Statistical Machine Translation and Language Barriers

Philipp Koehn, University of Edinburgh

23 September 2011

Overview

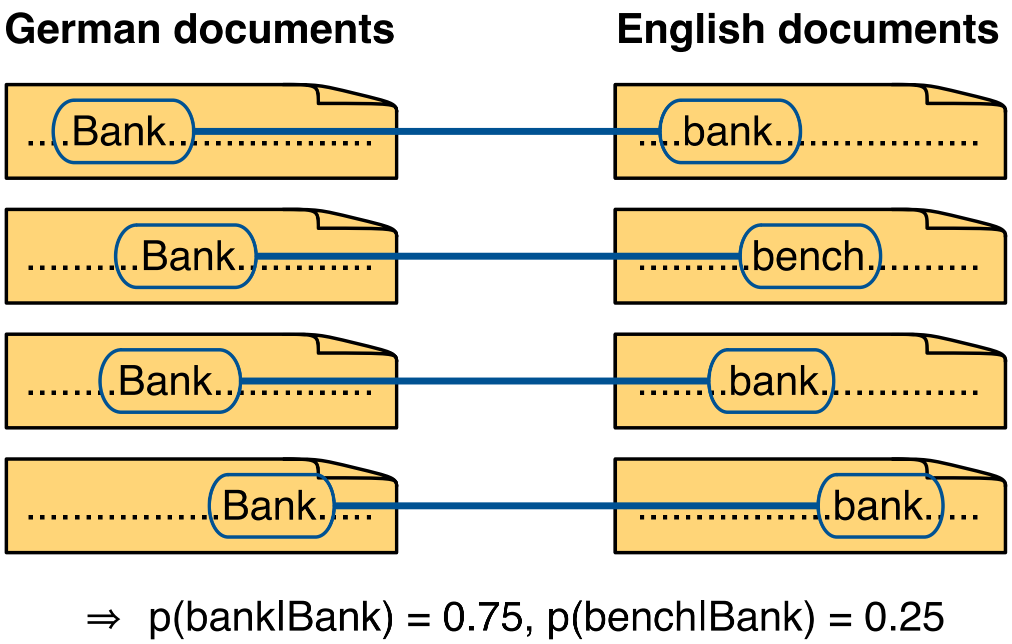
- How does statistical machine translation work?

- How well does statistical machine translation work?

- Patent translation

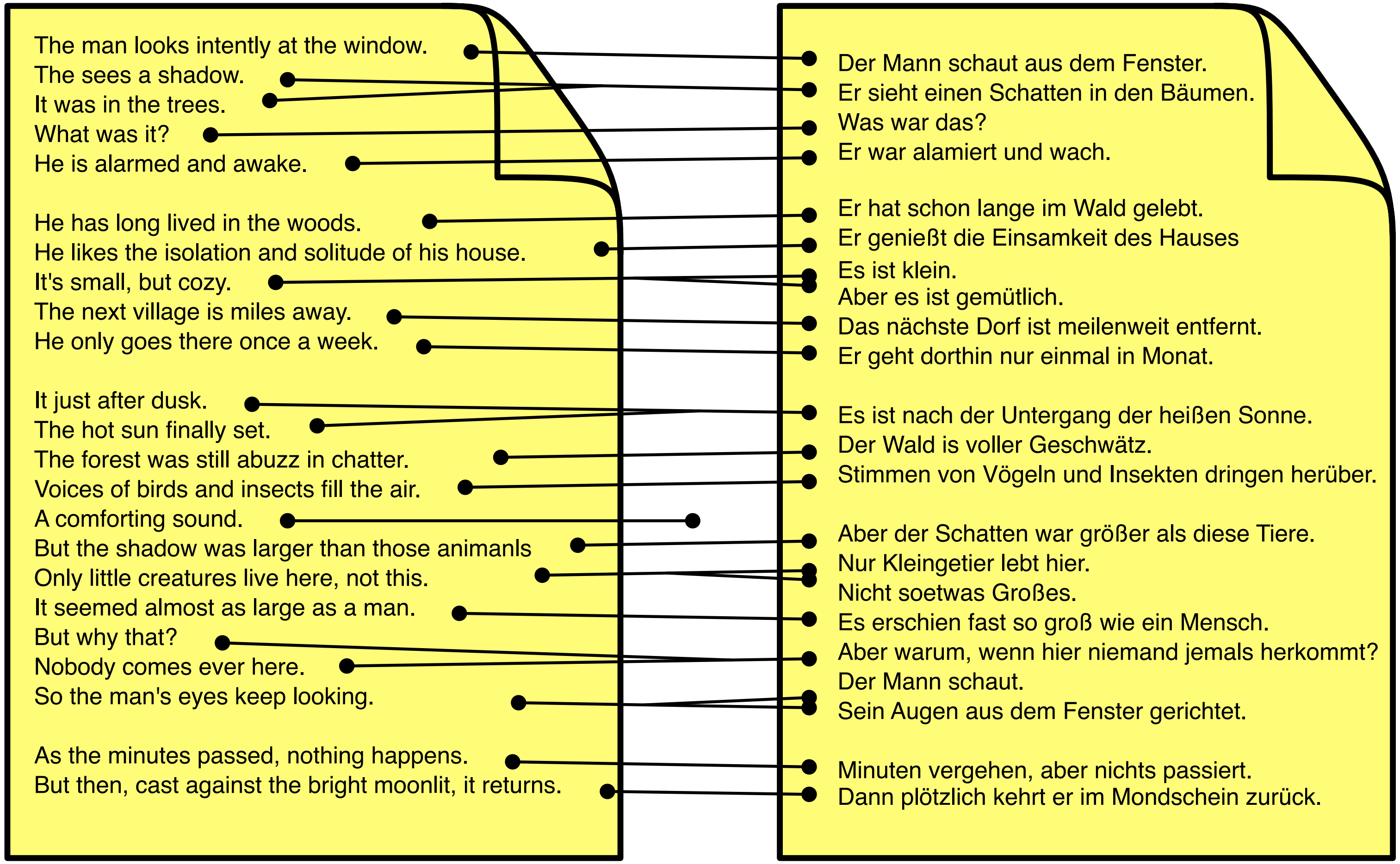
Statistical Machine Translation

- Learning from data (sentence-aligned translated texts)

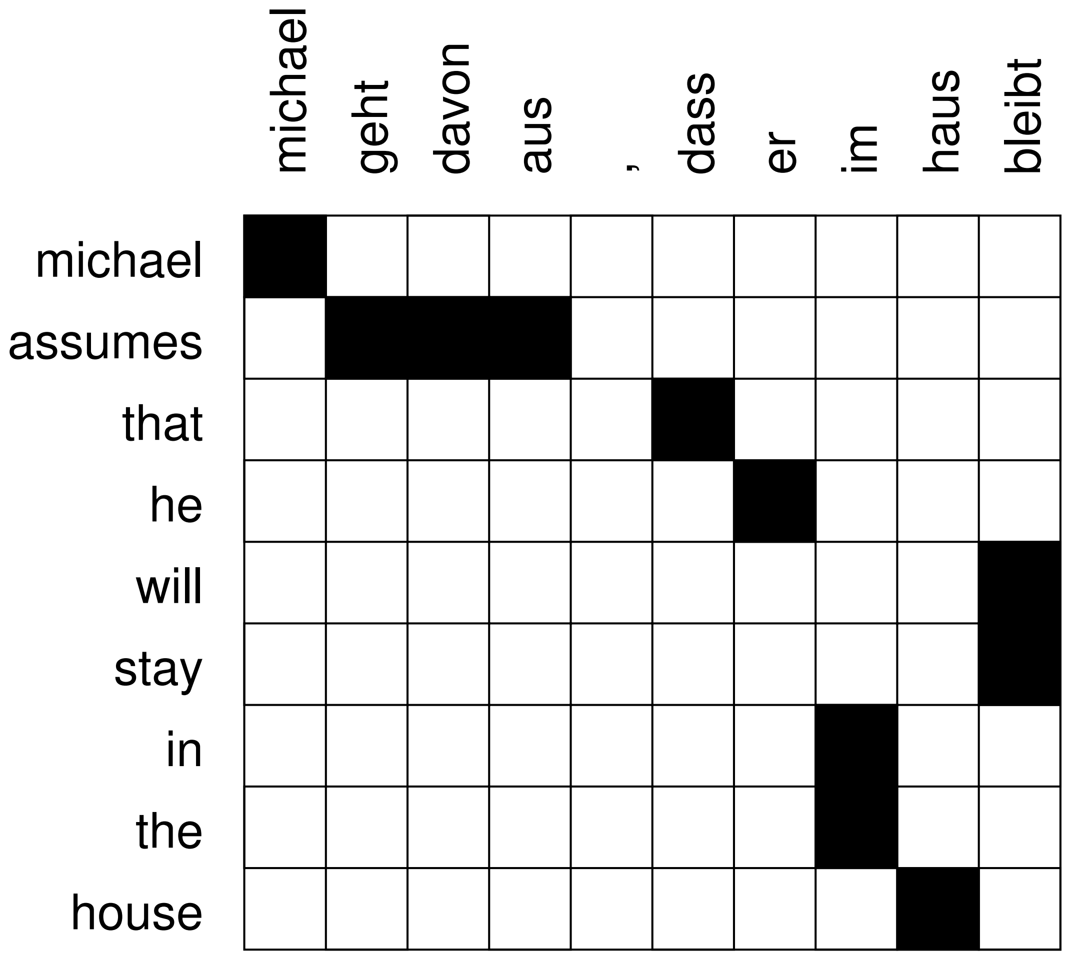


- New machine translation systems can be built automatically

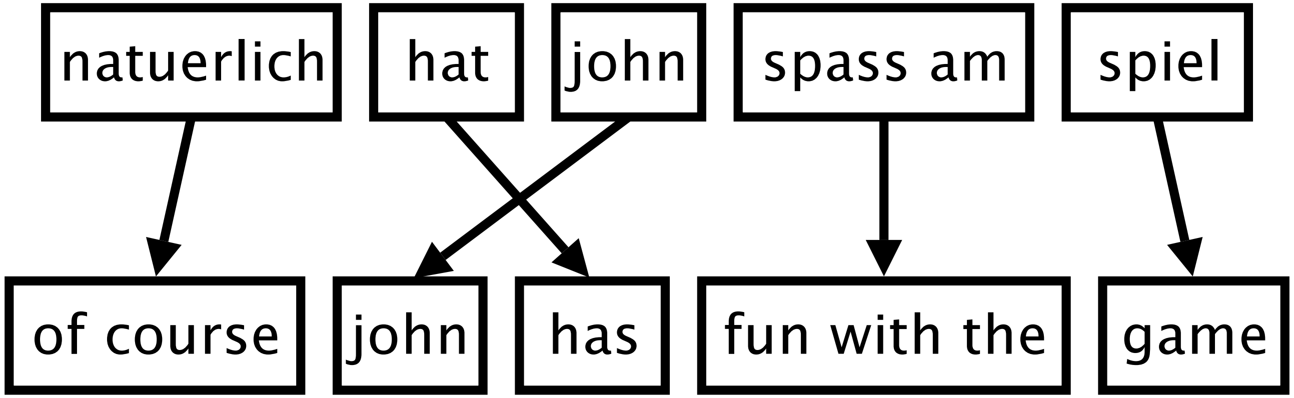
Preparing the Data: Sentence Alignment



Preparing the Data: Word Alignment

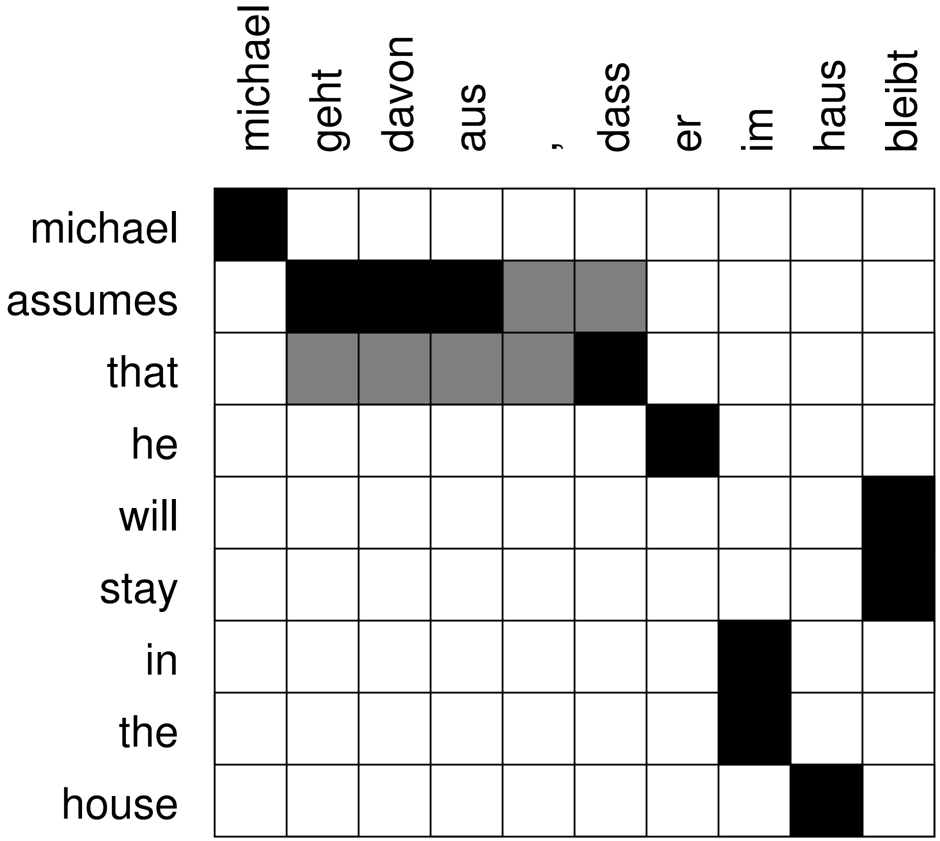


Phrase-Based Translation



- Foreign input is segmented in phrases
  - any sequence of words, not necessarily linguistically motivated
- Each phrase is translated into English
- Phrases are reordered

Extracting Phrases from Data



Given a word alignment: extract phrase pairs, estimate probabilities

Phrase Translation Table

- Phrase Translations for "den Vorschlag"

English	$\phi(e\|f)$	English	$\phi(e\|f)$
the proposal	0.6227	the suggestions	0.0114
's proposal	0.1068	the proposed	0.0114
a proposal	0.0341	the motion	0.0091
the idea	0.0250	the idea of	0.0091
this proposal	0.0227	the proposal ,	0.0068
proposal	0.0205	its proposal	0.0068
of the proposal	0.0159	it	0.0068
the proposals	0.0159	...	...

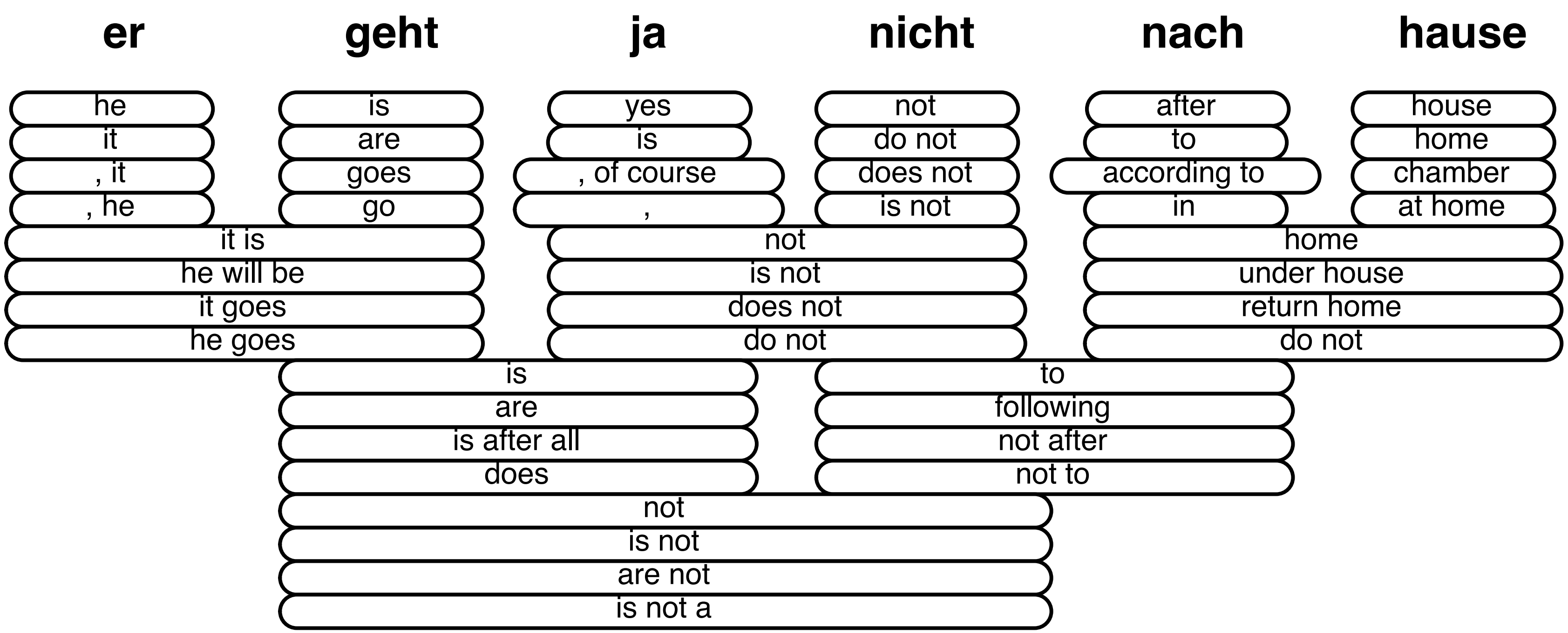
Language Model

- **Language models** answer the question: How likely is a string of English words good English?
  - the house is big → good
  - the house is xxl → worse
  - house big is the → bad
- Given: English words $W = w_1, w_2, w_3, ..., w_n$ — *what is* $p(W)$?
- Limited history: only previous k words matter (here: k=2)

$$p(w_1, w_2, w_3, ..., w_n) = p(w_1)\ p(w_2|w_1)\ p(w_3|w_2)...p(w_n|w_{n-1})$$

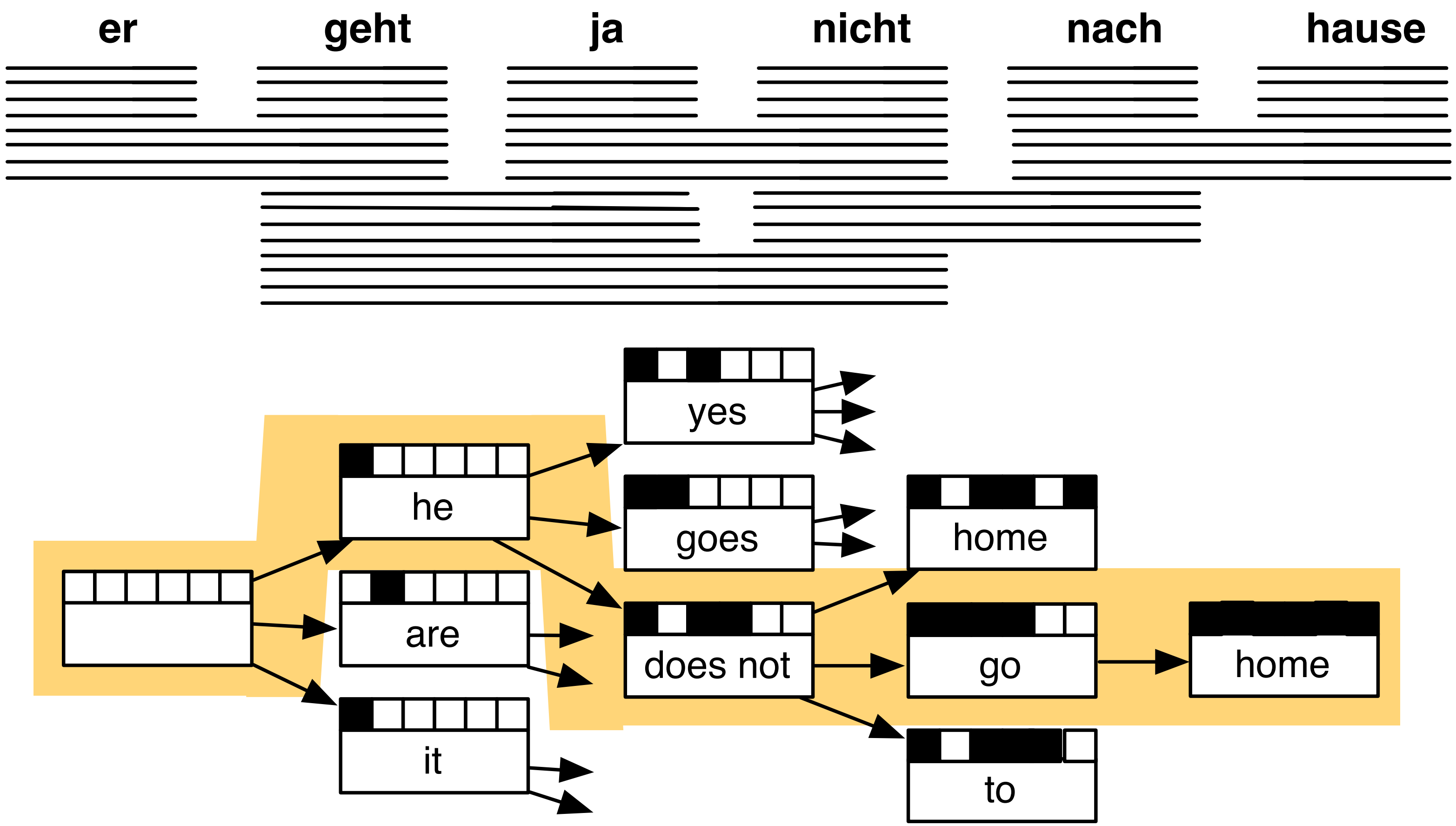
- Models trained on large amounts of monolingual text (billions of words)

Translation Options

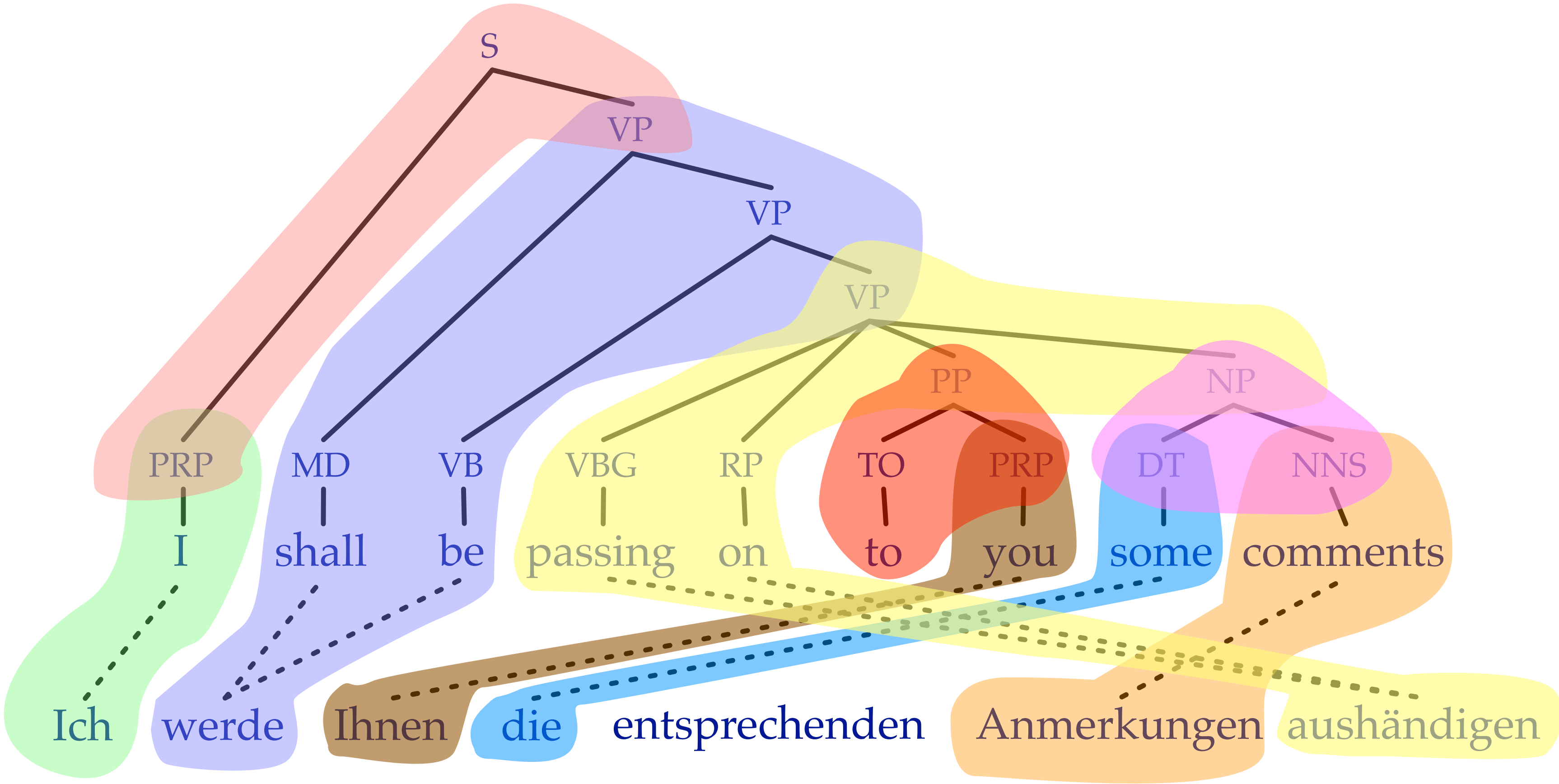


- Task: find the right output phrases, put them in the right order

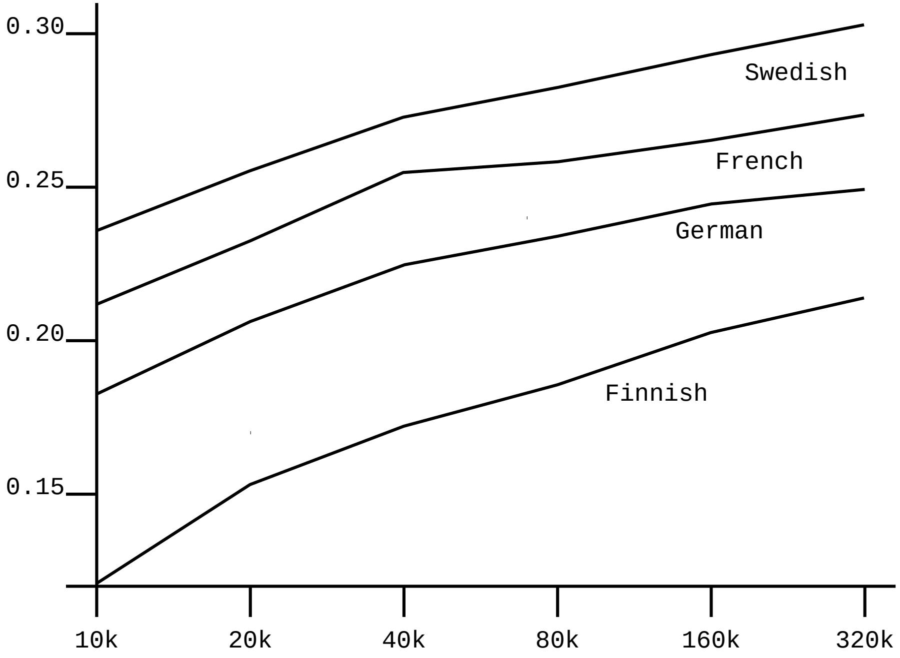
Decoding


er
geht
ja
nicht
nach
hause
yes
he
goes
home
are
does not
go
home
it
to

Syntactic Models



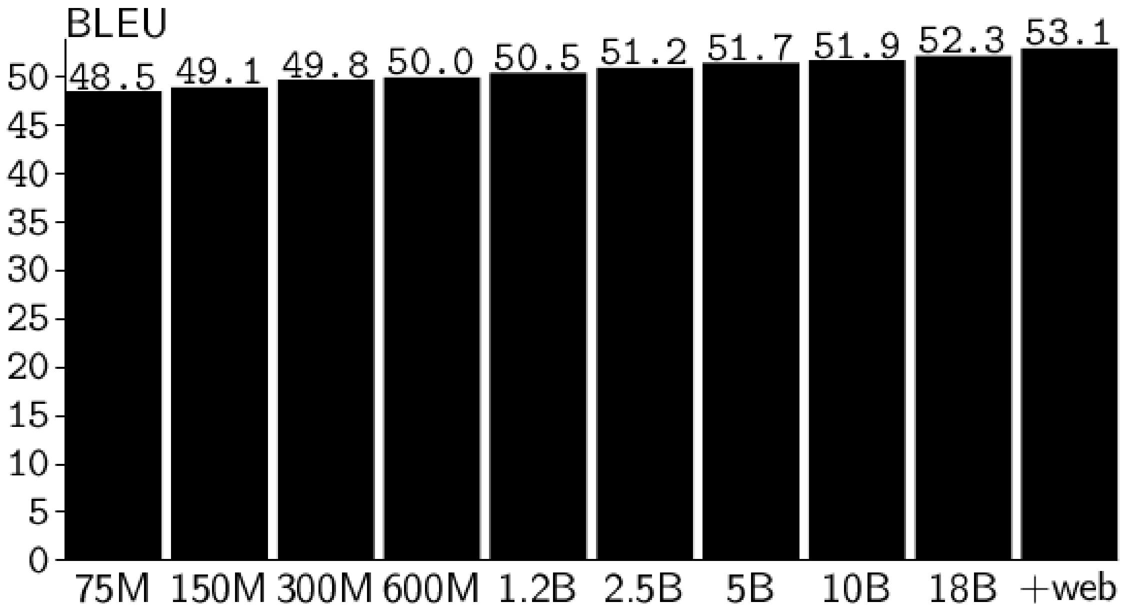
More Data, Better Translations



[from Koehn, 2003: Europarl]

- Log-scale improvements on BLEU:
  Doubling the training data gives constant improvement (+1 %BLEU)

More LM Data, Better Translations



[from Och, 2005: MT Eval presentation]

- Also log-scale improvements on BLEU:
  doubling the training data gives constant improvement (+0.5 %BLEU)
  (last addition is 218 billion words out-of-domain web data)

Overview

- How does statistical machine translation work?

- **How well does statistical machine translation work?**

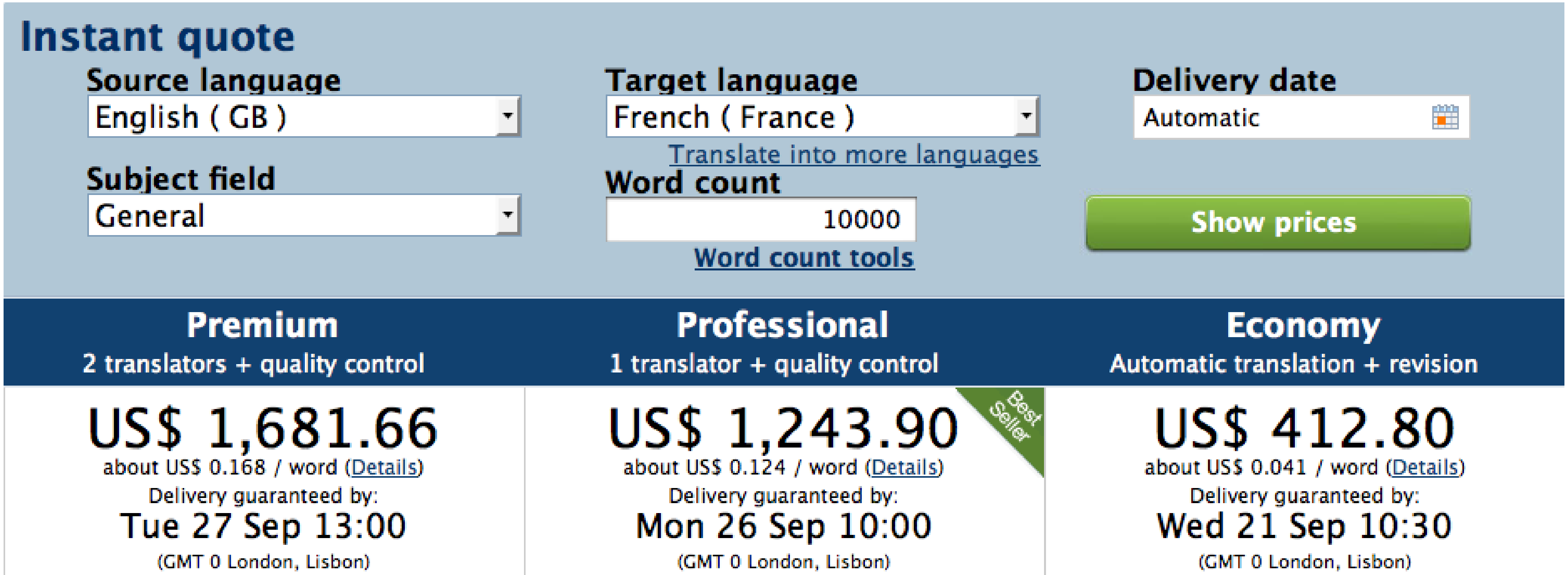
- Patent translation

How Good?

- It depends...

- Better question: Good enough for what?
  - understanding the general meaning of a document
  - understanding most of the details
  - humans faster when post-editing than translating from scratch
  - publication-quality

75% Cost Cut?



Instant quote
Source language
English (GB)
Subject field
General
Target language
French (France)
Translate into more languages
Word count
10000
Word count tools
Delivery date
Automatic
Show prices
Premium
2 translators + quality control
US$ 1,681.66
about US$ 0.168 / word (Details)
Delivery guaranteed by:
Tue 27 Sep 13:00
(GMT 0 London, Lisbon)
Professional
1 translator + quality control
US$ 1,243.90
Best Seller
about US$ 0.124 / word (Details)
Delivery guaranteed by:
Mon 26 Sep 10:00
(GMT 0 London, Lisbon)
Economy
Automatic translation + revision
US$ 412.80
about US$ 0.041 / word (Details)
Delivery guaranteed by:
Wed 21 Sep 10:30
(GMT 0 London, Lisbon)

Language Pairs Differ

Target Language

	en	bg	de	cs	da	el	es	et	fi	fr	hu	it	lt	lv	mt	nl	pl	pt	ro	sk	sl	sv
en	–	40.5	46.8	52.6	50.0	41.0	55.2	34.8	38.6	50.1	37.2	50.4	39.6	43.4	39.8	52.3	49.2	55.0	49.0	44.7	50.7	52.0
bg	61.3	–	38.7	39.4	39.6	34.5	46.9	25.5	26.7	42.4	22.0	43.5	29.3	29.1	25.9	44.9	35.1	45.9	36.8	34.1	34.1	39.9
de	53.6	26.3	–	35.4	43.1	32.8	47.1	26.7	29.5	39.4	27.6	42.7	27.6	30.3	19.8	50.2	30.2	44.1	30.7	29.4	31.4	41.2
cs	58.4	32.0	42.6	–	43.6	34.6	48.9	30.7	30.5	41.6	27.4	44.3	34.5	35.8	26.3	46.5	39.2	45.7	36.5	43.6	41.3	42.9
da	57.6	28.7	44.1	35.7	–	34.3	47.5	27.8	31.6	41.3	24.2	43.8	29.7	32.9	21.1	48.5	34.3	45.4	33.9	33.0	36.2	47.2
el	59.5	32.4	43.1	37.7	44.5	–	54.0	26.5	29.0	48.3	23.7	49.6	29.0	32.6	23.8	48.9	34.2	52.5	37.2	33.1	36.3	43.3
es	60.0	31.1	42.7	37.5	44.4	39.4	–	25.4	28.5	51.3	24.0	51.7	26.8	30.5	24.6	48.8	33.9	57.3	38.1	31.7	33.9	43.7
et	52.0	24.6	37.3	35.2	37.8	28.2	40.4	–	37.7	33.4	30.9	37.0	35.0	36.9	20.5	41.3	32.0	37.8	28.0	30.6	32.9	37.3
fi	49.3	23.2	36.0	32.0	37.9	27.2	39.7	34.9	–	29.5	27.2	36.6	30.5	32.5	19.4	40.6	28.8	37.5	26.5	27.3	28.2	37.6
fr	64.0	34.5	45.1	39.5	47.4	42.8	60.9	26.7	30.0	–	25.5	56.1	28.3	31.9	25.3	51.6	35.7	61.0	43.8	33.1	35.6	45.8
hu	48.0	24.7	34.3	30.0	33.0	25.5	34.1	29.6	29.4	30.7	–	33.5	29.6	31.9	18.1	36.1	29.8	34.2	25.7	25.6	28.2	30.5
it	61.0	32.1	44.3	38.9	45.8	40.6	26.9	25.0	29.7	52.7	24.2	–	29.4	32.6	24.6	50.5	35.2	56.5	39.3	32.5	34.7	44.3
lt	51.8	27.6	33.9	37.0	36.8	26.5	21.1	34.2	32.0	34.4	28.5	36.8	–	40.1	22.2	38.1	31.6	31.6	29.3	31.8	35.3	35.3
lv	54.0	29.1	35.0	37.8	38.5	29.7	8.0	34.2	32.4	35.6	29.3	38.9	38.4	–	23.3	41.5	34.4	39.6	31.0	33.3	37.1	38.0
mt	72.1	32.2	37.2	37.9	38.9	33.7	48.7	26.9	25.8	42.4	22.4	43.7	30.2	33.2	–	44.0	37.1	45.9	38.9	35.8	40.0	41.6
nl	56.9	29.3	46.9	37.0	45.4	35.3	49.7	27.5	29.8	43.4	25.3	44.5	28.6	31.7	22.0	–	32.0	47.7	33.0	30.1	34.6	43.6
pl	60.8	31.5	40.2	44.2	42.1	34.2	46.2	29.2	29.0	40.0	24.5	43.2	33.2	35.6	27.9	44.8	–	44.1	38.2	38.2	39.8	42.1
pt	60.7	31.4	42.9	38.4	42.8	40.2	60.7	26.4	29.2	53.2	23.8	52.8	28.0	31.5	24.8	49.3	34.5	–	39.4	32.1	34.4	43.9
ro	60.8	33.1	38.5	37.8	40.3	35.6	50.4	24.6	26.2	46.5	25.0	44.8	28.4	29.9	28.7	43.0	35.8	48.5	–	31.5	35.1	39.4
sk	60.8	32.6	39.4	48.1	41.0	33.3	46.2	29.8	28.4	39.4	27.4	41.8	33.8	36.7	28.5	44.4	39.0	43.3	35.3	–	42.6	41.8
sl	61.0	33.1	37.9	43.5	42.6	34.0	47.0	31.1	28.8	38.2	25.7	42.3	34.6	37.3	30.0	45.9	38.2	44.1	35.8	38.9	–	42.7
sv	58.5	26.9	41.0	35.6	46.6	33.3	46.6	27.4	30.9	38.9	22.7	42.0	28.2	31.0	23.7	45.6	32.2	44.2	32.7	31.3	33.5	–

(using the Acquis corpus)

[from Koehn et al., 2009]

What Makes MT Hard?

- Some language pairs more difficult than others

- Finding explanatory factors for diverging performance of Europarl systems

Explanatory Factor	R^2
Target vocabulary size ($\sim$ morphological complexity)	0.388
Reordering amount	0.384
Language similarity	0.366
Source vocabulary size ($\sim$ morphological complexity)	0.045

[from Birch et al., 2008]

- These factors explain together 75% of the differences in performance

- Similar results in study of Acquis systems [Koehn et al., 2009]

Overview

- How does statistical machine translation work?
- How well does statistical machine translation work?
- **Patent translation**

Specific Problems of Patent Translation

- Formulaic language
- Large vocabulary
- Long sentences
- Many specialized domains

Specific Problems of Patent Translation

- Formulaic language
  → easy to memorize

- Large vocabulary
  → learnable with sufficient data

- Long sentences
  → hard to translate for syntactically divergent languages

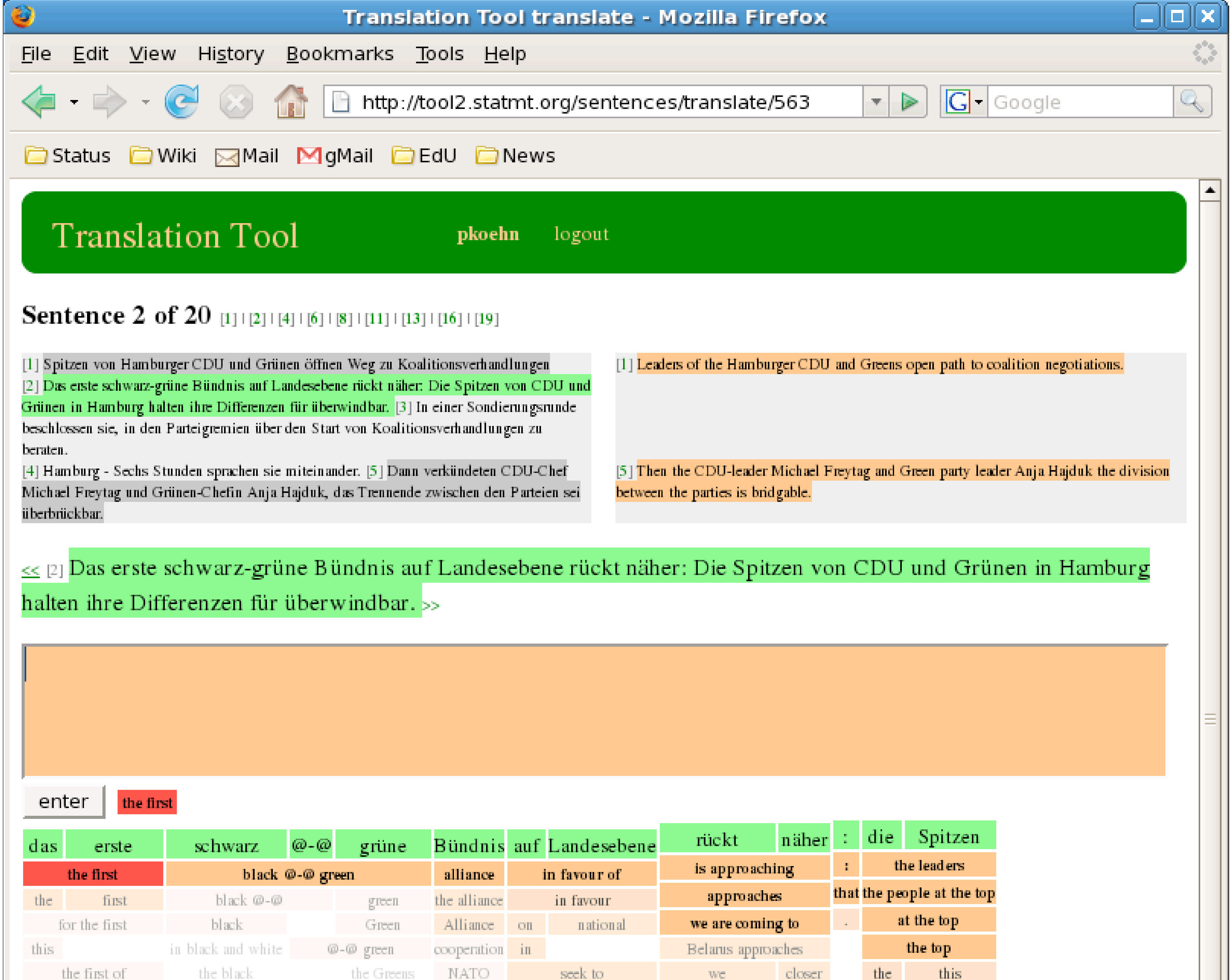
- Many specialized domains
  → challenge to develop better domain adaptation methods

Who is the Customer?

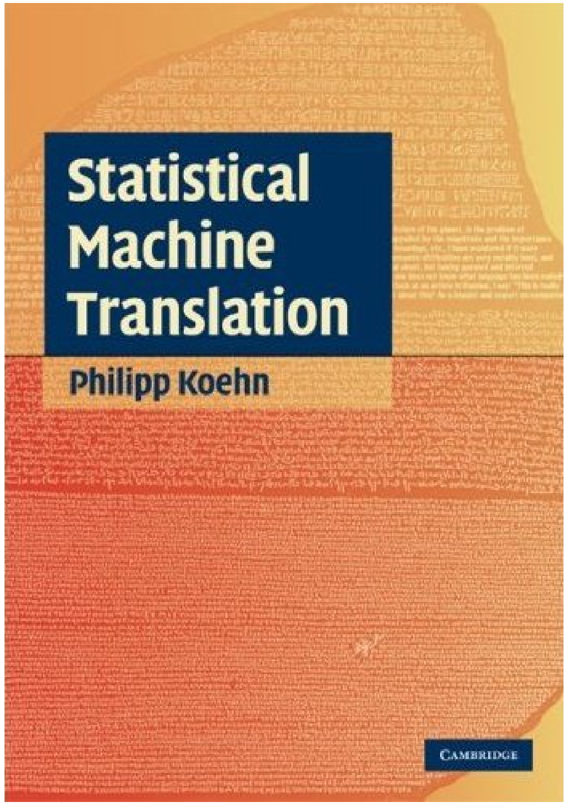
- Information seekers
  - user is tolerant of inferior quality
  - machine translation may be *good enough*
- Lawyers
  - high demands for quality
  - out of reach for machine translation

Computer Aided Translation

- Post-editing machine translation
- Other types of collaborations between human and machine
  - interactive machine translation
  - adapting machine translation to user's needs
  - interactive terminology database
  - bilingual concordancers
  - language model based fluency assistance
- Forthcoming EU projects: CASMACAT, MATECAT


Translation Tool translate - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://tool2.statmt.org/sentences/translate/563
Google
Status Wiki Mail gMail EdU News
Translation Tool
pkoehn
logout
Sentence 2 of 20 [1] | [2] | [4] | [6] | [8] | [11] | [13] | [16] | [19]
[1] Spitzen von Hamburger CDU und Grünen öffnen Weg zu Koalitionsverhandlungen
[2] Das erste schwarz-grüne Bündnis auf Landesebene rückt näher. Die Spitzen von CDU und Grünen in Hamburg halten ihre Differenzen für überwindbar. [3] In einer Sondierungsrunde beschlossen sie, in den Parteigremien über den Start von Koalitionsverhandlungen zu beraten.
[4] Hamburg - Sechs Stunden sprachen sie miteinander. [5] Dann verkündeten CDU-Chef Michael Freytag und Grünen-Chefin Anja Hajduk, das Trennende zwischen den Parteien sei überbrückbar.
[1] Leaders of the Hamburger CDU and Greens open path to coalition negotiations.
[5] Then the CDU-leader Michael Freytag and Green party leader Anja Hajduk the division between the parties is bridgable.
<< [2] Das erste schwarz-grüne Bündnis auf Landesebene rückt näher: Die Spitzen von CDU und Grünen in Hamburg halten ihre Differenzen für überwindbar. >>
enter
the first
das erste schwarz @-@ grüne Bündnis auf Landesebene rückt näher : die Spitzen
the first
black @-@ green
alliance
in favour of
is approaching
:
the leaders
the first
black @-@ green
the alliance
in favour
approaches
that
the people at the top
for the first
black
Green
Alliance
on
national
we are coming to
.
at the top
this
in black and white
@-@ green
cooperation
in
Belarus approaches
the top
the first of
the black
the Greens
NATO
seek to
we
closer
the
this

Questions?


Statistical
Machine
Translation
Philipp Koehn
CAMBRIDGE