment span of the S subtree

Reordering with ITG Constraints

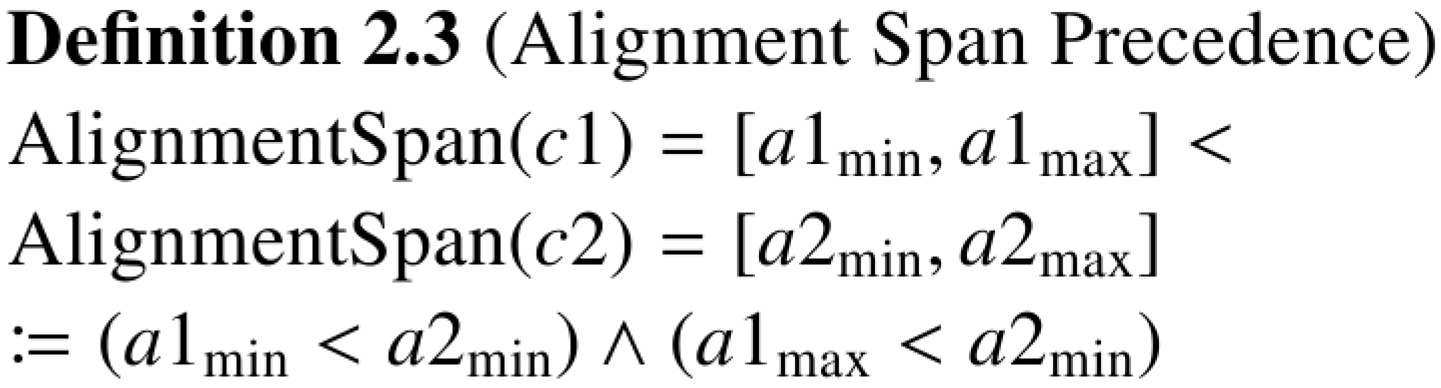
- Origin:
  Inversion Transduction Grammars (Wu, 1997):
  Bilingual Parsing

- Application   for Reordering:
  - Basic: child nodes binary tree may be inverted
  - General tree: permute child nodes arbitrarily
  - Restriction to constituency parse

Tree Constrained Reordering

- ***Monotonization:*** Reorder source to match target order

- Formal definition precedence used

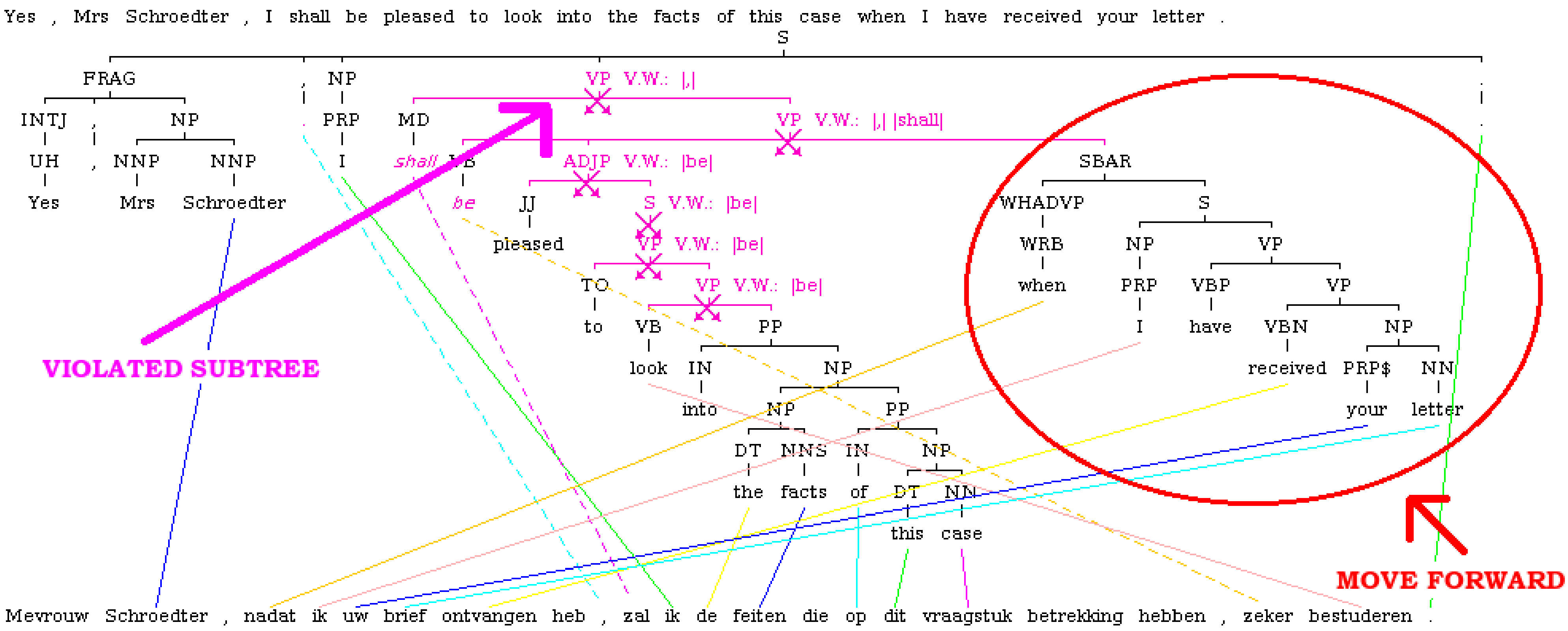


Definition 2.3 (Alignment Span Precedence)
$\text{AlignmentSpan}(c1) = [a1_{\min}, a1_{\max}] <$
$\text{AlignmentSpan}(c2) = [a2_{\min}, a2_{\max}]$
$:= (a1_{\min} < a2_{\min}) \wedge (a1_{\max} < a2_{\min})$

- All positions covered by *c1* strictly precede those covered by *c2*

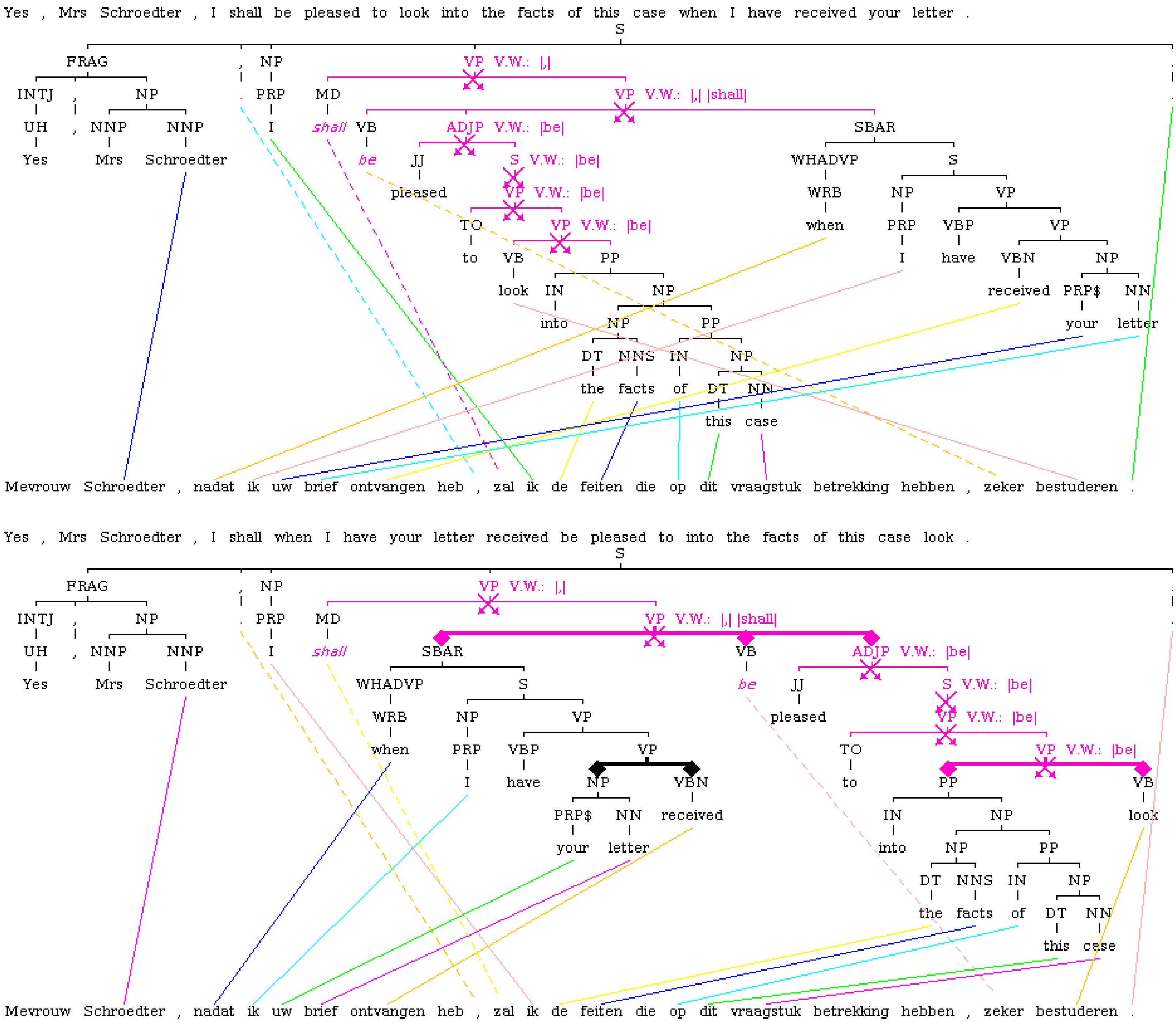
- ***Un-cohesiveness*** causes order problems

Tree-Constrained Reordering



- Right subtree moved forward, violated subtree not touched

Tree-Constrained Reordering



Yes , Mrs Schroedter , I shall be pleased to look into the facts of this case when I have received your letter .

Mevrouw Schroedter , nadat ik uw brief ontvangen heb , zal ik de feiten die op dit vraagstuk betrekking hebben , zeker bestuderen .

Yes , Mrs Schroedter , I shall when I have your letter received be pleased to into the facts of this case look .

Mevrouw Schroedter , nadat ik uw brief ontvangen heb , zal ik de feiten die op dit vraagstuk betrekking hebben , zeker bestuderen .

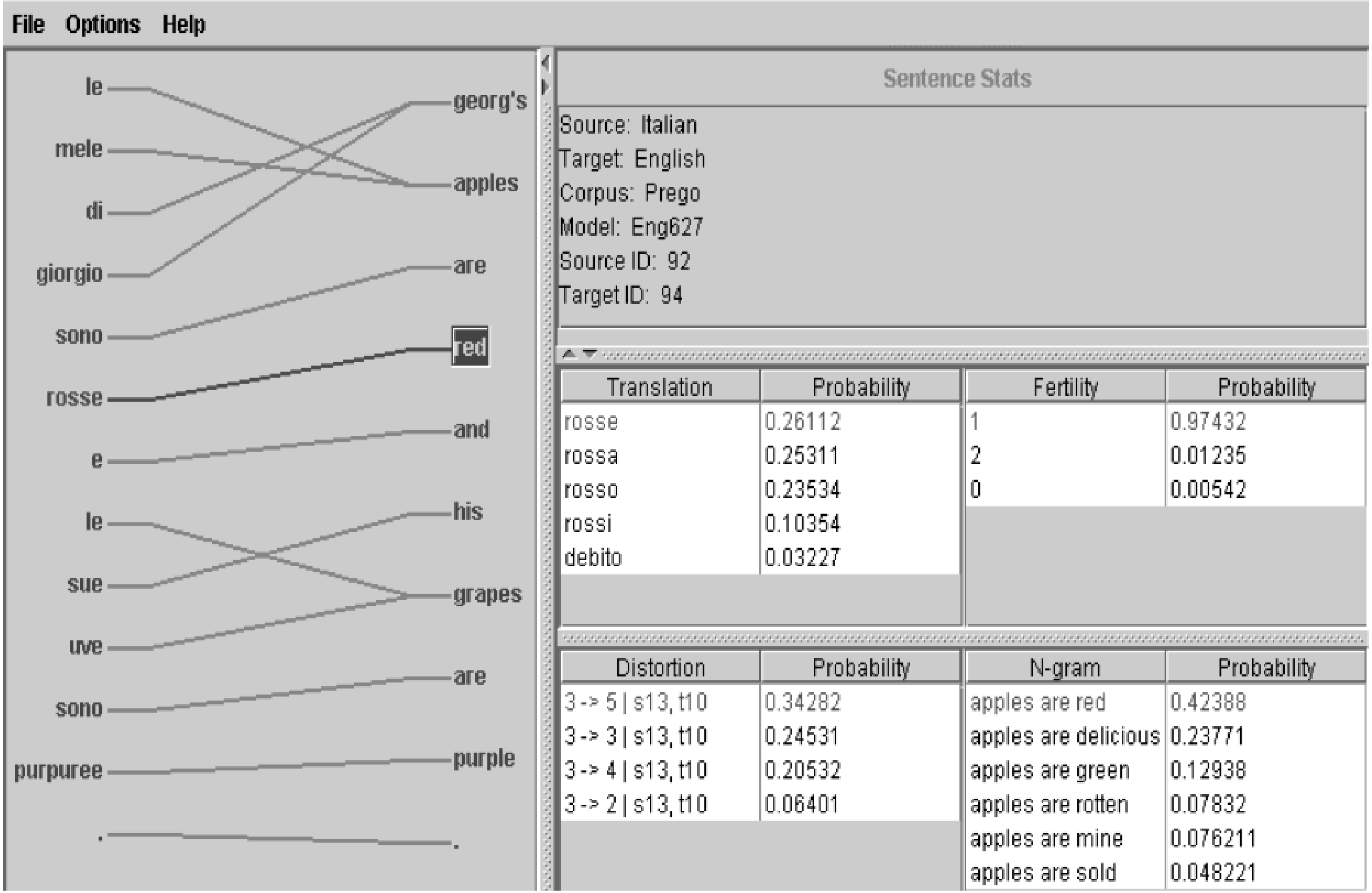
Usability

- Browsing through aligned sentences
- Insight into alignment mapping sub-trees
- Assess quality reordering tree-constrained ITG
- Get ideas for new tree-transduction operations

Related Work

- Visualization tools
  - *Cairo* (Smith and Jahr, 2000)
  - *Yawat* (Germann, 2008)
  - *Stockholm Tree Aligner (STA)* (Volk et al., 2007)

Cairo: IBM model visualization



Source: Italian
Target: English
Corpus: Prego
Model: Eng627
Source ID: 92
Target ID: 94

Translation	Probability
rosse	0.26112
rossa	0.25311
rosso	0.23534
rossi	0.10354
debito	0.03227

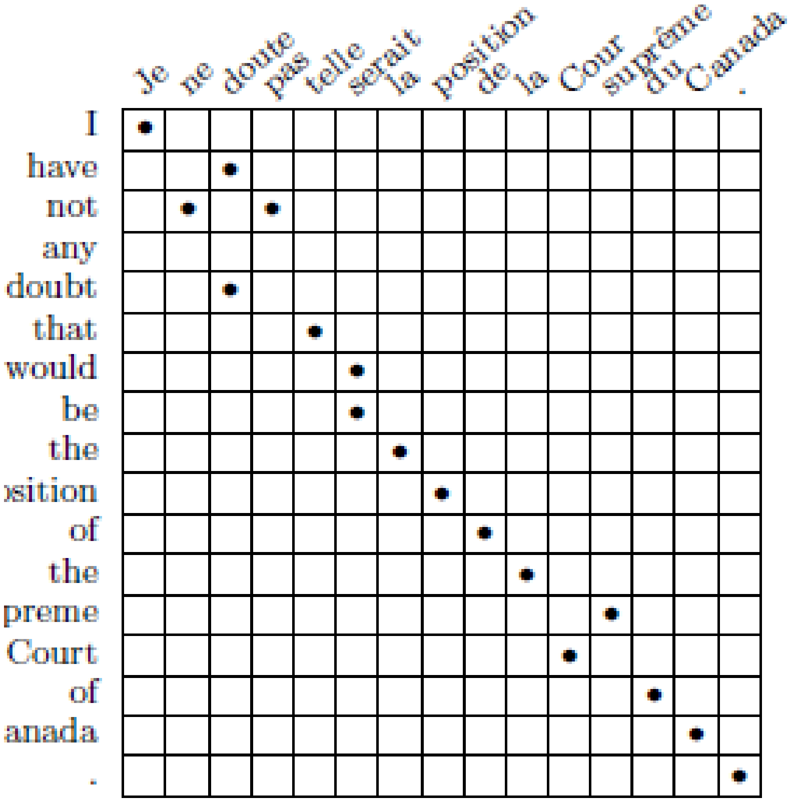
Fertility	Probability
1	0.97432
2	0.01235
0	0.00542

Distortion	Probability
3 -> 5 | s13, t10	0.34282
3 -> 3 | s13, t10	0.24531
3 -> 4 | s13, t10	0.20532
3 -> 2 | s13, t10	0.06401

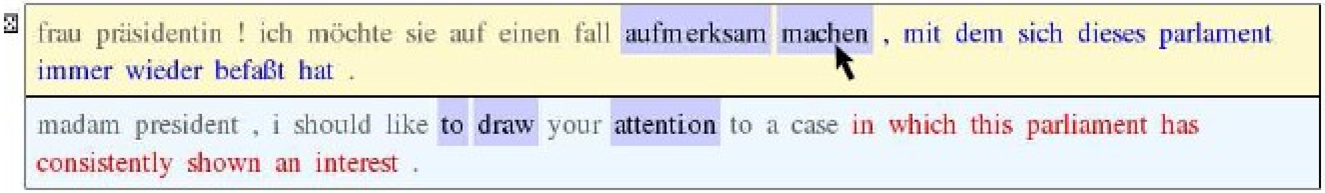
N-gram	Probability
apples are red	0.42388
apples are delicious	0.23771
apples are green	0.12938
apples are rotten	0.07832
apples are mine	0.076211
apples are sold	0.048221

Yawat: focus on clearity

- Visualization alignment matrix
- Support manual annotation
- Dynamic highlighting



	Je	ne	doute	pas	telle	serait	la	position	de	la	Cour	suprême	du	Canada	.
I	●														
have			●												
not		●		●											
any															
doubt			●												
that					●										
would						●									
be						●									
the							●								
osition								●							
of									●						
the										●					
preme												●			
Court											●				
of													●		
anada														●	
.															●

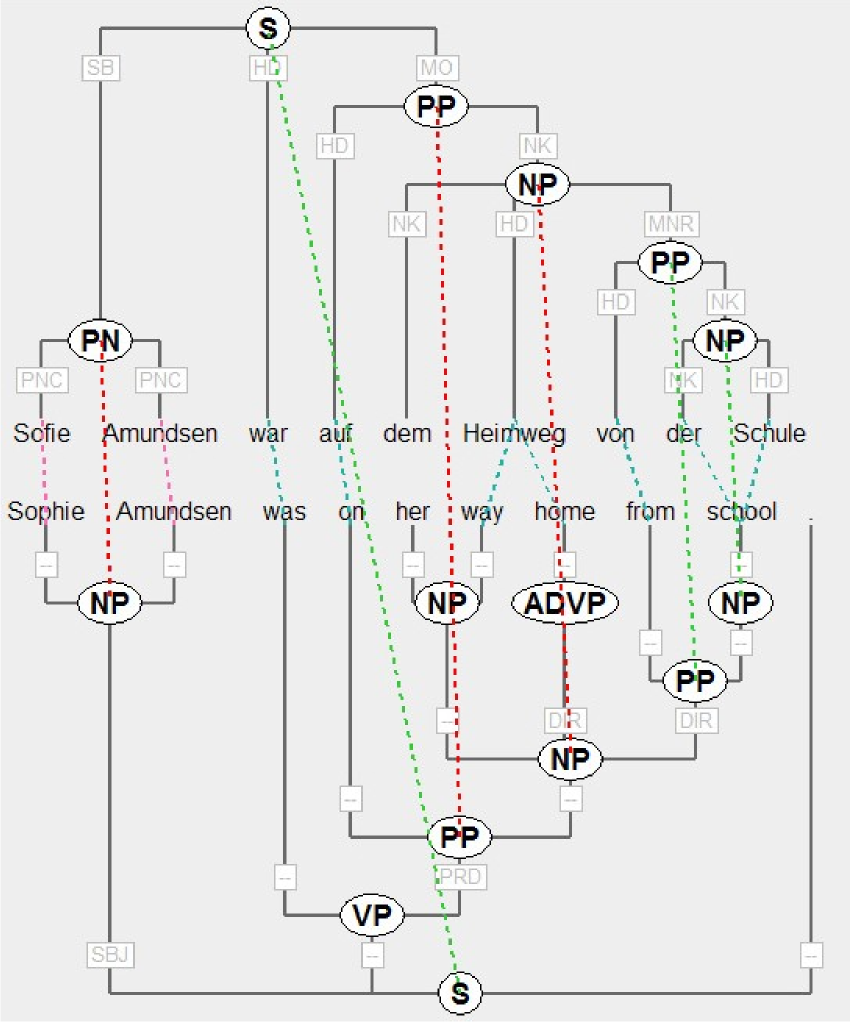


frau präsidentin ! ich möchte sie auf einen fall aufmerksam machen , mit dem sich dieses parlament immer wieder befaßt hat .

madam president , i should like to draw your attention to a case in which this parliament has consistently shown an interest .

Stockholm Tree Aligner (STA)

- Visualization parallel treebanks
- Focus on hand-annotated trees



Conclusions

- Toolkit targeted especially SMT people
- Focus on automatically generated resources and syntactic SMT
- Offers new functionality
- Goal: support SMT research

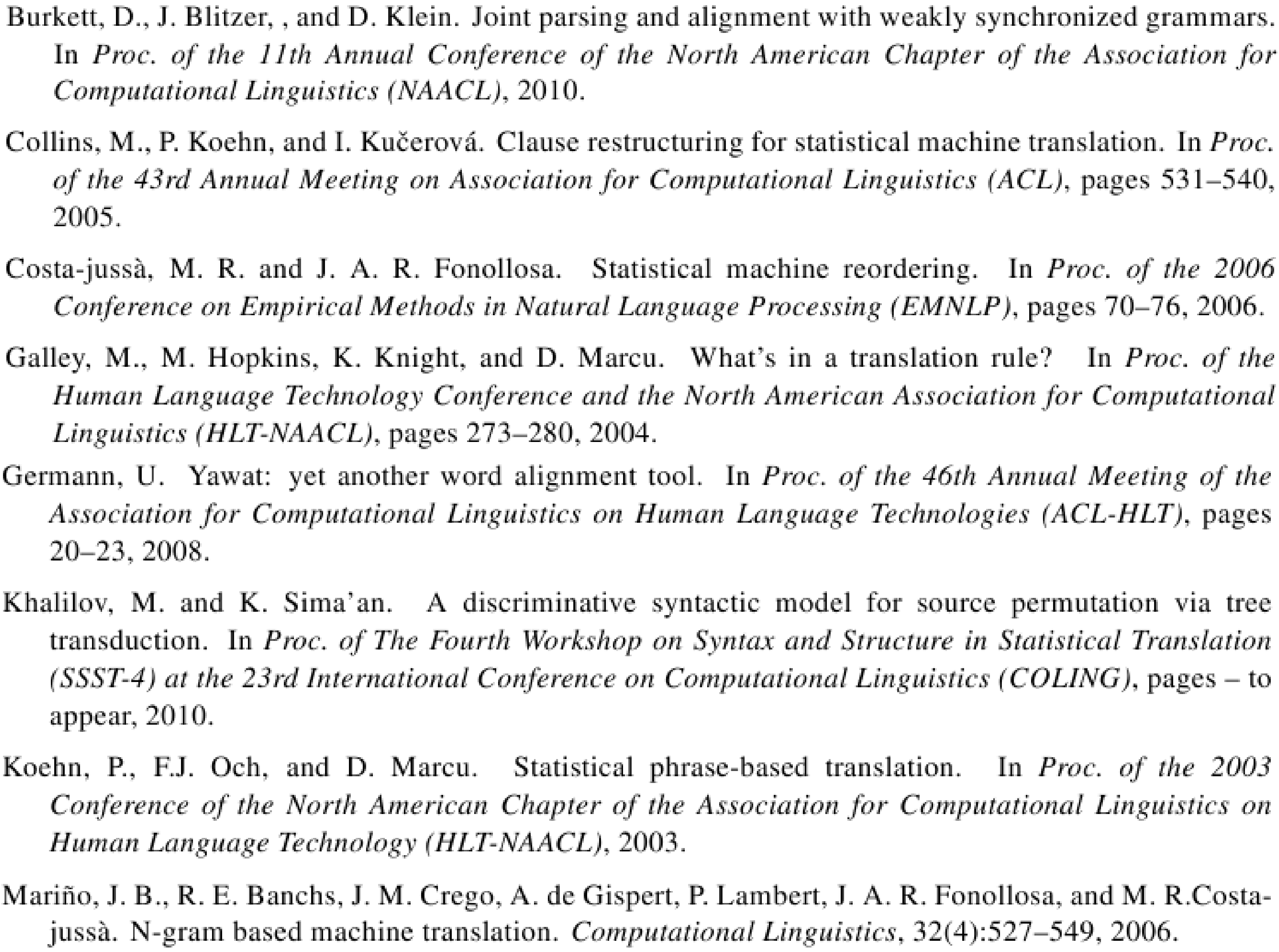
Future Work

- Heuristics in subtree reordering
- Tree modification
- Alignment refinement

Demo

Questions?

Bibliography(1/2)



Burkett, D., J. Blitzer, , and D. Klein. Joint parsing and alignment with weakly synchronized grammars. In *Proc. of the 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2010.

Collins, M., P. Koehn, and I. Kučerová. Clause restructuring for statistical machine translation. In *Proc. of the 43rd Annual Meeting on Association for Computational Linguistics (ACL)*, pages 531–540, 2005.

Costa-jussà, M. R. and J. A. R. Fonollosa. Statistical machine reordering. In *Proc. of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 70–76, 2006.

Galley, M., M. Hopkins, K. Knight, and D. Marcu. What's in a translation rule? In *Proc. of the Human Language Technology Conference and the North American Association for Computational Linguistics (HLT-NAACL)*, pages 273–280, 2004.

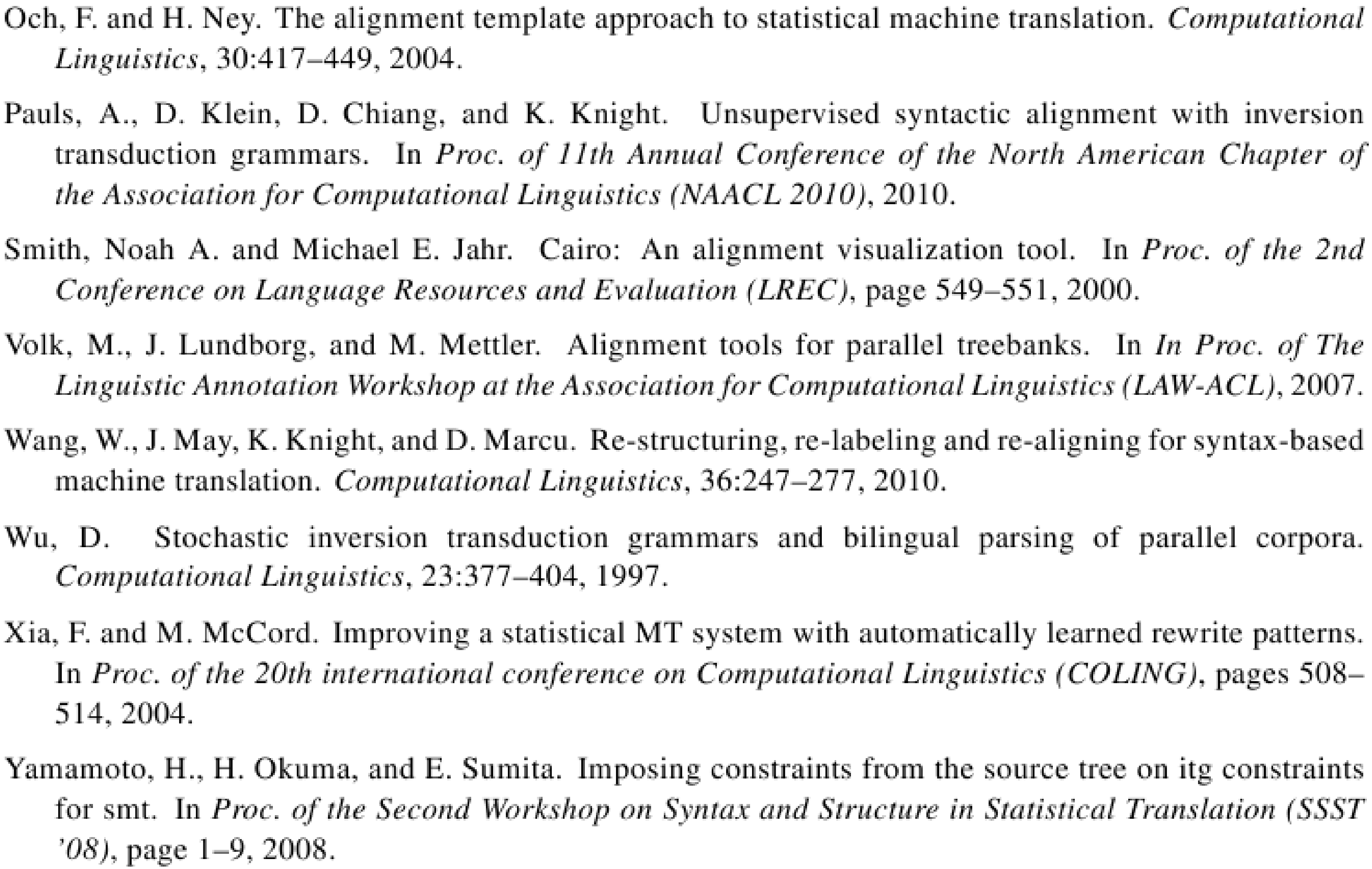
Germann, U. Yawat: yet another word alignment tool. In *Proc. of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies (ACL-HLT)*, pages 20–23, 2008.

Khalilov, M. and K. Sima'an. A discriminative syntactic model for source permutation via tree transduction. In *Proc. of The Fourth Workshop on Syntax and Structure in Statistical Translation (SSST-4) at the 23rd International Conference on Computational Linguistics (COLING)*, pages – to appear, 2010.

Koehn, P., F.J. Och, and D. Marcu. Statistical phrase-based translation. In *Proc. of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (HLT-NAACL)*, 2003.

Mariño, J. B., R. E. Banchs, J. M. Crego, A. de Gispert, P. Lambert, J. A. R. Fonollosa, and M. R.Costa-jussà. N-gram based machine translation. *Computational Linguistics*, 32(4):527–549, 2006.

Bibliography(2/2)



Och, F. and H. Ney. The alignment template approach to statistical machine translation. *Computational Linguistics*, 30:417–449, 2004.

Pauls, A., D. Klein, D. Chiang, and K. Knight. Unsupervised syntactic alignment with inversion transduction grammars. In *Proc. of 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2010)*, 2010.

Smith, Noah A. and Michael E. Jahr. Cairo: An alignment visualization tool. In *Proc. of the 2nd Conference on Language Resources and Evaluation (LREC)*, page 549–551, 2000.

Volk, M., J. Lundborg, and M. Mettler. Alignment tools for parallel treebanks. In *In Proc. of The Linguistic Annotation Workshop at the Association for Computational Linguistics (LAW-ACL)*, 2007.

Wang, W., J. May, K. Knight, and D. Marcu. Re-structuring, re-labeling and re-aligning for syntax-based machine translation. *Computational Linguistics*, 36:247–277, 2010.

Wu, D. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23:377–404, 1997.

Xia, F. and M. McCord. Improving a statistical MT system with automatically learned rewrite patterns. In *Proc. of the 20th international conference on Computational Linguistics (COLING)*, pages 508–514, 2004.

Yamamoto, H., H. Okuma, and E. Sumita. Imposing constraints from the source tree on itg constraints for smt. In *Proc. of the Second Workshop on Syntax and Structure in Statistical Translation (SSST '08)*, page 1–9, 2008.