Moses

Machine Translation with Open Source Software

Marcin Junczys-Dowmunt and Hieu Hoang

2 July 2014

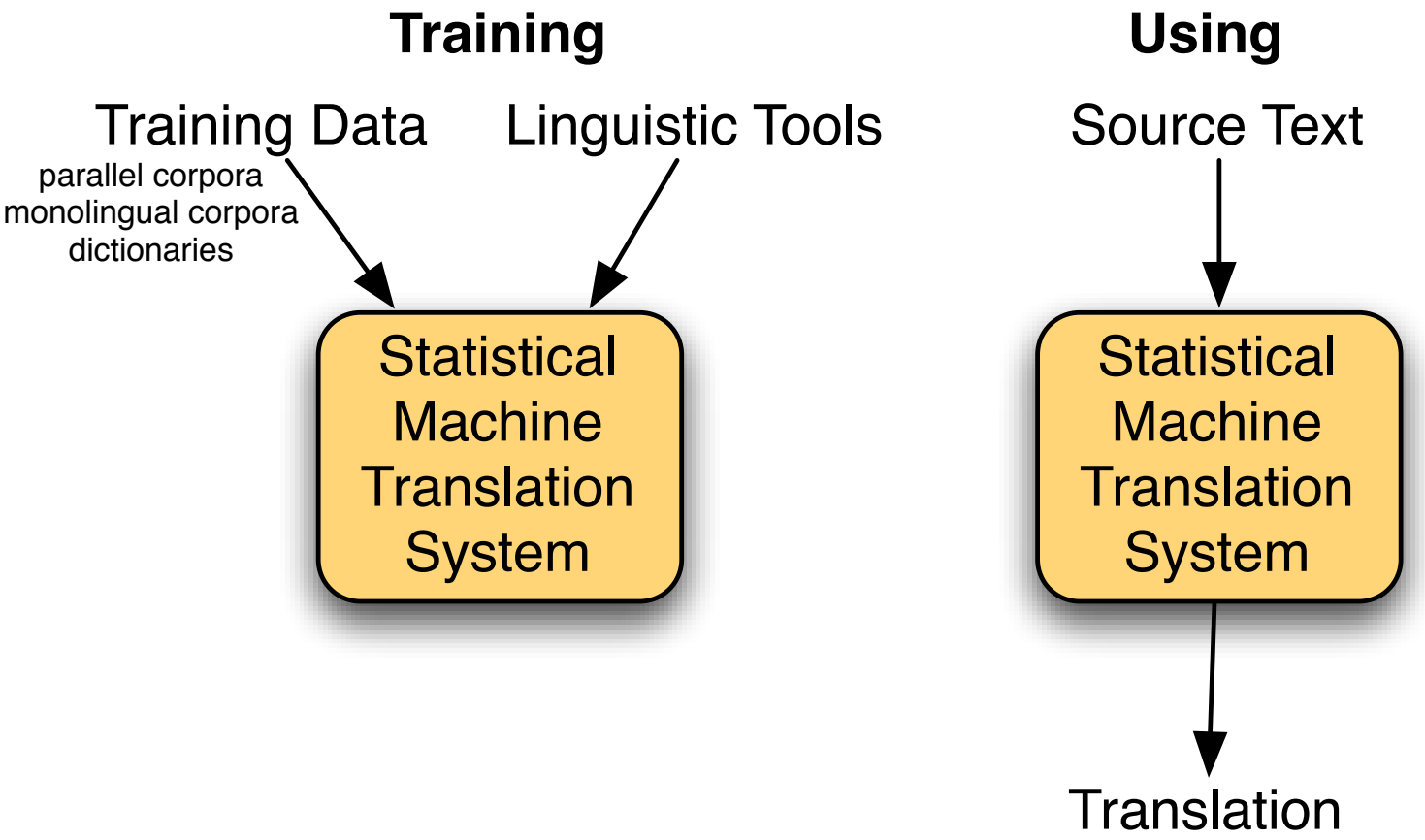
Outline

09:30-10:00 **Introduction**

11:00-11:30 Break

11:30-12:30 **Advanced Topics**

Basic Idea


Training
Training Data
parallel corpora
monolingual corpora
dictionaries
Linguistic Tools
Statistical Machine Translation System
Using
Source Text
Statistical Machine Translation System
Translation

Statistical Machine Translation History

around 1990

Pioneering work at IBM, inspired by success in speech recognition

1990s

Dominance of IBM's word-based models, support technologies

early 2000s

Phrase-based models

late 2000s

Tree-based models

Moses History

2002 Pharaoh decoder, precursor to Moses (phrase-based models)

2005 Moses started by Hieu Hoang and Philipp Koehn (factored models)

2006 JHU workshop extends Moses significantly

2006-2012 Funding by EU projects EuroMatrix, EuroMatrixPlus

2009 Tree-based models implemented in Moses

2012-2015 MosesCore project. Full-time staff to maintain and enhance Moses

Moses in Academia

- Built by academics, for academics
- Reference implementation of state of the art
  - researchers develop new methods on top of Moses
  - developers re-implement published methods
  - used by other researchers as black box
- Baseline to beat
  - researchers compare their method against Moses

Moses in Industry

- Distributed with LGPL — free to use
- Competitive with commercial SMT solutions
  (Language Weaver, Google, ...)
- But:
  - not easy to use
  - requires significant expertise for optimal performance
  - integration into existing workflow not straight-forward

Case Studies

European Commission —
uses Moses in-house to aid human translators

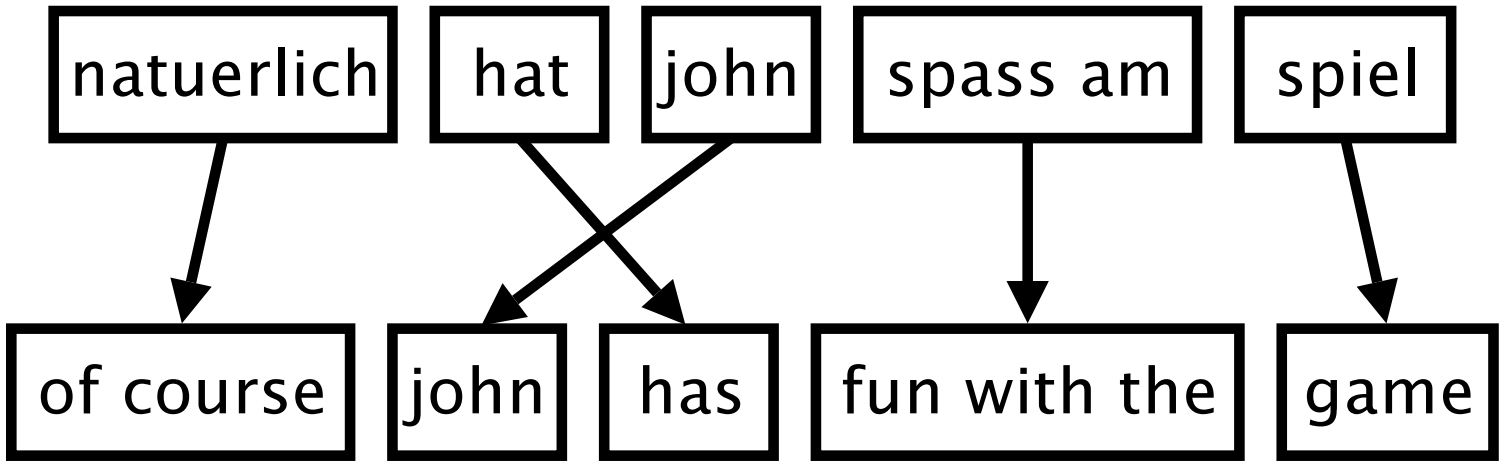
Autodesk —
showed productivity increases in translating manuals when post-editing output from a custom-build Moses system

Systran —
developed statistical post-editing using Moses

Asia Online —
offers translation technology and services based on Moses

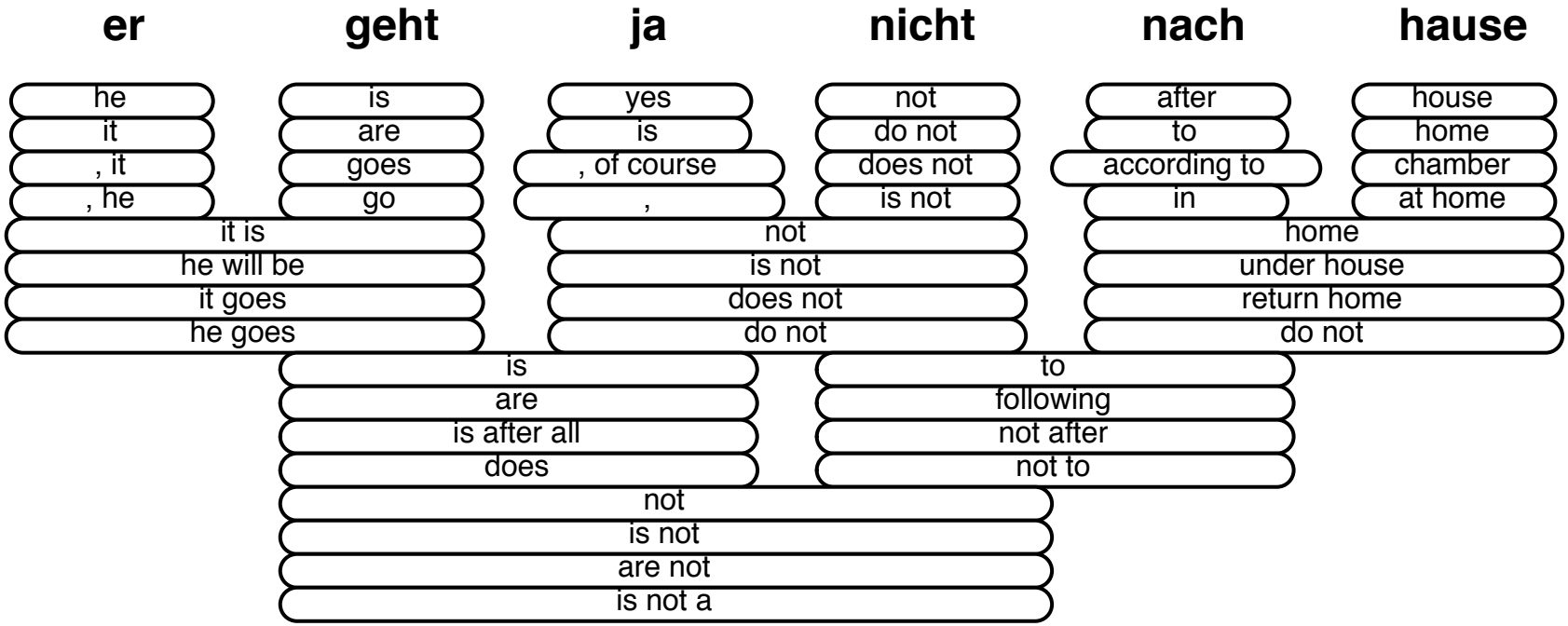
Many others ...
World Trade Organisation, Adobe, Symantec, WIPO, Sybase, Safaba, Bloomberg, Pangeanic, KatanMT, Capita, ...

Phrase-Based Model



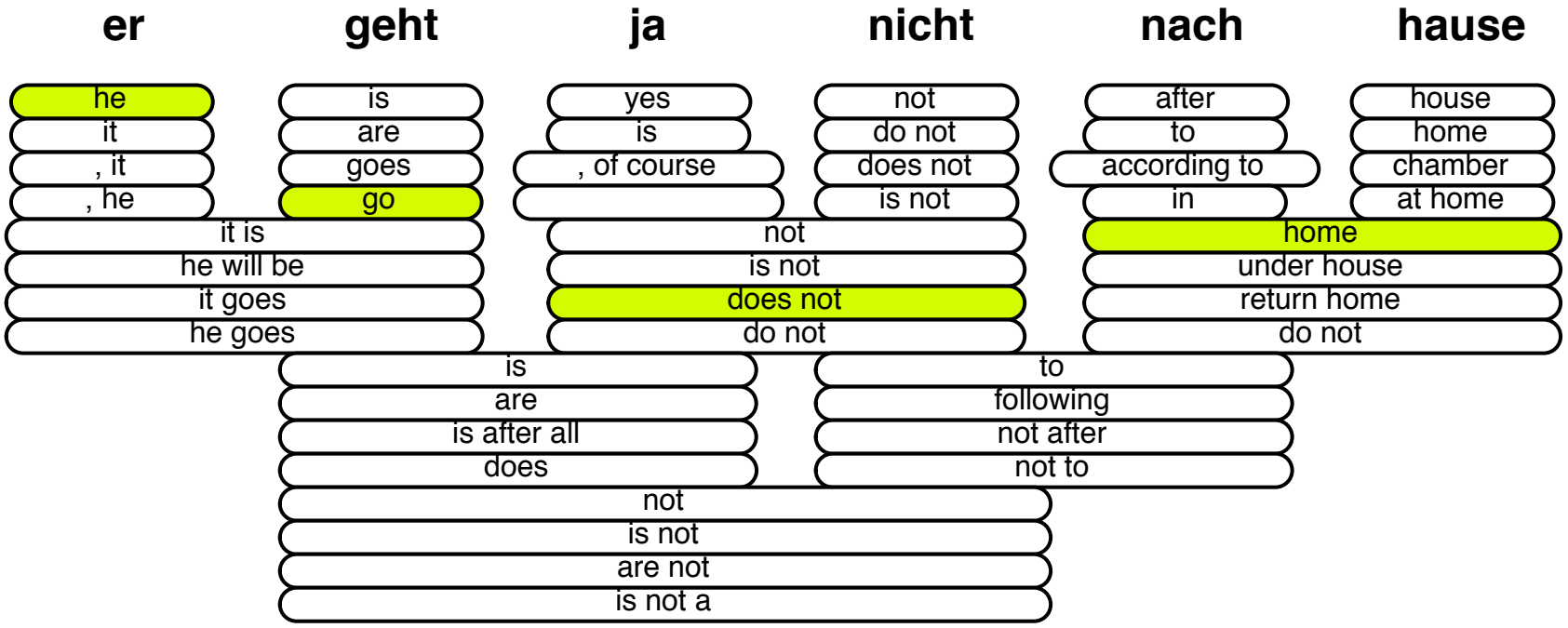
- Foreign input is segmented in phrases
- Each phrase is translated into English
- Phrases are reordered

Phrase Translation Options



- Many translation options to choose from

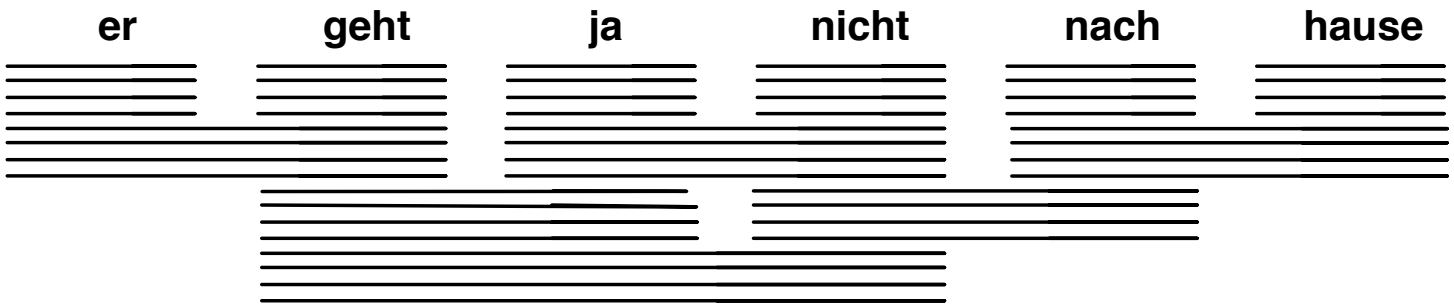
Phrase Translation Options



- The machine translation decoder does not know the right answer
  - picking the right translation options
  - arranging them in the right order

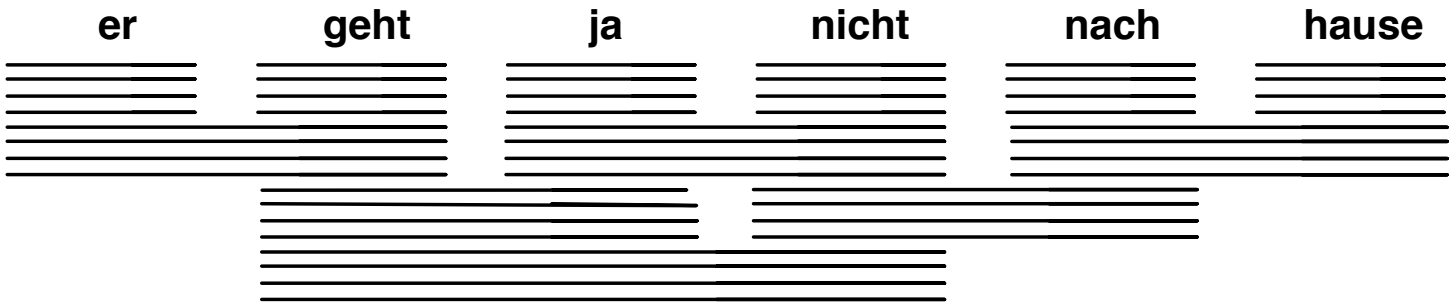
→ Search problem solved by heuristic beam search

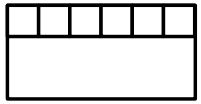
Decoding: Precompute Translation Options



consult phrase translation table for all input phrases

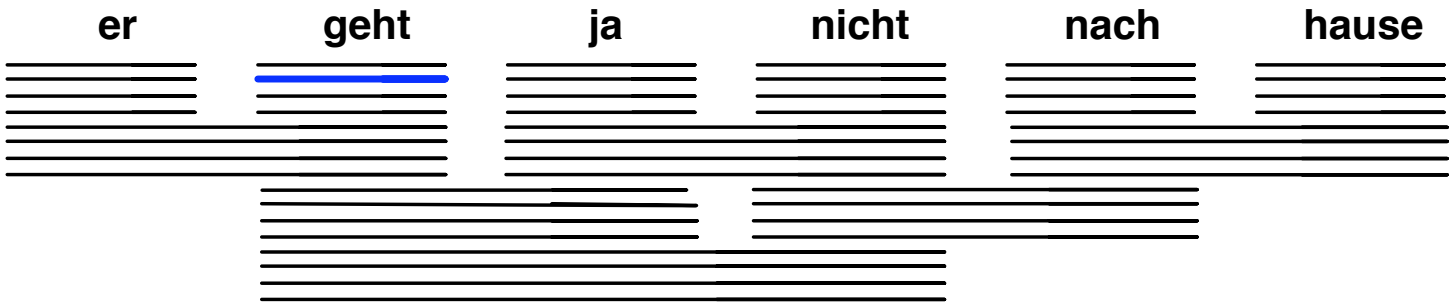
Decoding: Start with Initial Hypothesis

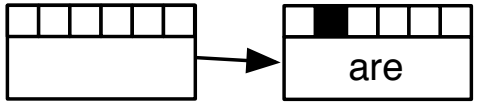




initial hypothesis: no input words covered, no output produced

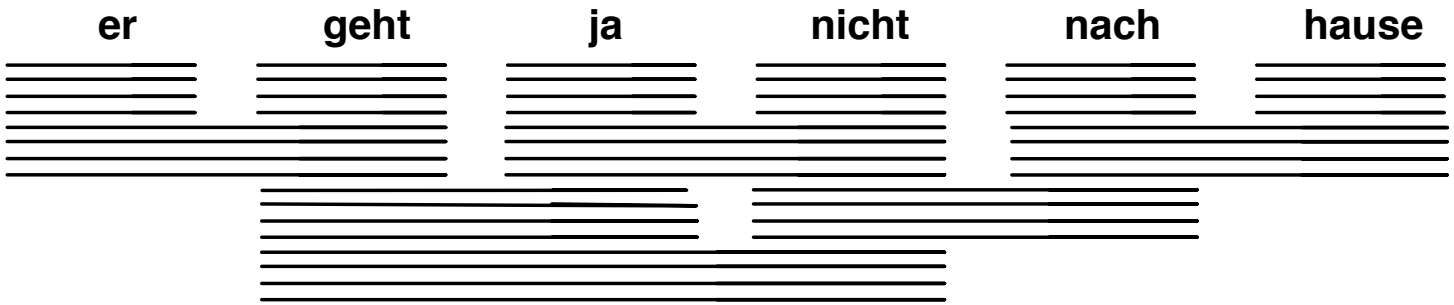
Decoding: Hypothesis Expansion

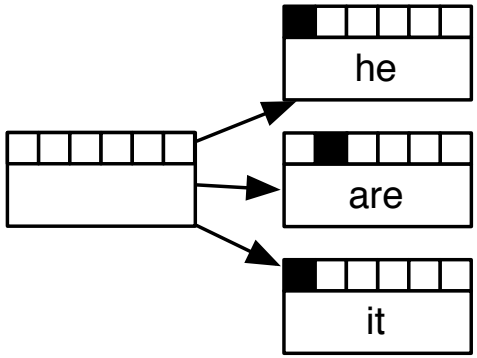




pick any translation option, create new hypothesis

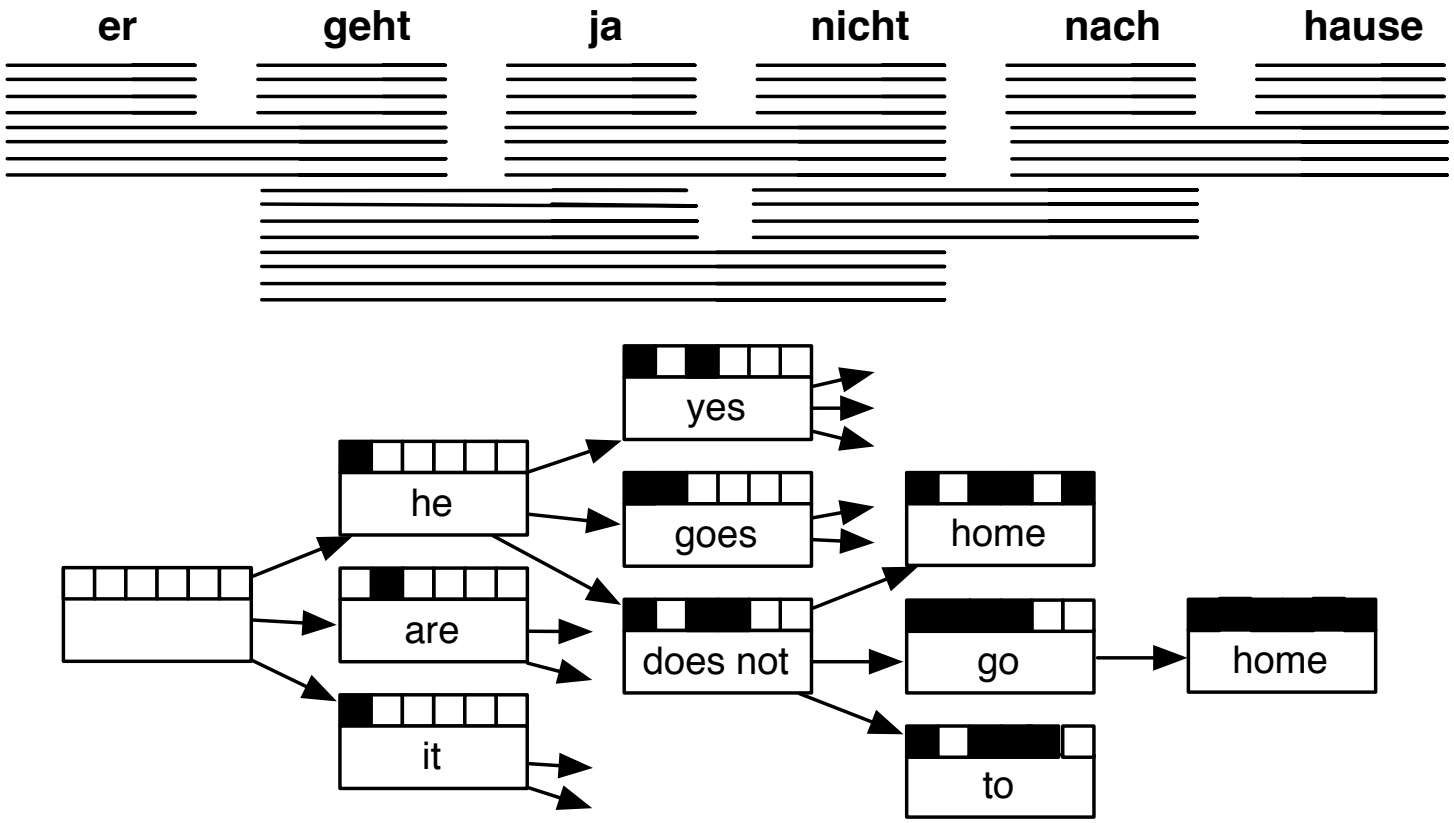
Decoding: Hypothesis Expansion





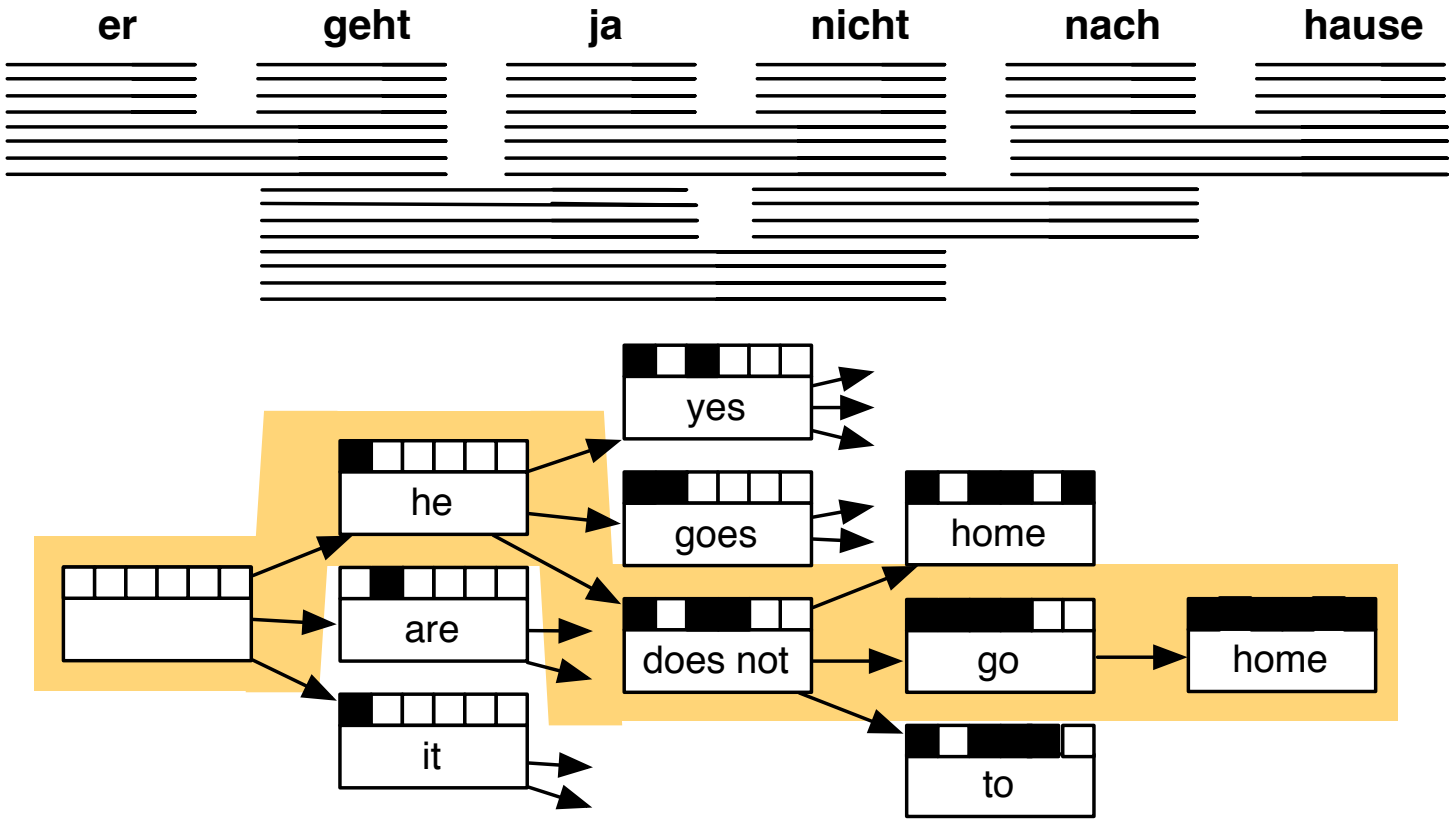
create hypotheses for all other translation options

Decoding: Hypothesis Expansion



also create hypotheses from created partial hypothesis

Decoding: Find Best Path



backtrack from highest scoring complete hypothesis

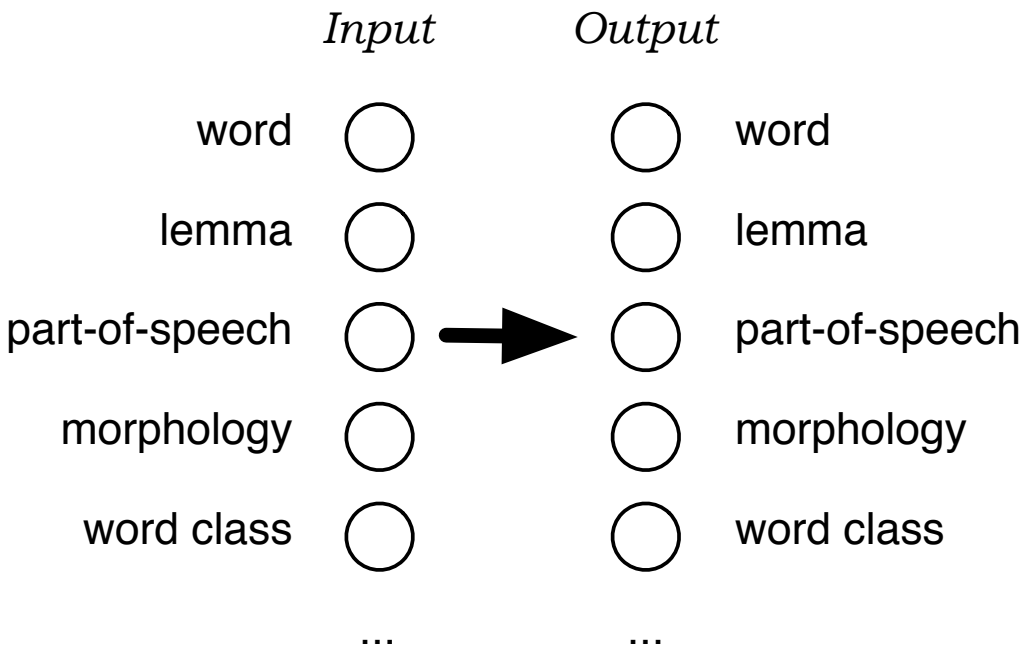
Computational Complexity

- The suggested process creates exponential number of hypothesis
- Reduction of search space: pruning

$\rightarrow$ Decoder may not find the model-best translation

Factored Represention

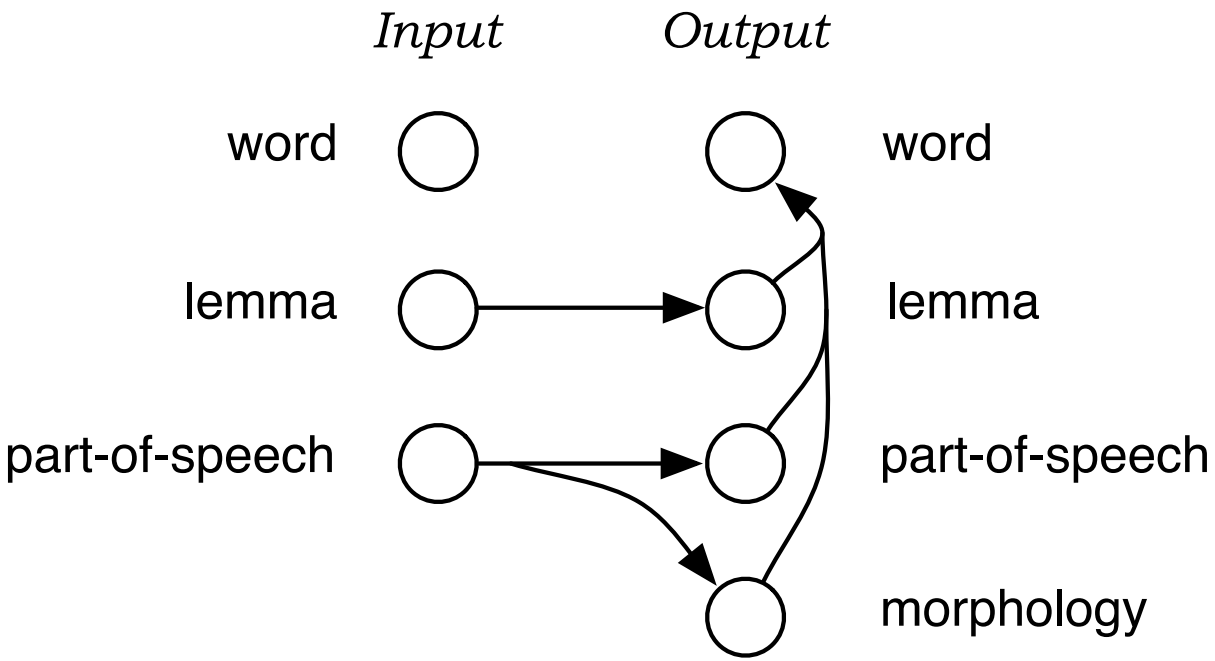
- Factored represention of words



- Goals
  - generalization, e.g. by translating lemmas, not surface forms
  - richer model, e.g. using syntax for reordering, language modeling)

Factored Model

Example:



Decomposing the translation step
Translating lemma and morphological information more robust

Syntax Models

String to String

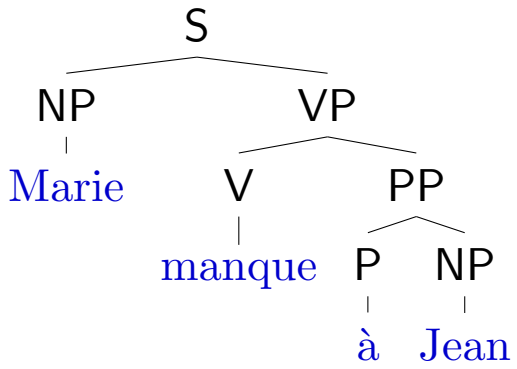
John misses Mary

⇒ Marie manque à Jean

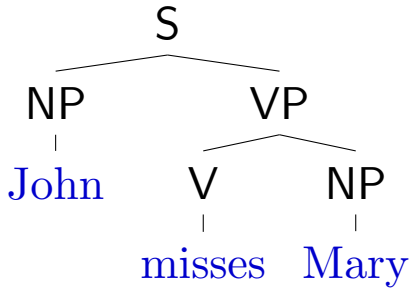
String to Tree

John misses Mary

⇒

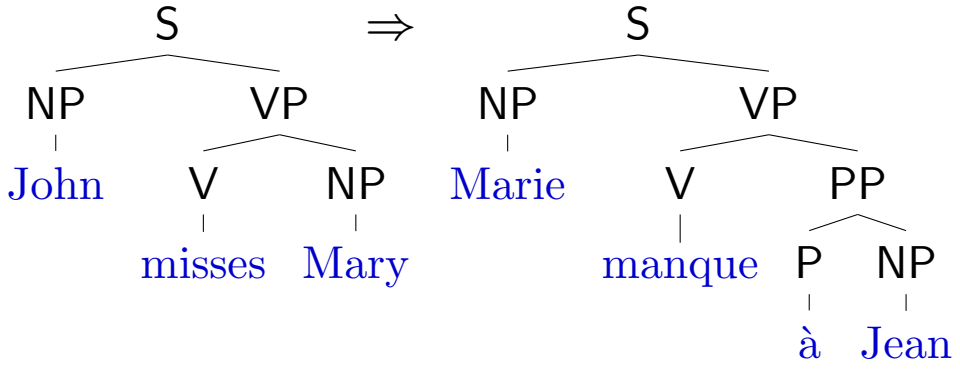


Tree to String

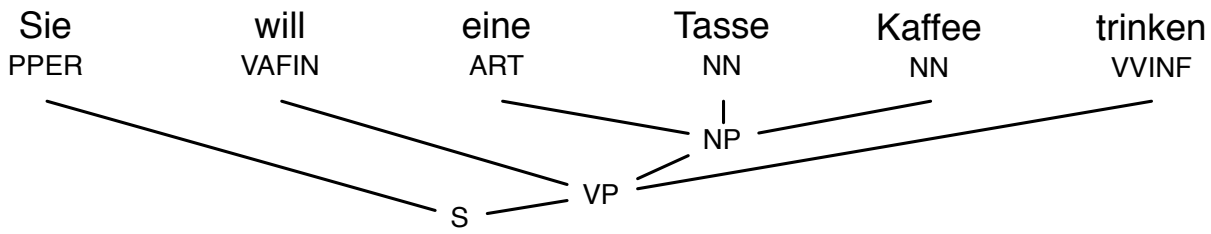


⇒ Marie manque à Jean

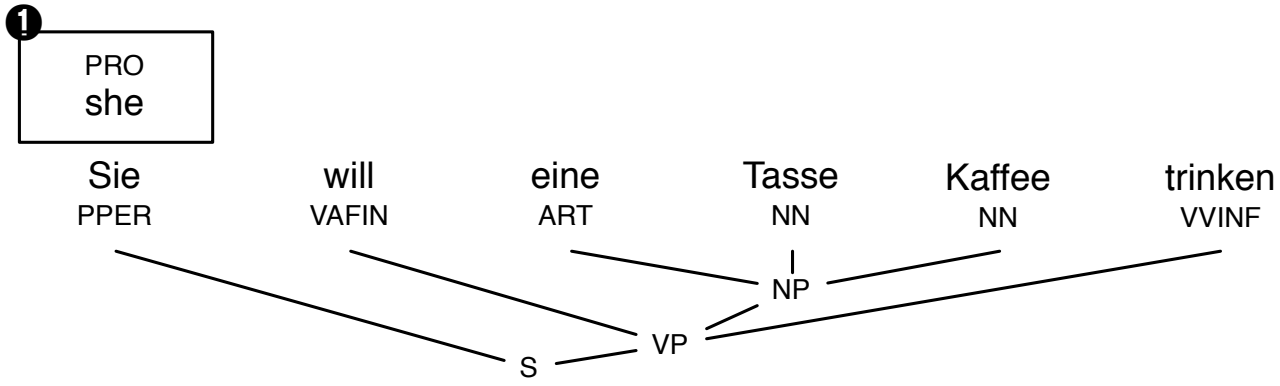
Tree to Tree



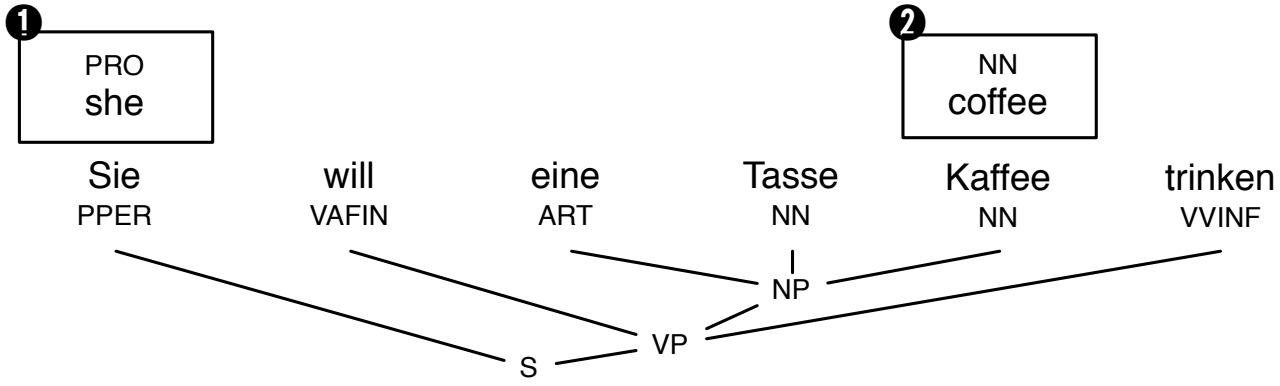
Syntax Decoding



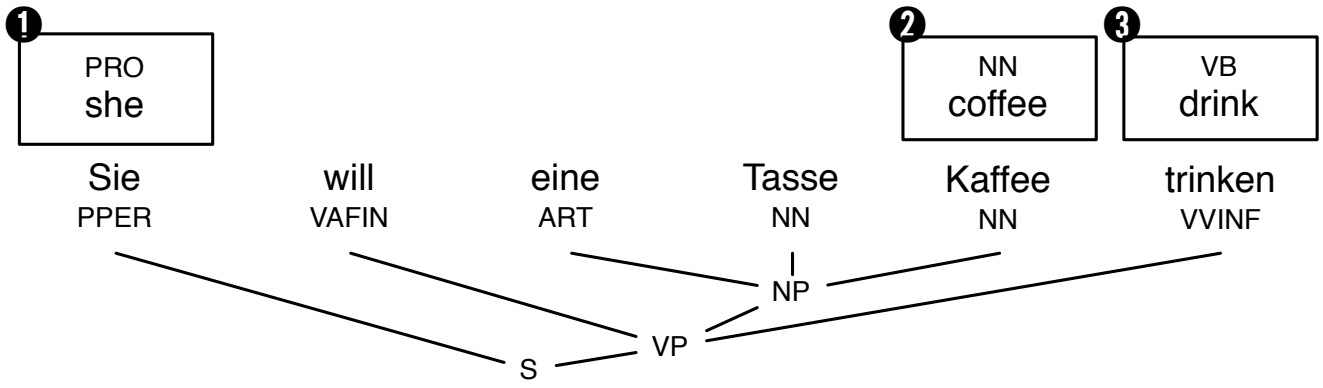
Syntax Decoding



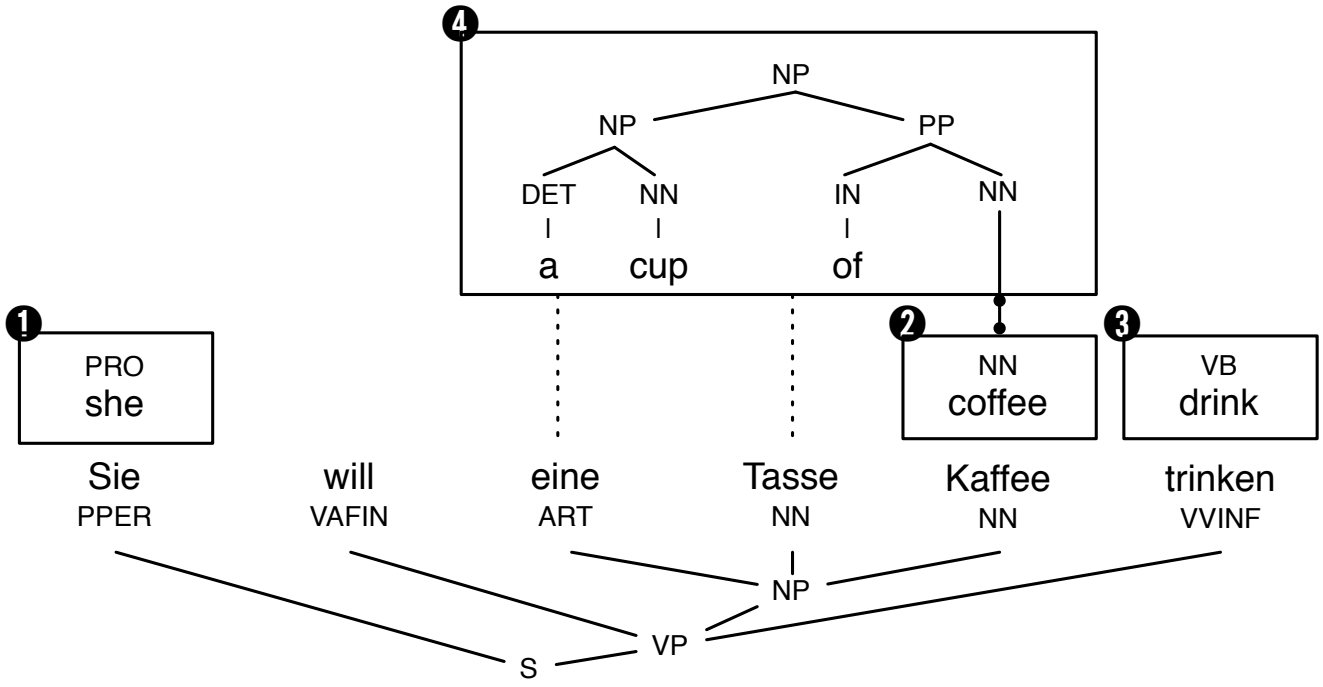
Syntax Decoding



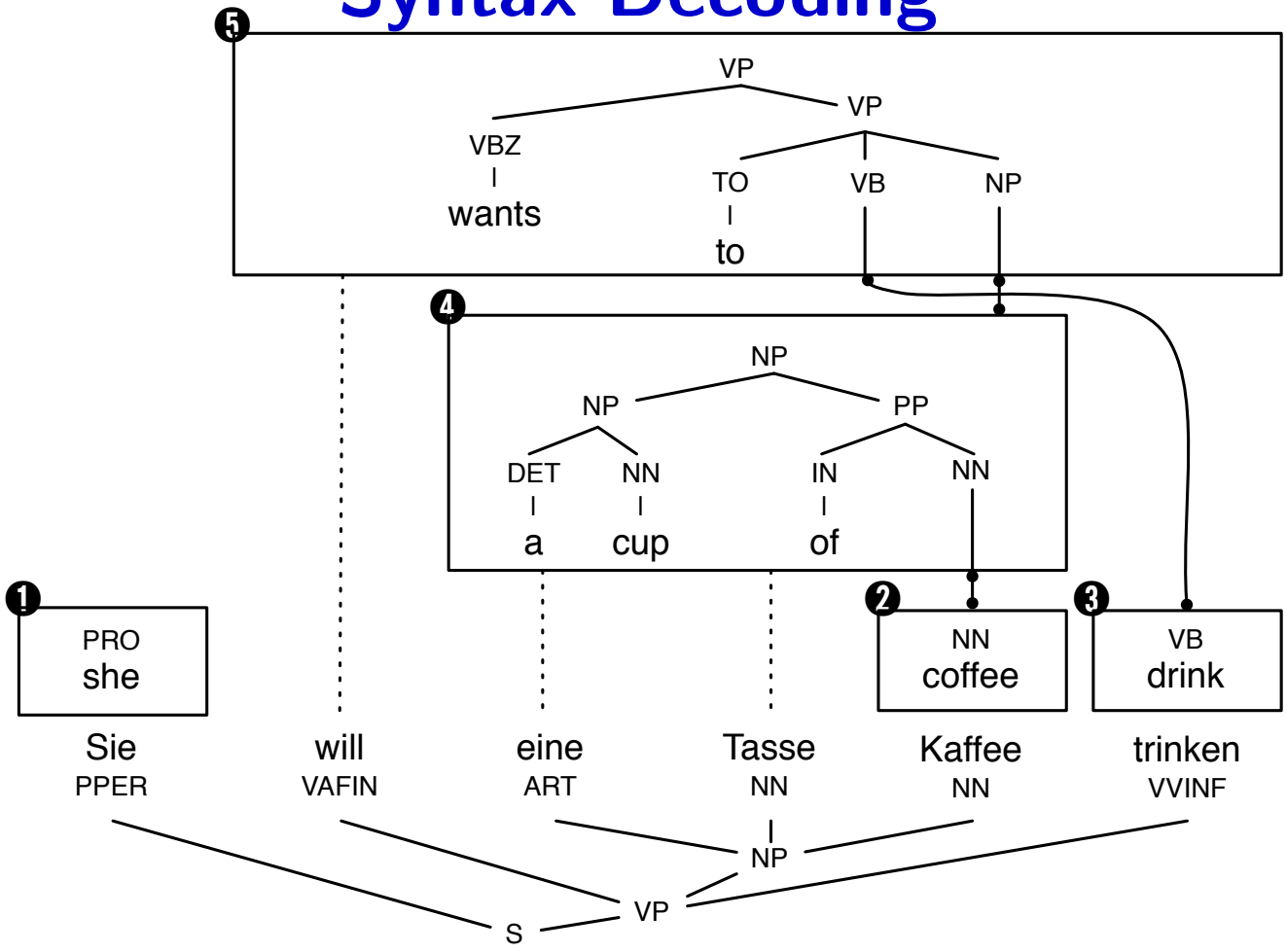
Syntax Decoding

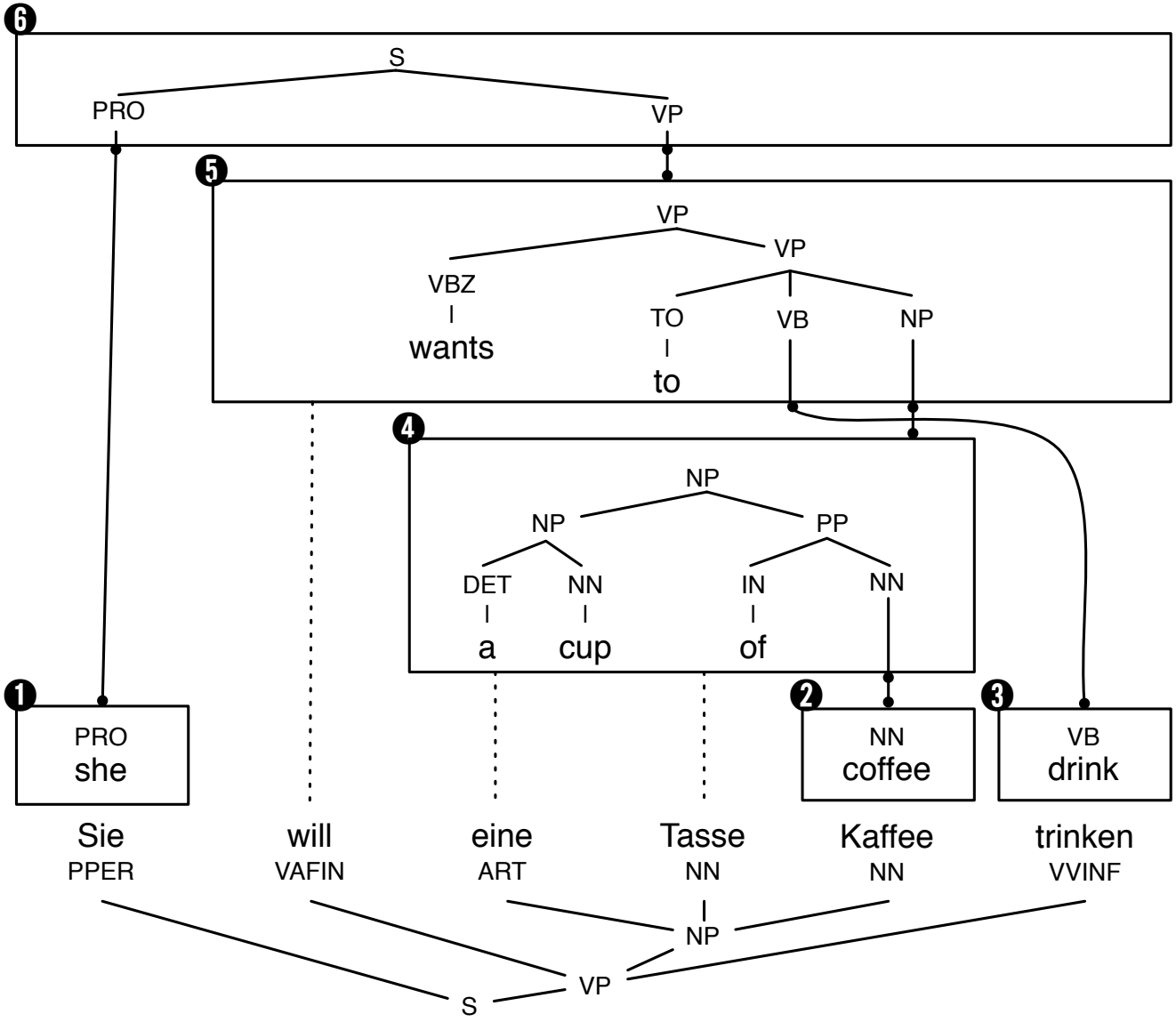


Syntax Decoding



Syntax Decoding




6
S
PRO
VP
5
VP
VP
VBZ
wants
TO
to
VB
NP
4
NP
NP
PP
DET
a
NN
cup
IN
of
NN
1
PRO
she
2
NN
coffee
3
VB
drink
Sie
PPER
will
VAFIN
eine
ART
Tasse
NN
Kaffee
NN
trinken
VVINF
NP
VP
S

Advanced Features

Advanced Features

- **How do I get started?**
- Experiment Management System
- Faster Decoding
- Data and domain adaptation
- Instructions to decoder
- Input formats
- Output formats

How do I get started?

- Collect your data
  - Parallel data
  - Translation memories
  - Open-sourced data, eg. Europarl, UN, TAUS Data Association
  - Monolingual data
- Set up Moses
  - Download source code for Moses, GIZA++, MGIZA
  - Compile, install
  - More info: http://www.statmt.org/moses/
  - Prepackaged Moses: Precision Tools, MacPorts, Debian packages, M4Loc

How do I get started?

Execute a lot of scripts

```
tokenize < corpus.en > corpus.en.tok
lowercase < corpus.en.tok > corpus.en.lc
...
mert.perl ....
moses ...
mteval-v13.pl ...
```

Change a part of the process, execute everything again

```
tokenize < corpus.en > corpus.en.tok
lowercase < corpus.en.tok > corpus.en.lc
...
mert.perl ....
moses ...
mteval-v13.pl ...
```

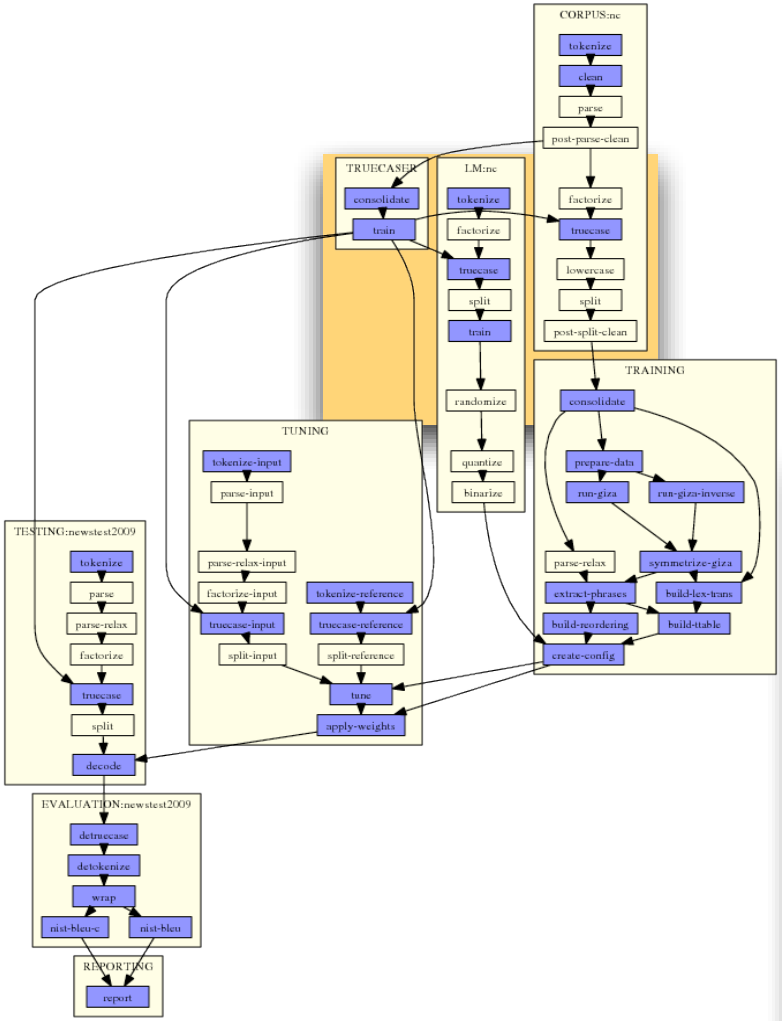
Advanced Features

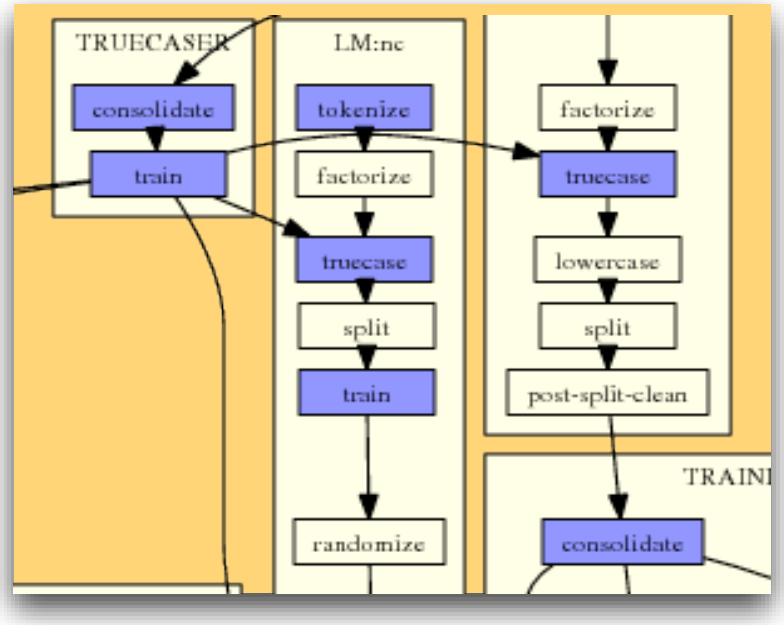
- How do I get started?
- **Experiment Management System**
- Faster Decoding
- Data and domain adaptation
- Instructions to decoder
- Input formats
- Output formats

Experiment Management System

- One configuration file for all settings: record of all experimental details
- Scheduler of individual steps in pipeline
  - automatically keeps track of dependencies
  - runs on single machine, multi-core machine, GridEngine cluster
  - parallel execution
  - crash detection
  - automatic re-use of prior results
- Fast to use
  - set up a new experiment in minutes
  - set up a variation of an experiment in seconds
- Disadvantage — not all Moses features are integrated

Workflow automatically generated by experiment.perl





How does it work?

- Write a configuration file (typically by adapting an existing file)
- Test:

```
experiment.perl -config config
```

- Execute:

```
experiment.perl -config config -exec
```

Web Interface

All Experimental Setups

ID	User	Task	Directory
97	pkoehn	Acquis Truecased	/group/project/statmt2/pkoehn/acquis-truecase
96	pkoehn	Chinese-English AGILE 2008	/group/project/statmt2/pkoehn/agile08-chinese
95	miles	Randlm testing	/group/project/statmt7/miles/experiments /ep-enfr/work
94	joseph	Proj2008 Impl.Adapted experiment(fr-en)for News Comm.	/group/project/statmt2/joseph/experimentJo/task6
93	joseph	Proj2008 Impl.Baseline experiment(fr-en)for News Comm.	/group/project/statmt2/joseph/experimentJo/task5
92	jschroe1	FR-EN System Combination Components	/group/project/statmt9/josh/experiments /fr-syscomb/work

List of experiments

List of Runs

Task: WMT10 German-English (pkoehn)

Wiki Notes | Overview of experiments | /fs/bragi2/pkoehn-experiment/wmt10-de-en

compare	ID	start	end	avg	newstest2009		newstest2010	
cfg|par|img	[1042-16] 11+analysis	16 May	16 May	BLEU-c: 21.74 BLEU: 22.91	21.03 (1.002) 22.30 (1.002)	Ⓐ	22.45 (1.041) 23.51 (1.041)	Ⓐ
cfg|par|img	[1042-15] 11+Internal emplus test set	21 Apr	crashed	-	-		-	
cfg|par|img	[1042-14] 9+interpolated-tm.lm-weighted	21 Feb	21 Feb 9: 0.239258 -> 0.239296	-	20.81 (1.003) 22.06 (1.003)	Ⓐ	-	
cfg|par|img	[1042-13] 9+only-ep	21 Feb	21 Feb 13: 0.235046 -> 0.235053	-	20.42 (1.002) 21.69 (1.002)	Ⓐ	-	
cfg|par|img	[1042-12] 9+only-nc	21 Feb	21 Feb 7: 0.222237 ->	-	18.96 (1.002) 20.16	Ⓐ	-	

Analysis: Basic Statistics

Coverage

	model	corpus
0	2047 (3.1%)	1708 (2.6%)
1	738 (1.1%)	518 (0.8%)
2-5	1483 (2.2%)	818 (1.2%)
6+	61745 (93.5%)	62969 (95.4%)

by token / by type / details

Phrase Segmentation

	1	2	3	4+
1 to	26897 (40.7%)	2145 (3.2%)	278 (0.4%)	90 (0.1%)
2 to	4144 (6.3%)	14414 (21.8%)	2518 (3.8%)	432 (0.7%)
3 to	639 (1.0%)	3522 (5.3%)	4821 (7.3%)	1272 (1.9%)
4+ to	158 (0.2%)	855 (1.3%)	1693 (2.6%)	2135 (3.2%)

by word / by phrase

- Basic statistics
  - n-gram precision
  - evaluation metrics
  - coverage of the input in corpus and translation model
  - phrase segmentations used

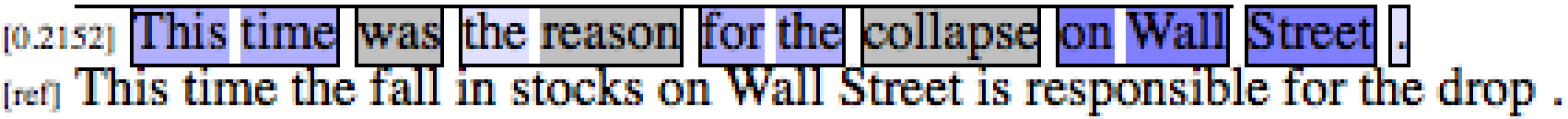
Analysis: Unknown Words

grouped by count in test set

unknown words

18 Eatonville	**4:** Eatonvilles, Együtt, Garver, Harmadik, Hurstons, Jobb, Jol, Jos, Jövőért, Kovalev, Krever, Lados, Mercandelli, Stehplätze, Tauro, Tórtola, Zenobia, fon, Évezredért, Ózd	**3:** Anmil, Atlasz, BR23C, BSA, Bayón, Biztos, Bt., Butch, Casado, Dal, Embraer, FT, Faymann, Fiatal, Gregg, Gélineau, HSV, Hanzelka, Illhäusern, Iván, Jansen, Jančura, Joanne, Kemrová, Kid, Llamazares, Loafs, Mangas, Medikamentes, Mobil.cz, Mutual,	**2:** Abfertigungen, Albums, Alondra, Andoh, Anm., Armiñon, Ashford, BZÖ, Baloldal, Bani, Baugesellschaften, Bedienkomfort, Bento, Bentos, Bingleys, Bojen, Bowens, Bowery, Boyd, Bringley, Browser, Bělohlávek, CBGB, Carci, Cera, Charts, Chemical, Chigi, Cineast, Comics, Commerzbank, Coppola, Corker, Cowon, DF, Dinkins, Download, Drehbewegung, Drzewiecki, Drápal, Düsseldorfer, Ella,	**1:** -Ach, -Minister, -Pakets, -weiss, .docx, .pptx, .xlsx, 1,45, 1.106,55, 1.983,73, 10.365,45, 10.579, 10.809,25, 106,85, 11,9, 11.743,61, 12.595.75, 14,2, 14,7, 145.29, 16,8, 17.9, 18,6, 18.286,90, 1802, 1834, 1880ern, 1920ern, 1925, 19252008, 199,61, 2,178, 2,37, 2.400, 26,3, 270.000, 29,2, 3,30, 3,632, 3,827, 3.0.0, 4,161, 4,357, 42,2, 43,4, 499, 49sten, 5.839, 506,43, 6,98, 684,81, 729,700, 75,5, 777,68, 8,25, 8,81, 9,14, 99.80, AAC, ADQ, ART, Aareal, Abbremsens, Abhöraktion, Absenzen, Abwesenheiten, Abwiegen, Abwärtssog, Achronot, Actor, AdSense, AdWords, Aday, Adobe, Adressverzeichnisses, Adwards, Adélard, Agazio, Akku, Akron, Aktuálně.cz, Alameda, Alatriste, Alcolock, Aleš, Alhambra, Alleinregierer, Amazonengebiet, Amil, Aminei, Amministrazione, Amway, Andalusierin, Andik, Android, Anděl, Angeklagtem, Ansa, Anthologie, Antiasthmatika, Apnoe, Aquel, Arabija, Arbeiternehmers, Arcandor, Arriaga, Asiana, Askale, Astronomen, Aufeislegen, Augäpfel, Ausdrückstärke, Ausführungs-, Ausgeruhter, Ausscheidungsspiele,
16 Hurston				
12 Barrick				
12 Hema				
12 Stewards				
11 Gebrselassie				
10 Flamenco				
10 Mango				
9 Glitter				
9 ÚOHS				
9 ČTÚ				
8 Coles				
8 Deka				
8 Garci				
8 ITV				

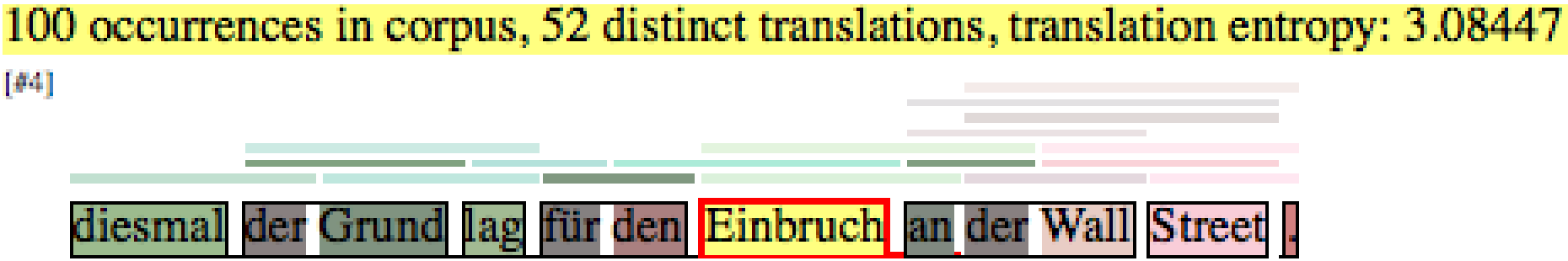
Analysis: Output Annotation



Color highlighting to indicate n-gram overlap with reference translation

darker bleu = word is part of larger n-gram match

Analysis: Input Annotation



- For each word and phrase, color coding and stats on
  - number of occurrences in training corpus
  - number of distinct translations in translation model
  - entropy of conditional translation probability distribution $\phi(e|f)$ (normalized)

Analysis: Bilingual Concordancer

entre autres(560/1554)

...d and made recommendations , " **inter alia** " , with respect to the follow...	...**entre autres** , les questions spécifiques suiva...
...on (EC) No 1995 / 2000 imposing , **inter alia** , a definitive anti @-@ dumping dut...	...995 / 2000 du Conseil instituant , **entre autres** , un droit antidumping définitif ...
...ervices . this increase , arising , **inter alia** , as a result of economic growth , ...	...nsports . cette augmentation , due **entre autres** facteurs à la croissance économi...
...of paragraph 1 the Commission may , **inter alia** , bring forward :	...aragraphe 1 , la Commission peut , **entre autres** , présenter :
... of stocks of obsolete pesticides , **inter alia** , by supporting projects aimed at s...	...r les stocks de vieux pesticides , **entre autres** en soutenant des projets à cet ef...
...wn rules of procedure which shall , **inter alia** , contain provisions for convening ...	...lement intérieur , qui contient , **entre autres** dispositions , les modalités de c...
...uch specific agreements may cover , **inter alia** , financing provisions , assignment...	...ords spécifiques peuvent porter , **entre autres** , sur les mécanismes financiers s...
...he internal market and concerning , **inter alia** , health and environmental protecti...	...hé intérieur et qui concernent , **entre autres** , la santé et la protection de l&...
...e product concerned) originating , **inter alia** , in Belarus and Russia (the count...	...it concerné ") originaire , **entre autres** , du Belarus et de Russie (ci @-@...
...e product concerned) originating , **inter alia** , in India .	...t concerné ") originaires , **entre autres** , de l ' Inde .

notamment(447/1554)

... the EU budget by addressing " **inter alia** " the problems of accountabili...	...get de l' Union , ce qui passe **notamment** par la résolution du problème de r...
...ates , the Commission has adopted , **inter alia** , Decision 2003 / 526 / EC (3) wh...	...es États membres , la Commission a **notamment** arrêté la décision 2003 / 526 / C...
...d equitable development involving , **inter alia** , access to productive resources , ...	... durable et équitable , impliquant **notamment** l' accès aux ressources produc...
...ertain products which could be used **inter alia** , as equipment on board ships but w...	...usceptibles d' être utilisés **notamment** comme équipements mis à bord , mai...
...nexes , taking into consideration , **inter alia** , available scientific , technical ...	...ion et à ses annexes , compte tenu **notamment** des informations scientifiques , tec...
...w that it is absolutely necessary , **inter alia** , because of enlargement , to find ...	...os; il est absolument nécessaire , **notamment** en raison de l' élargissement ...
...paragraphs 1 and 2 as appropriate , **inter alia** , by conducting studies and compili...	...ragraphes 1 et 2 le cas échéant , **notamment** en menant des études et en compilan...
...liability and efficiency , caused , **inter alia** , by insufficient technical and adm...	... et d' efficacité en raison , **notamment** , d' une interopérabilité tec...
...in the Programme shall be pursued , **inter alia** , by the following means :	...nis dans le programme , il convient **notamment** de mettre en oeuvre les moyens ci @-...

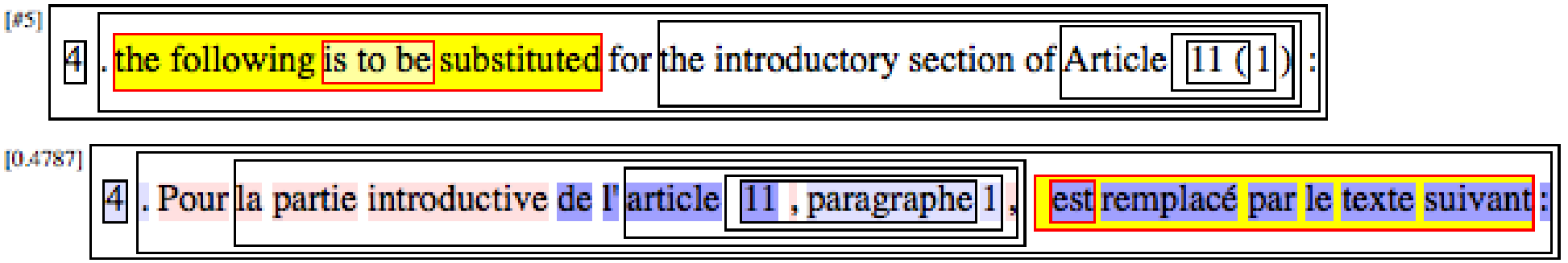
translation of input phrase in training data context

Analysis: Alignment

Phrase alignment of the decoding process

(red border, interactive)

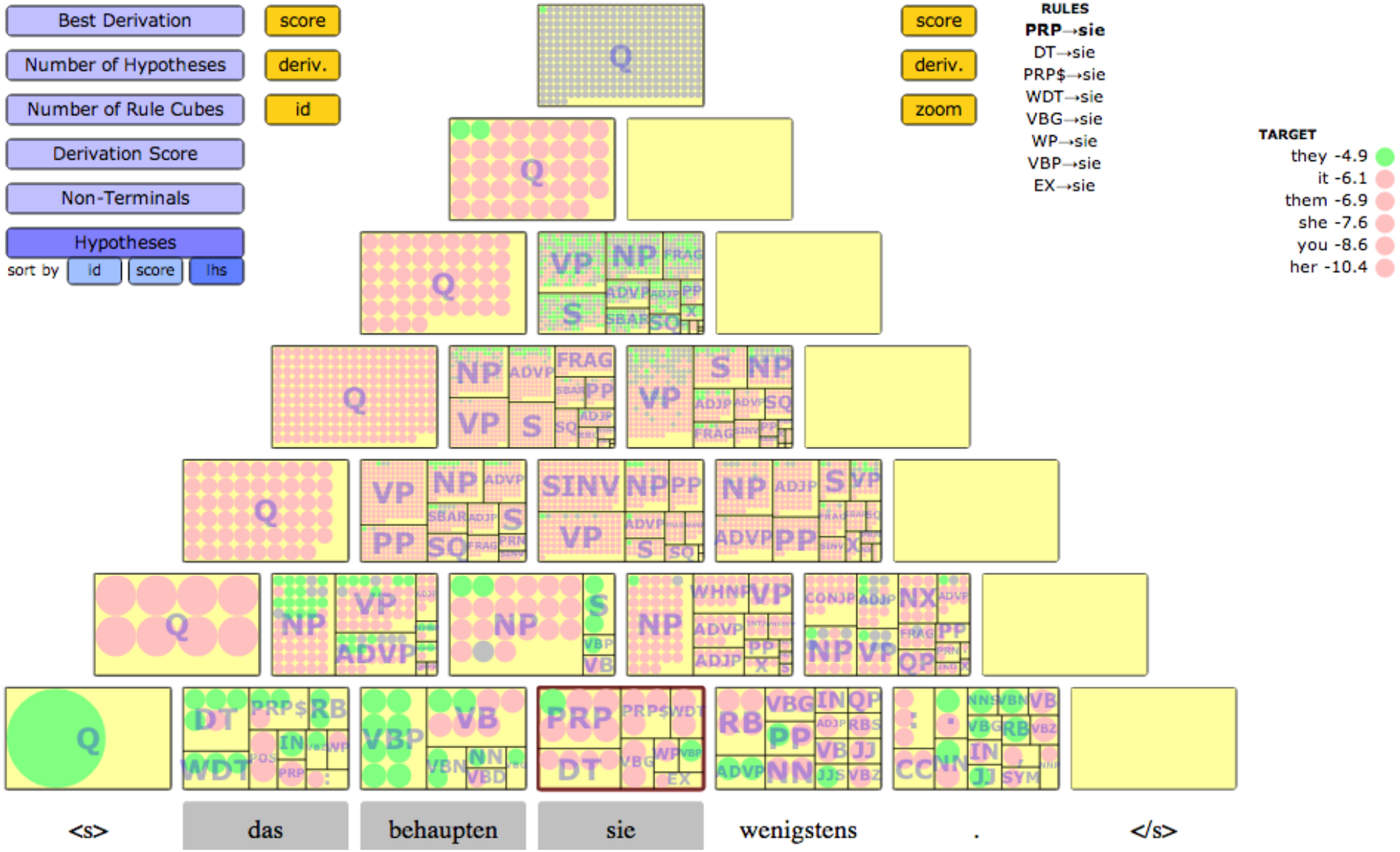
Analysis: Tree Alignment



Uses nested boxes to indicate tree structure

(red border, yellow shaded spans in focus, interactive)

for syntax model, non-terminals are also shown


Best Derivation
Number of Hypotheses
Number of Rule Cubes
Derivation Score
Non-Terminals
Hypotheses
sort by id score lhs
score
deriv.
id
score
deriv.
zoom
RULES
PRP→sie
DT→sie
PRP$→sie
WDT→sie
VBG→sie
WP→sie
VBP→sie
EX→sie
TARGET
they -4.9
it -6.1
them -6.9
she -7.6
you -8.6
her -10.4
<s>
das
behaupten
sie
wenigstens
.
</s>

Analysis: Comparison of 2 Runs

annotated sentences

sorted by order order worse display fullscreen showing 5 more all

identical	same	better	worse
2348	51	57	69
93%	2%	2%	3%

[2143:0.2974] In Austria , Haider and Co. are ready to govern to prevent a red and black coalition .
[2143:0.1754] In Austria , Haider and Co. are prepared to rule to prevent a red and black coalition .
[ref] Haider and his party are ready to govern Austria in order to avoid red @-@ black coalition .

[2165:0.3174] The SPÖ wants to show that the cooperation of both parties is possible - in some countries and in the social partnership that is already the case .
[2165:0.2061] The SPÖ wants to show that a cooperation of both parties is possible - in some countries and in the social partnership that is already the case .
[ref] SPÖ would like to show that the cooperation of the two parties is possible - it does exist in some of the provinces as well as in social partnership .

Different words are highlighted

sortable by most improvement, deterioration

Advanced Features

- How do I get started?
- Experiment Management System
- Faster Decoding
- Data and domain adaptation
- Instructions to decoder
- Input formats
- Output formats

Advanced Features

- How do I get started?
- Experiment Management System
- **Faster Decoding**
- Data and domain adaptation
- Instructions to decoder
- Input formats
- Output formats

Advanced Features

- How do I get started?
- Experiment Management System
- **Faster Decoding**
  - Speed vs. Memory
  - Speed vs. Quality

...

Advanced Features

- How do I get started?
- Experiment Management System
- Faster Decoding
  - Speed vs. Memory
  - Speed vs. Quality

...

Advanced Features

- How do I get started?
- Experiment Management System
- Faster Training
- Faster Decoding
  - **Speed vs. Memory**
  - Speed vs. Quality

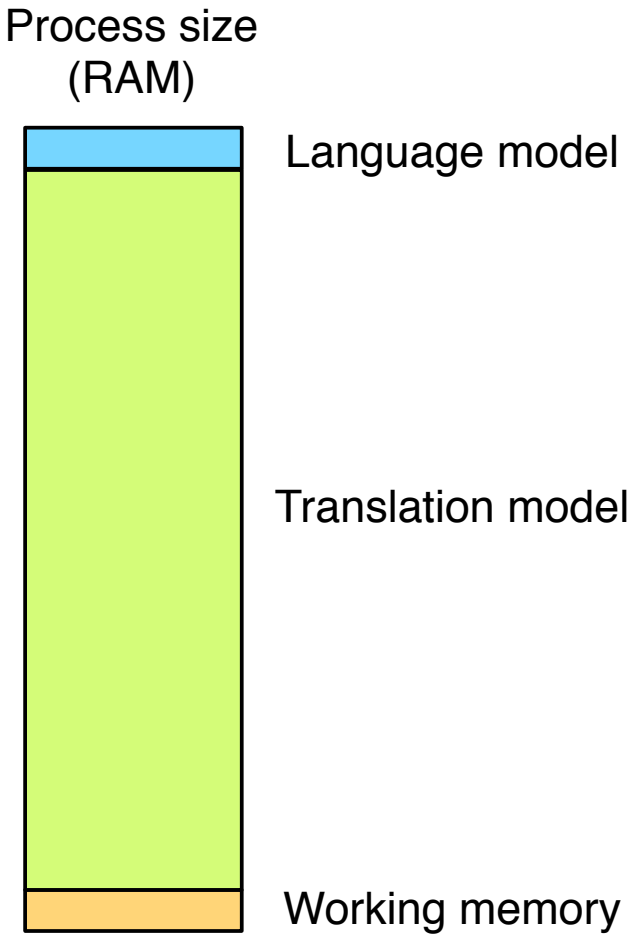
...

Speed vs. Memory Use

Typical Europarl file sizes:

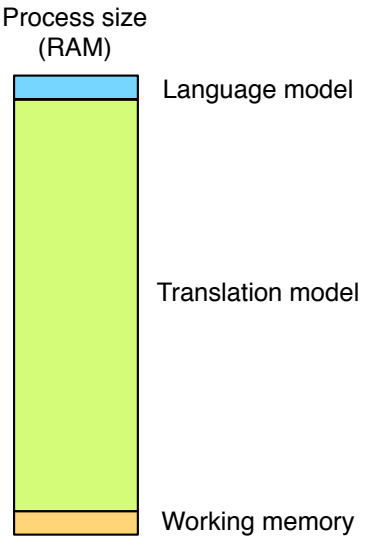
- Language model
  - 170 MB (trigram)
  - 412 MB (5-gram)
- Phrase table
  - 11GB
- Lexicalized reordering
  - 9.4GB

→ total = 20.8 GB



Speed vs. Memory Use

- Load into memory
  - long load time
  - large memory usage
  - fast decoding
- Load-on-demand
  - store indexed model on disk
  - binary format
  - minimal start-up time, memory usage
  - slower decoding



New: Compact Phrase Table

- Memory-efficient data structure
  - phrase table 6–7 times smaller than on-disk binary table
  - lexical reordering table 12–15 times smaller than on-disk binary table
- Stored in RAM
- May be memory mapped

IRSTLM

- Developed by FBK-irst, Trento, Italy
- Create a binary format which can be read from disk as needed
  - reduces memory but slower decoding
- Quantization of probabilities
  - reduces memory but lose accuracy
  - probability stored in 1 byte instead of 4 bytes

KENLM

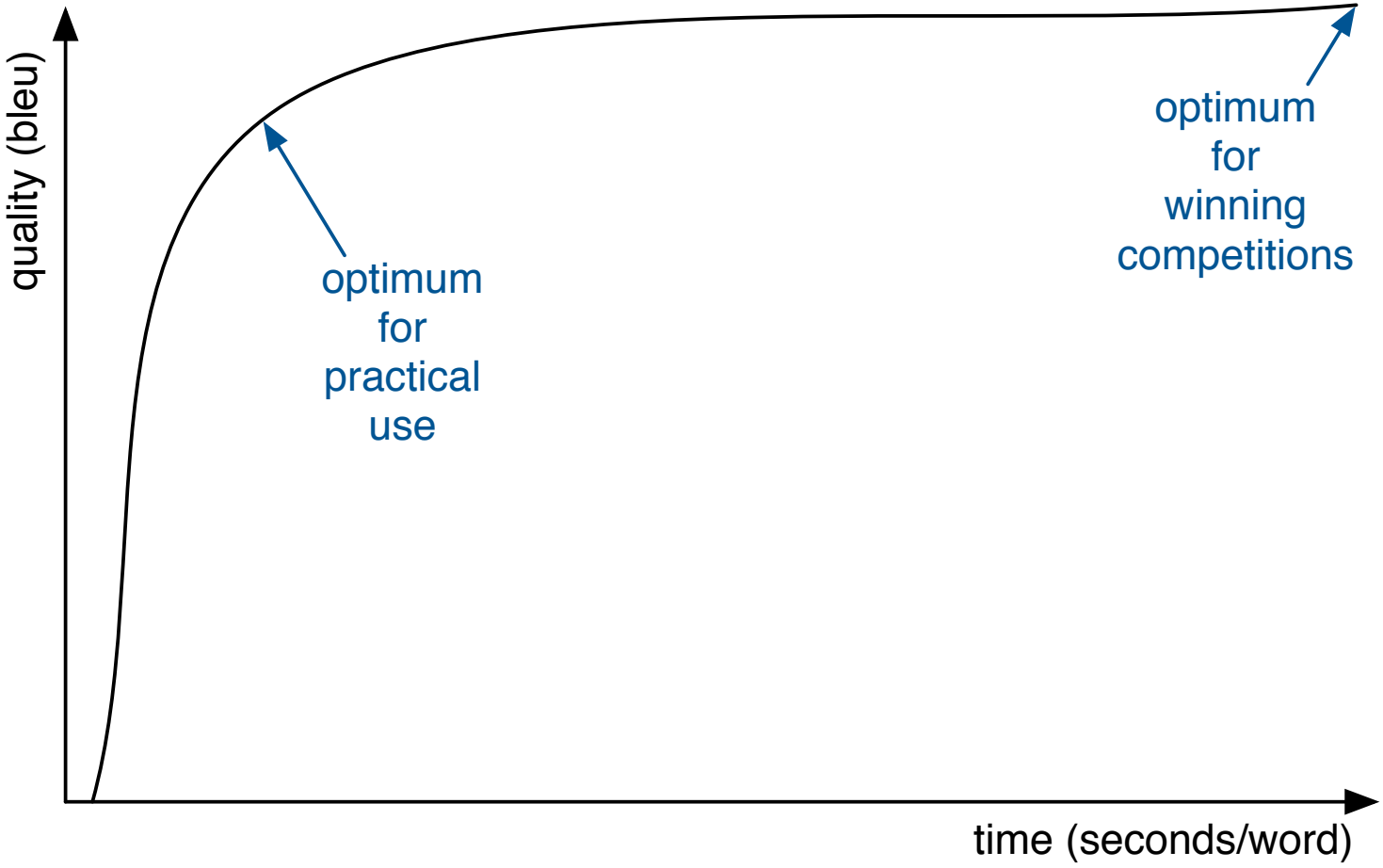
- Developed by Kenneth Heafield (CMU / Edinburgh / Stanford)
- Fastest and smallest language model implementation

Advanced Features

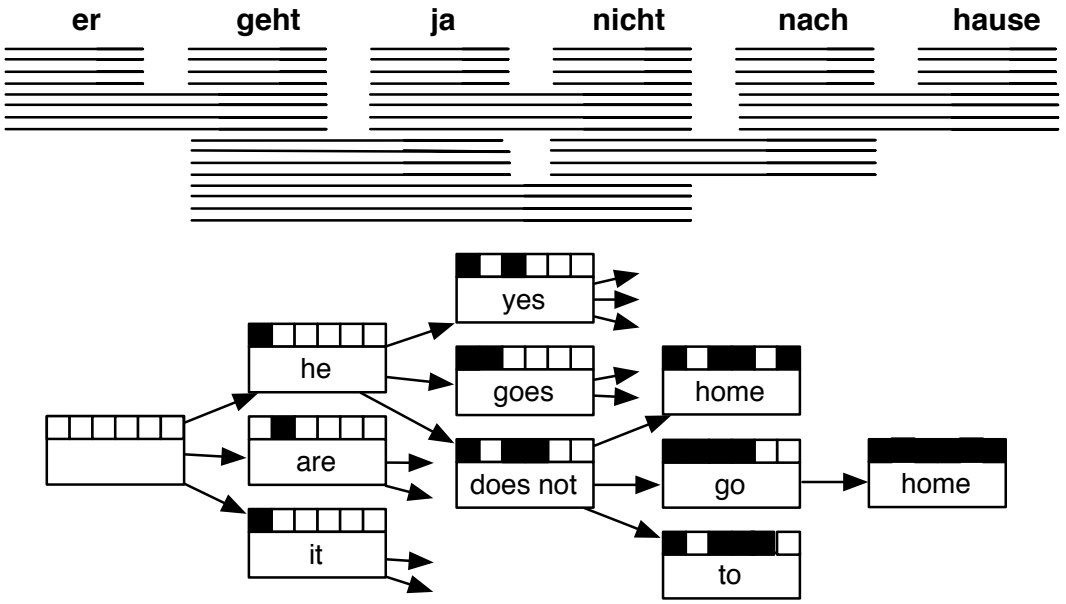
- How do I get started?
- Experiment Management System
- Faster Decoding
  - Speed vs. Memory
  - **Speed vs. Quality**

...

Speed vs. Quality

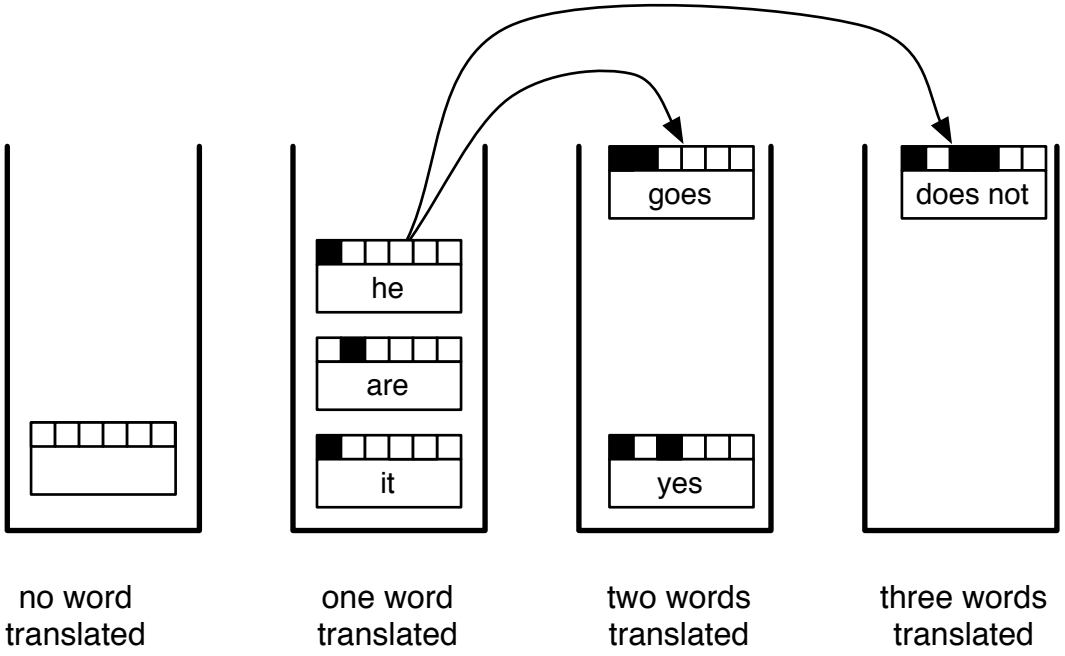


Speed vs. Quality



- Decoder search creates very large number of partial translations ("hypotheses")
- Decoding time $\sim$ number of hypotheses created
- Translation quality $\sim$ number of hypothesis created

Hypothesis Stacks



- Phrase-based: One stack per number of input words covered
- Number of hypothesis created =
  sentence length × stack size × applicable translation options

Advanced Features

- How do I get started?
- Experiment Management System
- Faster Decoding
- Data and domain adaptation
- Instructions to decoder
- Input formats
- Output formats

Advanced Features

- How do I get started?
- Experiment Management System
- Faster Decoding
- **Data and domain adaptation**
  - Train everything together
  - Secondary phrase table
  - Domain indicator features
  - Interpolated language models
  - TM-MT integration

Data

- Parallel corpora → translation model
  - sentence-aligned translated texts
  - translation memories are parallel corpora
  - dictionaries are parallel corpora
- Monolingual corpora → language model
  - text in the target language
  - billions of words easy to handle

Domain Adaptation

- The more data, the better
- The more in-domain data, the better
  (even in-domain monolingual data very valuable)
- Always tune towards target domain

Advanced Features

- How do I get started?
- Experiment Management System
- Faster Decoding
- Data and domain adaptation
  - **Train everything together**
  - Secondary phrase table
  - TM-MT integration

Default: Train Everything Together

- Easy to implement
  - Concatenate new data with existing data
  - Retrain
- Disadvantages:
  - Slower training for large amount of data
  - Cannot weight old and new data separately

Advanced Features

- How do I get started?
- Experiment Management System
- Faster Decoding
- Data and domain adaptation
  - Train everything together
  - **Secondary phrase table**
  - Domain indicator features
  - Interpolated language models
  - TM-MT integration

TM-MT Integration

- Input sentence:

The second paragraph of Article 21 is deleted .

- Fuzzy match in translation memory:

The second paragraph of Article 5 is deleted .

=

À l' article 5 , le texte du deuxiéme alinéa est supprimé .

Output word(s) taken from the target TM

Input word(s) that still need to be translated by SMT

Secondary Phrase Table

- Train initial phrase table and LM on baseline data
- Train secondary phrase table and LM new/in-domain data
- Use both in Moses

TM-MT Integration

- Input sentence:

The second paragraph of Article 21 is deleted .

- Fuzzy match in translation memory:

The second paragraph of Article 5 is deleted .

=

À l' article 5 , le texte du deuxiéme alinéa est supprimé .

Output word(s) taken from the target TM

Input word(s) that still need to be translated by SMT

Advanced Features

- How do I get started?
- Experiment Management System
- Faster Decoding
- Data and domain adaptation
- **Instructions to decoder**
- Input formats
- Output formats

Specifying Translations with XML

- Translation tables for numbers?

f	e	$p(f\|e)$
2003	2003	0.7432
2003	2000	0.0421
2003	year	0.0212
2003	the	0.0175
2003	...	...

- Instruct the decoder with XML instruction

```
the revenue for <num translation="2003"> 2003 </num> is higher than ...
```

- Deal with different number formats

```
er erzielte <num translation="17.55"> 17,55 </num> Punkte .
```

Specifying Translations with XML


```
./moses -xml-input [exclusive | inclusive | constraint ]
```

the revenue for <num translation="2003"> 2003 </num> is higher than ...

Three types of XML input:

- Exclusive
  Only possible translation is given in XML
- Inclusive
  Translation is given in XML is in addition to phrase-table
- Constraint
  Only use translations from phrase-table if it match XML specification

Constraint XML

- Specifically for translating terminology
  - consistently translate particular phrase in a document
  - may have learned larger phrase pairs that contain terminology term
- Example:

```
Microsoft <option translation="Windows"> Windows </option> 8 ...
```

- Allows use of phrase pair only if maps Windows to Windows

Placeholders

- Translate:
  - You owe me 100 dollars !
  - You owe me 200 dollars !
  - You owe me 9.56 dollars !
- Problem: need translations for
  - 100
  - 200
  - 9.56
- Some things are better off being handled by simple rules:
  - Numbers
  - Dates
  - Currency
  - Named entities

Placeholders

- Input
  You owe me 100 dollars !
- Replace numbers with `@num@`

```
You owe me @num@ dollars !
```

- Specification

```
You owe me <ne translation="@num@" entity="100">@num@</ne> dollars !
```

Walls and Zones

- Specification of reordering constraints
- Zone
  sequence to be translated without reordering with outside material
- Wall
  hard reordering constraint, no words may be reordered across
- Local wall
  wall within a zone, not valid outside zone

Walls and Zones: Examples

- Requiring the translation of quoted material as a block

  He said <zone> ” yes ” </zone> .

- Hard reordering constraint

  Number 1 : <wall/> the beginning .

- Local hard reordering constraint within zone

  A new plan <zone> (<wall/> maybe not new <wall/>) </zone> emerged .

- Nesting

  The <zone> ” new <zone> (old) </zone> ” </zone> proposal .

Preserving Markup

- How do you translate this:

<h1>My Home Page</h1>
I really like to eat chicken!

- Solution 1: XML translations, walls and zones

```
<x translation="<h1>"/> <wall/> My Home Page <wall/>
<x translation="</h1>"/>
```

```
I really like to <zone><x translation="<b>"/> <wall/> eat <wall/>
<x translation="</b>"/> </zone> chicken !
```

(note: special XML characters like < and > need to be escaped)

Preserving Markup

- Solution 2: Handle markup externally
  - track word positions and their markup

I	really	like	to	eat	chicken	!
1	2	3	4	5	6	7
-	-	-	-		-	-

  - translate without markup

I really like to eat chicken !

  - keep word alignment to source

Ich	esse	wirklich	gerne	Hühnchen	!
1	5	2	3-4	6	7

  - re-insert markup

Ich esse wirklich gerne Hühnchen!

Advanced Features

- How do I get started?
- Experiment Management System
- Faster Decoding
- Moses Server
- Data and domain adaptation
- Instructions to decoder
- **Input formats**
- Output formats

Example: Misspelt Words

- Misspelt sentence:

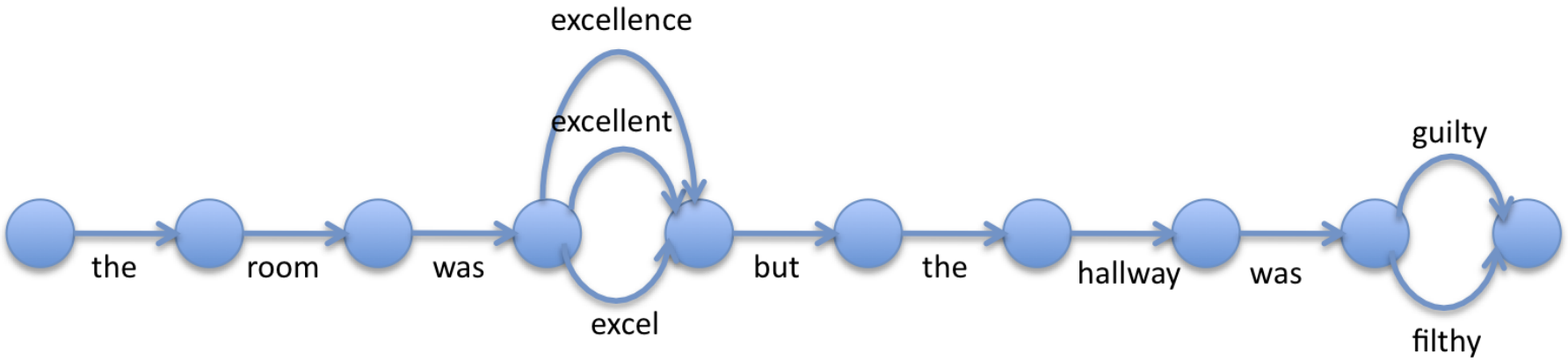
  The room was *exellent but the hallway was *filty .

- Strategies for dealing with spelling errors:
  - Create correct sentence with correction
    × problem: if not corrected properly, adds more errors
  - Create many sentences with different corrections
    × problem: have to decode each sentence, slow

Confusion Network

The room was *exellent but the hallway was *filty .

Input to decoder:



Let the decoder decide

Example: Diacritics

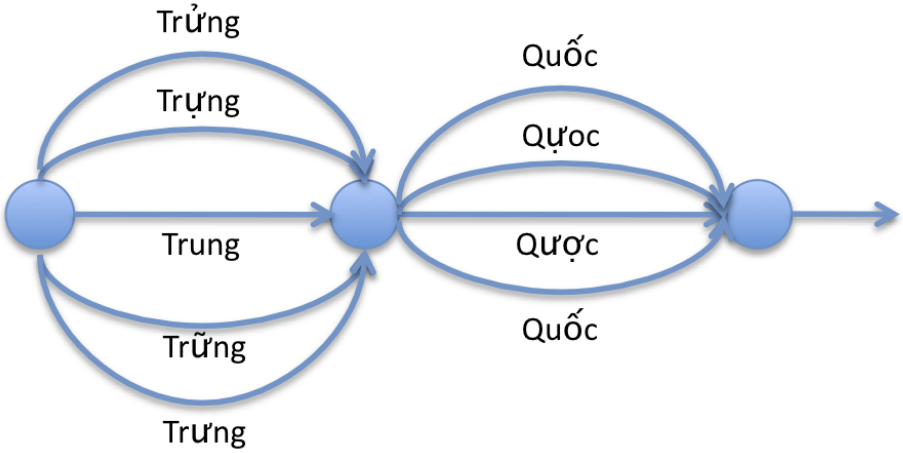
- Correct sentence

  Trung Quốc cảnh báo Mỹ về luật tiền tệ

- Something a non-native person might type

  Trung Quoc canh bao My ve luat tien te

- Confusion network



Confusion Network Specification

Argument on command line

```
./moses -inputtype 1
```

Input to moses

```
the 1.0
room 1.0
was 1.0
excel 0.33 excellent 0.33 excellence 0.33
but 1.0
the 1.0
hallway 1.0
was 1.0
guilty 0.5 filthy 0.5
```

Lattice

Example: Chinese Word Segmentation

- Unsegmented sentence

硬质合金号称”工业牙齿"

- Incorrect segmention

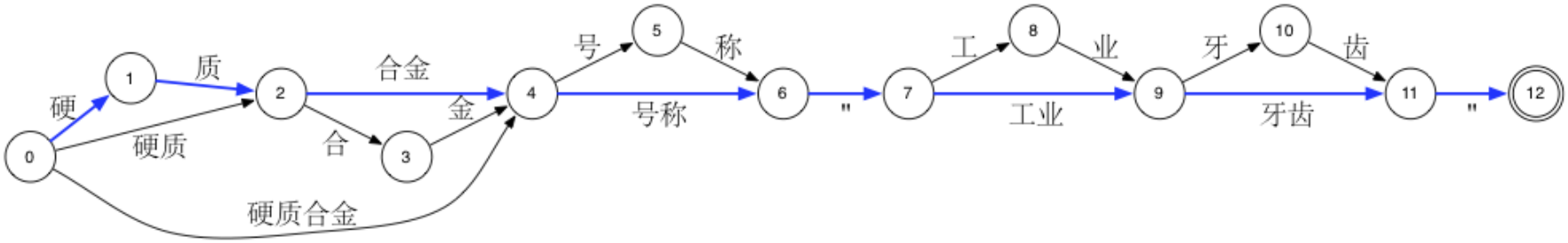
硬质 合 金 号称 " 工 业牙 齿 "

- Correct segmention

硬 质 合金 号称 " 工业 牙齿 "

Lattice

Input to decoder:



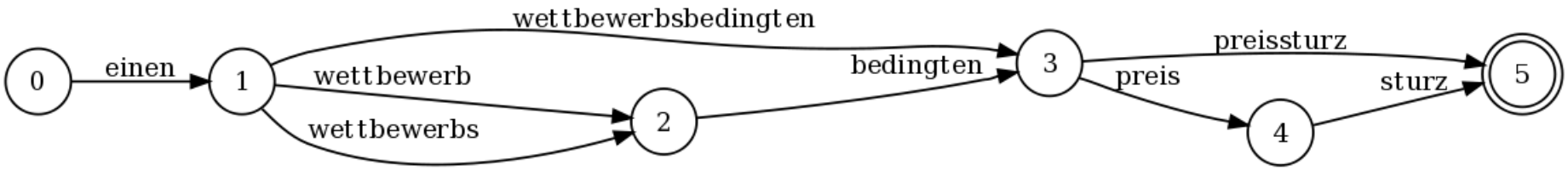
Let the decoder decide

Example: Compound Splitting

- Input sentence

einen wettbewerbsbedingten preissturz

- Different compound splits



- Let the decoder decide

Lattice Specification

Command line argument

```
./moses -inputtype 1
```

Input to Moses (PLF format - Python Lattice Format)

```
(
 (
  ('einen', 1.0, 1),
 ),
 (
  ('wettbewerbsbedingten', 0.5, 2),
  ('wettbewerbs', 0.25, 1),
  ('wettbewerb', 0.25, 1),
 ),
 (
  ('bedingten', 1.0, 1),
 ),
 (
  ('preissturz', 0.5, 2),
  ('preis', 0.5, 1),
 ),
 (
  ('sturz', 1.0, 1),
 ),
)
```

Advanced Features

- How do I get started?
- Experiment Management System
- Faster Decoding
- Data and domain adaptation
- Instructions to decoder
- Input formats
- **Output formats**

N-Best List

- Input

es gibt verschiedene andere meinungen .

- Best Translation

there are various different opinions .

- Next nine best translations

there are various other opinions .
there are different different opinions .
there are other different opinions .
we are various different opinions .
there are various other opinions of .
it is various different opinions .
there are different other opinions .
it is various other opinions .
it is a different opinions .

Uses of N-Best Lists

- Let the translator choose from possible translations
- Reranker
  - add more knowledge sources
  - can take global view
  - coherency of whole sentence
  - coherency of document
- Used to tune component weights

N-Best Lists in Moses

Argument to command line

```
./moses -n-bestlist n-best.file.txt [distinct] 100
```

Output

```
0 ||| there are various different opinions .  ||| d: 0 lm: -21.6664 w: -6 ...  ||| -113.734
0 ||| there are various other opinions .  ||| d: 0 lm: -25.3276 w: -6 ... ||| -114.004
0 ||| there are different different opinions .  ||| d: 0 lm: -27.8429 w: -6 ...  ||| -117.738
0 ||| there are other different opinions .  ||| d: -4 lm: -25.1666 w: -6 ...  ||| -118.007
0 ||| we are various different opinions .  ||| d: 0 lm: -28.1533 w: -6 ...  ||| -118.142
0 ||| there are various other opinions of .  ||| d: 0 lm: -33.7616 w: -7 ...  ||| -118.153
0 ||| it is various different opinions .  ||| d: 0 lm: -29.8191 w: -6  ... ||| -118.222
0 ||| there are different other opinions .  ||| d: 0 lm: -30.426 w: -6 ...  ||| -118.236
0 ||| it is various other opinions .  ||| d: 0 lm: -32.6824 w: -6 ... ||| -118.395
0 ||| it is a different opinions .  ||| d: 0 lm: -20.1611 w: -6 ...  ||| -118.434
```

Acknowledgements

Moses Developers

Abhishek Arun
Amittai Axelrod
Barry Haddow
Christian Hardmeier
Edmund Huber
Frederic Blain
Hieu Hoang
Jean-Baptiste Fouet
Abby Levenberg
Mauro Cettolo
Mark Fishel
Nicola Bertoldi
Phil Williams
Joao Lus Rosas
Sara Stymne
Yizhao Ni
Suzy Howlett
Alexander Fraser

Adam Lopez
Ankit Srivastava
Ondrej Bojar
Christian Federmann
Evan Herbst
Brooke Cowan
H. Leal Fontes
Joern Wuebker
Alexandra Birch
Marcello Federico
Gabriele Antonio Musillo
Oliver Wilson
Bruno Pouliquen
Rico Sennrich
Steven B. Parks
David Kolovratnak
Wade Shen
Richard Zens

Ales Tamchyna
Anthony Rousseau
Chris Callison-Burch
Lane Schwartz
Andreas Eisele
Grace M. Ngai
Holger Schwenk
Jorge Civera
Bo Fu
Michael Auli
Miles Osborne
Pascual Martinez
Raphael Payen
Herve Saint-Amand
Steven Buraje Poggel
Sergio Penkale
Yang Gao

Alex
Benjamin Gottesman
Christine Corbett
David Talbot
Eva Hasler
Kenneth Heafield
Josh Schroeder
Konrad Rawlik
M.J.Bellino-Machado
John Joseph Morgan
Nadi Tomeh
Philipp Koehn
Chris Dyer
Felipe Sanchez Martinez
Andre Lynum
Stephan
Tsuyoshi Okita