Introduction

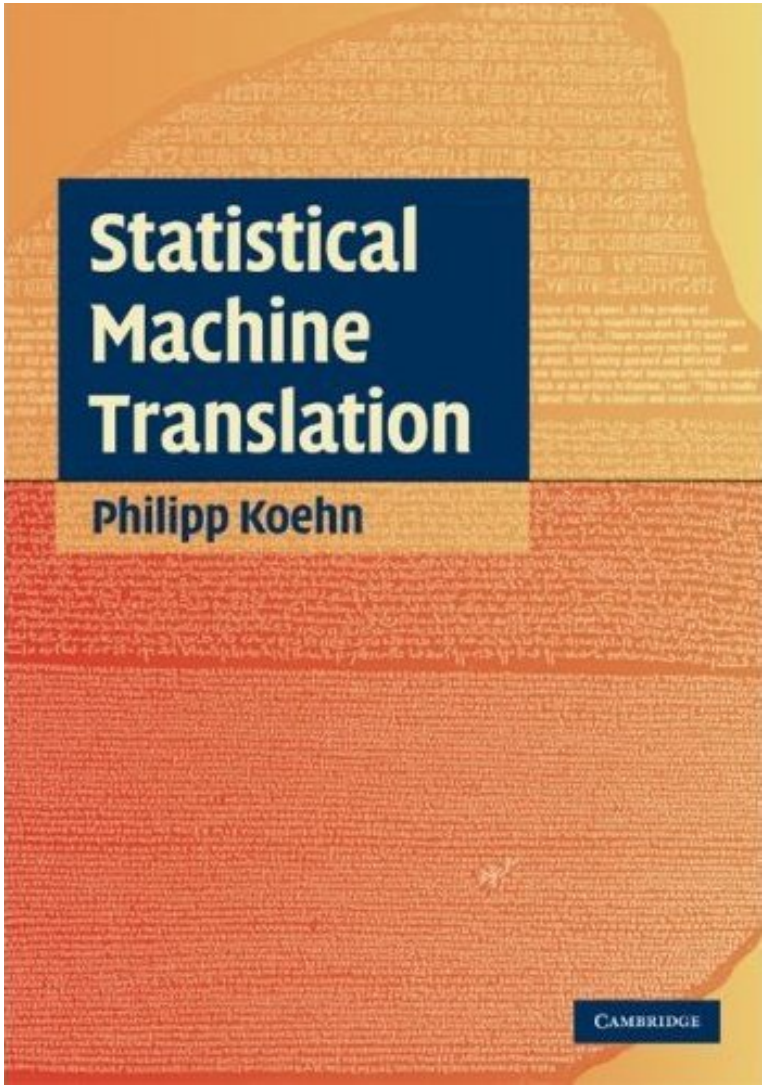
Philipp Koehn

28 January 2016

Administrativa

- Class web site: `http://www.mt-class.org/jhu/`
- Tuesdays and Thursdays, 1:30-2:45, Hodson 313
- Instructor: Philipp Koehn (with help from Matt Post)
- Grading
  - five programming assignments (12% each)
  - final project (30%)
  - in-class presentation: language in ten minutes (10%)

Textbook


Statistical
Machine
Translation
Philipp Koehn
CAMBRIDGE

Machine Translation: Chinese

因为一项2011年赤字削减协议，如果无法与共和党达成折中方案，奥巴马总统可能在年底面临联邦预算自动减少1000多亿美元的局面。在外交政策辩论中，奥巴马说，他的军事预算不会“减少”而将“维持”。

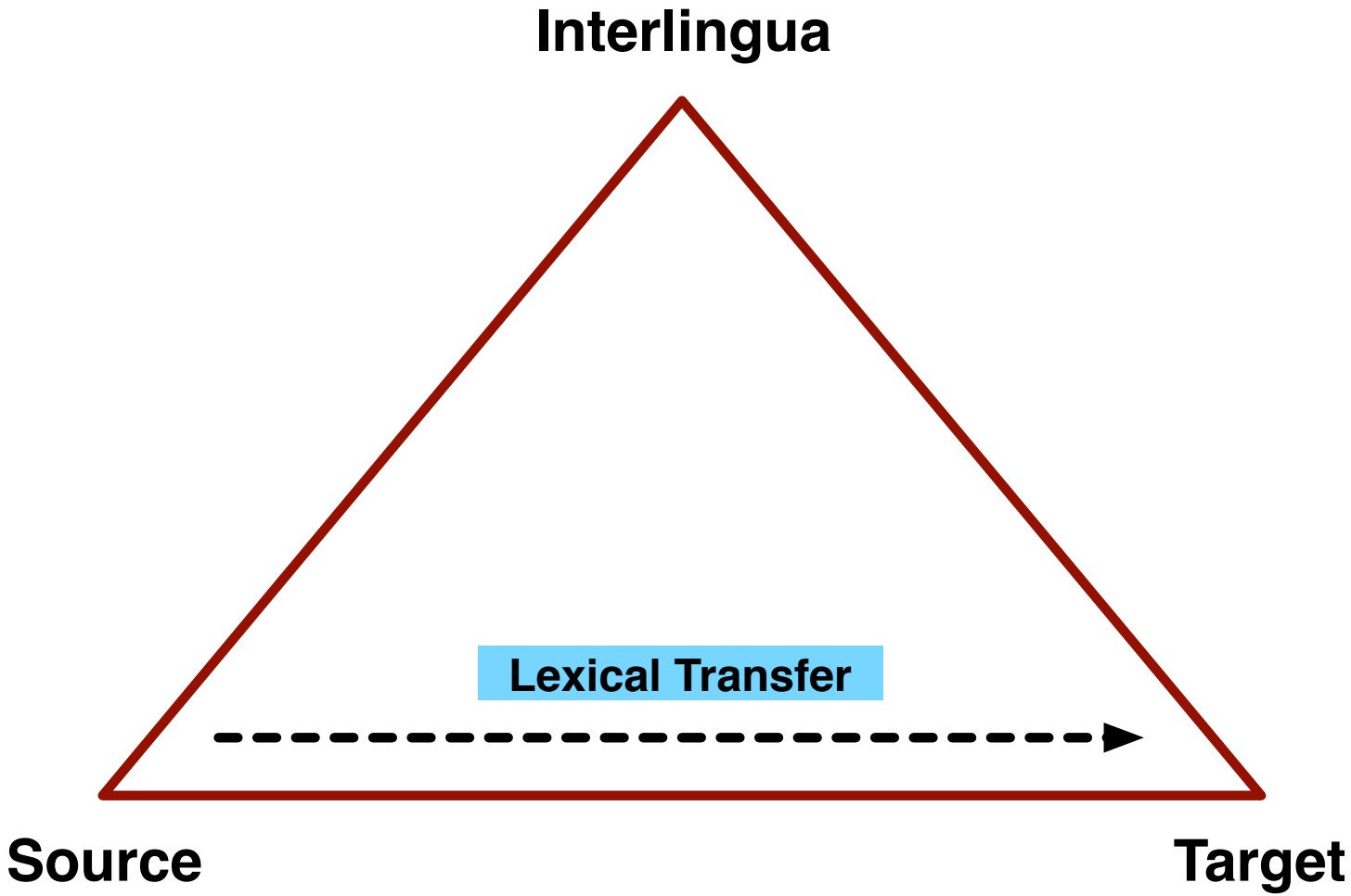
A 2011 deficit reduction agreement, if a compromise can not be reached with the Republican Party, President Obama may face at the end of the federal budget situation automatically reduced by more than 1000 billion dollars. In the foreign policy debate, Obama said, his military budget will not "reduce" and "maintain".

Machine Translation: French

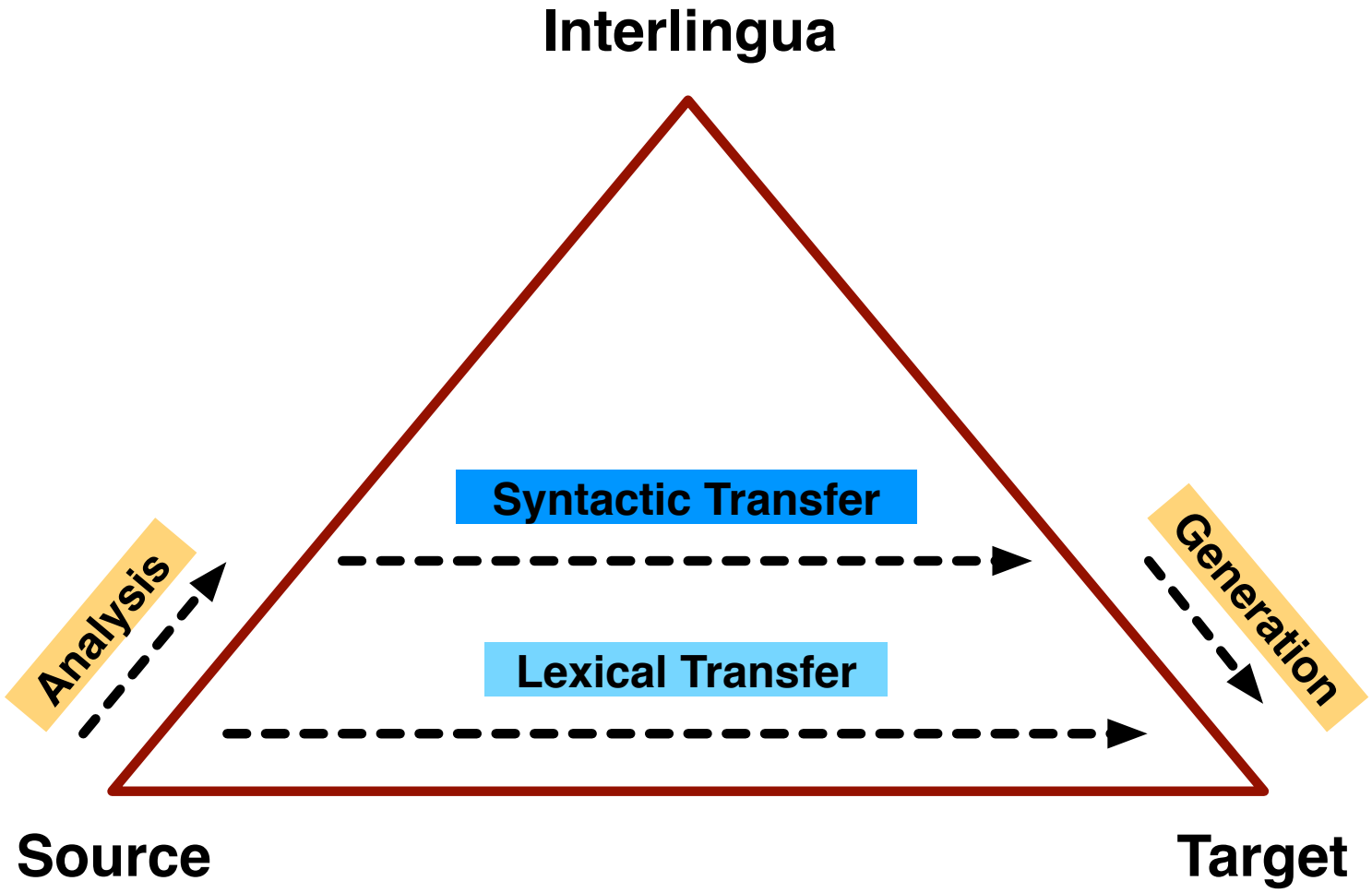
Obama et Romney prévoient de mener campagne dans les «swing states» à un rythme effréné pour les quatre derniers jours avant l'élection. L'Ohio se présente comme l'Etat le plus disputé du pays.

Obama and Romney plan to campaign in the "swing states" at a breakneck pace for the last four days before the election. The Ohio State presents itself as the most played country.

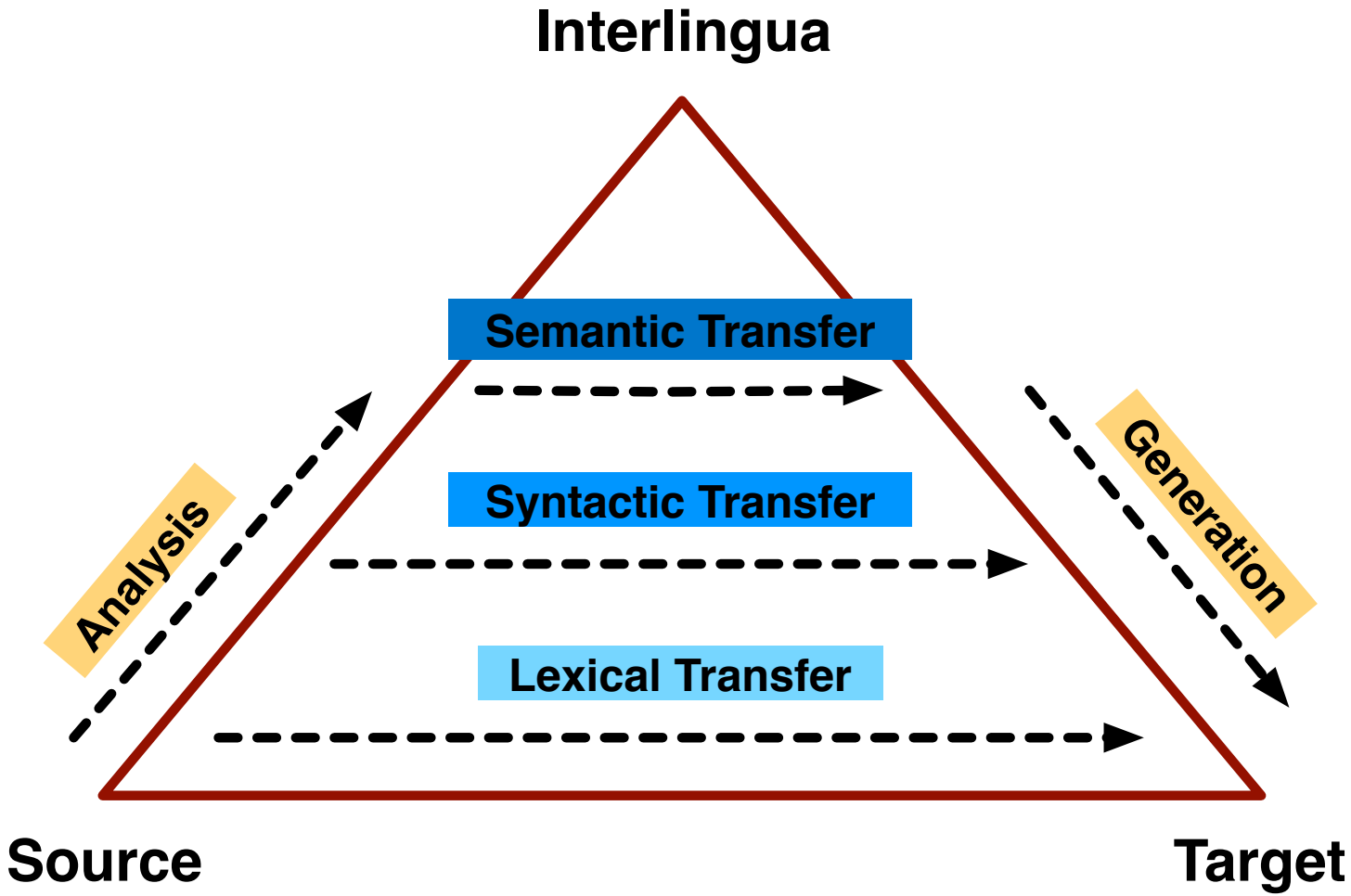
A Clear Plan



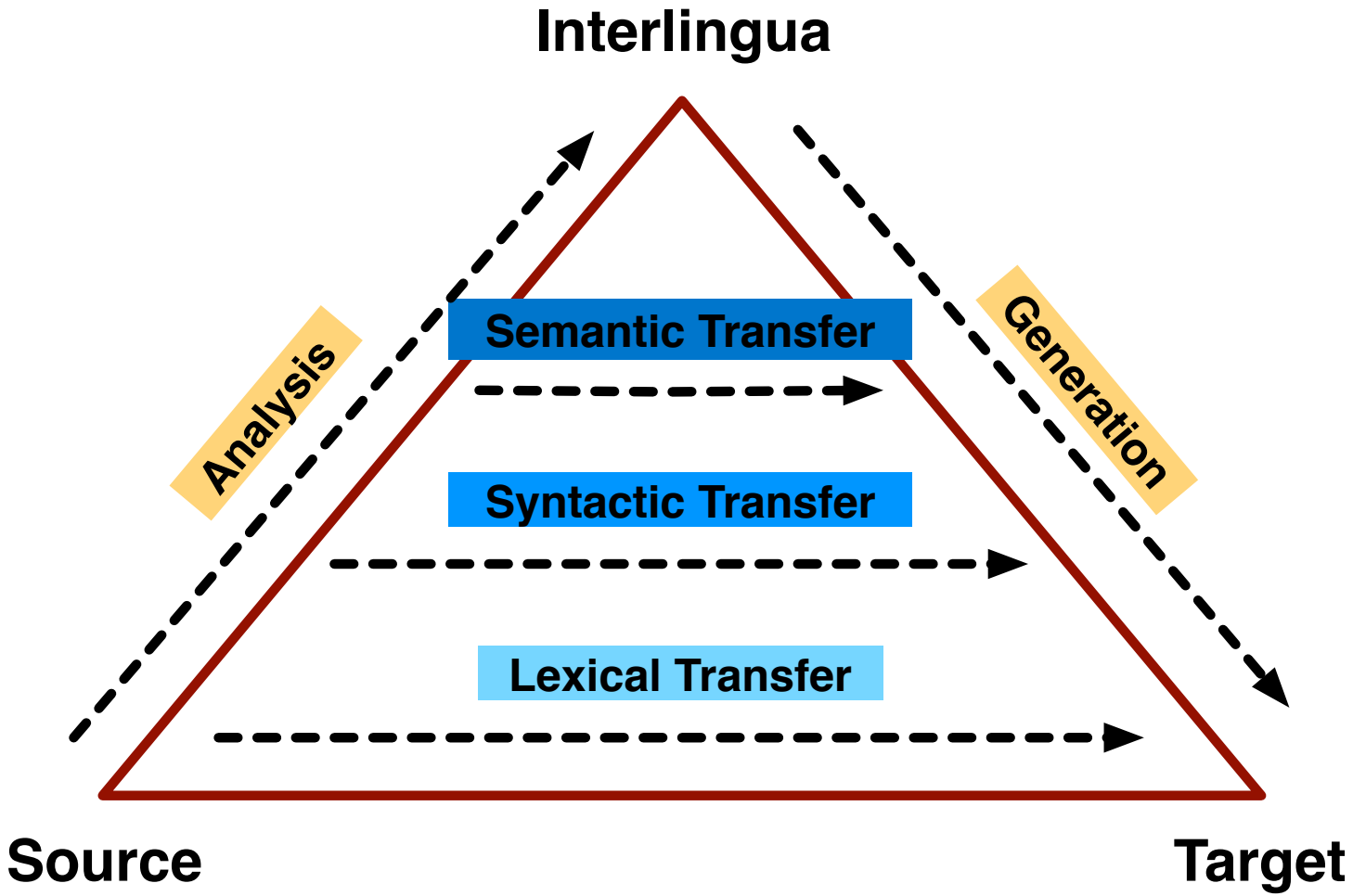
A Clear Plan


Interlingua
Syntactic Transfer
Analysis
Generation
Lexical Transfer
Source
Target

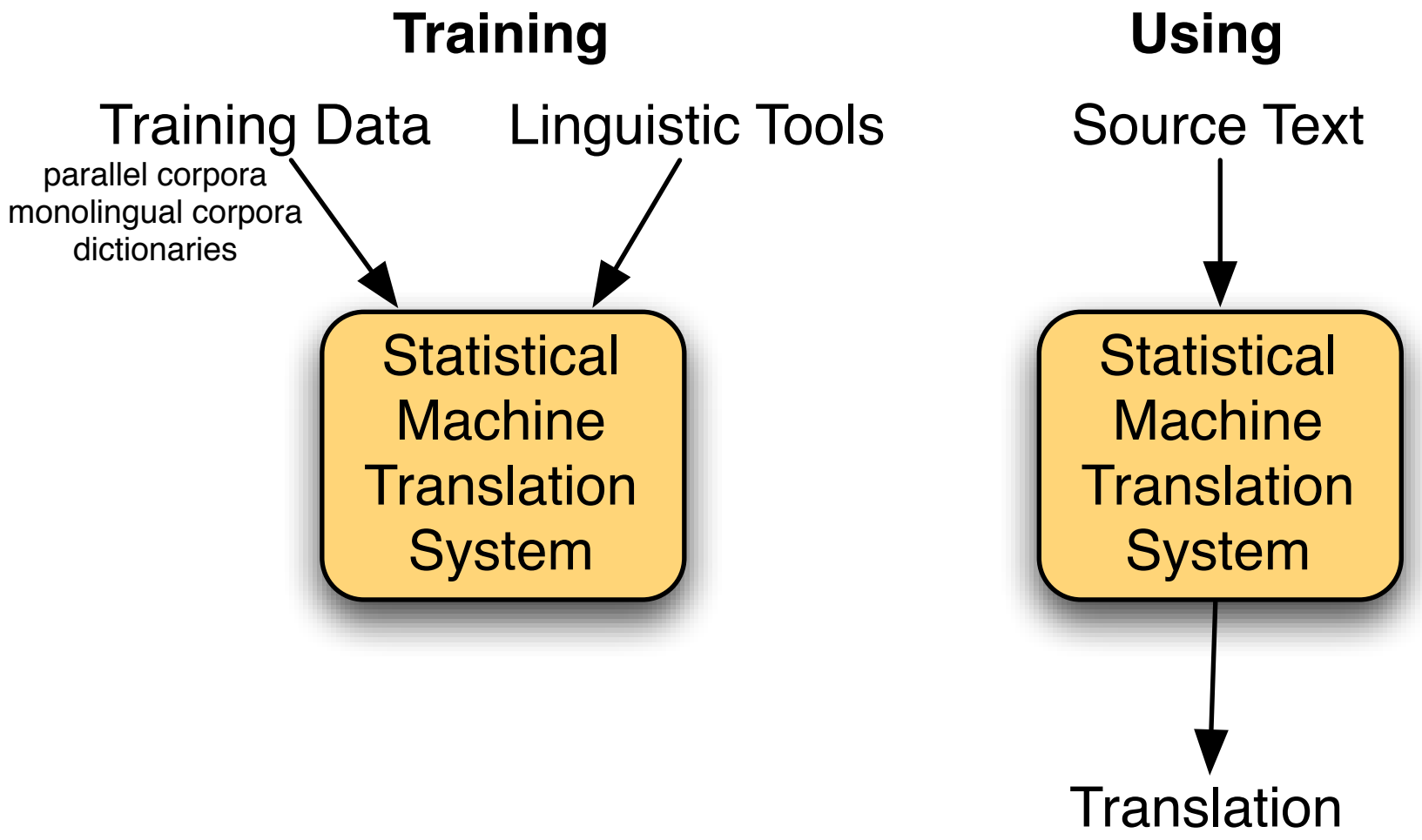
A Clear Plan


Interlingua
Semantic Transfer
Syntactic Transfer
Lexical Transfer
Analysis
Generation
Source
Target

A Clear Plan


Interlingua
Analysis
Semantic Transfer
Generation
Syntactic Transfer
Lexical Transfer
Source
Target

Learning from Data



why is that a good plan?

Word Translation Problems

- Words are ambiguous

He deposited money in a bank account
with a high interest rate.

Sitting on the bank of the Mississippi,
a passing ship piqued his interest.

- How do we find the right meaning, and thus translation?
- Context should be helpful

Syntactic Translation Problems

- Languages have different sentence structure

das	behaupten	sie	wenigstens
this	claim	they	at least
the		she	

- Convert from object-verb-subject (OVS) to subject-verb-object (SVO)

- Ambiguities can be resolved through syntactic analysis
  - the meaning the of das not possible (not a noun phrase)
  - the meaning she of sie not possible (subject-verb agreement)

Semantic Translation Problems

- Pronominal anaphora

I saw the movie and it is good.

- How to translate it into German (or French)?
  - it refers to movie
  - movie translates to Film
  - Film has masculine gender
  - ergo: it must be translated into masculine pronoun er

- We are not handling this very well [Le Nagard and Koehn, 2010]

Semantic Translation Problems

- Coreference

Whenever I visit my uncle and his daughters,
I can't decide who is my favorite cousin.

- How to translate cousin into German? Male or female?

- Complex inference required

Semantic Translation Problems

- Discourse

Since you brought it up, I do not agree with you.

Since you brought it up, we have been working on it.

- How to translated since? Temporal or conditional?

- Analysis of discourse structure — a hard problem

Learning from Data

- What is the best translation?

Sicherheit → security
Sicherheit → safety
Sicherheit → certainty

Learning from Data

- What is the best translation?

Sicherheit → security 14,516
Sicherheit → safety 10,015
Sicherheit → certainty 334

- Counts in European Parliament corpus

Learning from Data

- What is the best translation?

Sicherheit → security 14,516
Sicherheit → safety 10,015
Sicherheit → certainty 334

- Phrasal rules

Sicherheitspolitik → security policy 1580
Sicherheitspolitik → safety policy 13
Sicherheitspolitik → certainty policy 0

Lebensmittelsicherheit → food security 51
Lebensmittelsicherheit → food safety 1084
Lebensmittelsicherheit → food certainty 0

Rechtssicherheit → legal security 156
Rechtssicherheit → legal safety 5
Rechtssicherheit → legal certainty 723

Learning from Data

- What is most fluent?

a problem for translation

a problem of translation

a problem in translation

Learning from Data

- What is most fluent?

a problem for translation 13,000

a problem of translation 61,600

a problem in translation 81,700

- Hits on Google

Learning from Data

- What is most fluent?

a problem for translation 13,000

a problem of translation 61,600

a problem in translation 81,700

a translation problem 235,000

Learning from Data

- What is most fluent?

police disrupted the demonstration

police broke up the demonstration

police dispersed the demonstration

police ended the demonstration

police dissolved the demonstration

police stopped the demonstration

police suppressed the demonstration

police shut down the demonstration

Learning from Data

- What is most fluent?

police disrupted the demonstration 2,140

police broke up the demonstration 66,600

police dispersed the demonstration 25,800

police ended the demonstration 762

police dissolved the demonstration 2,030

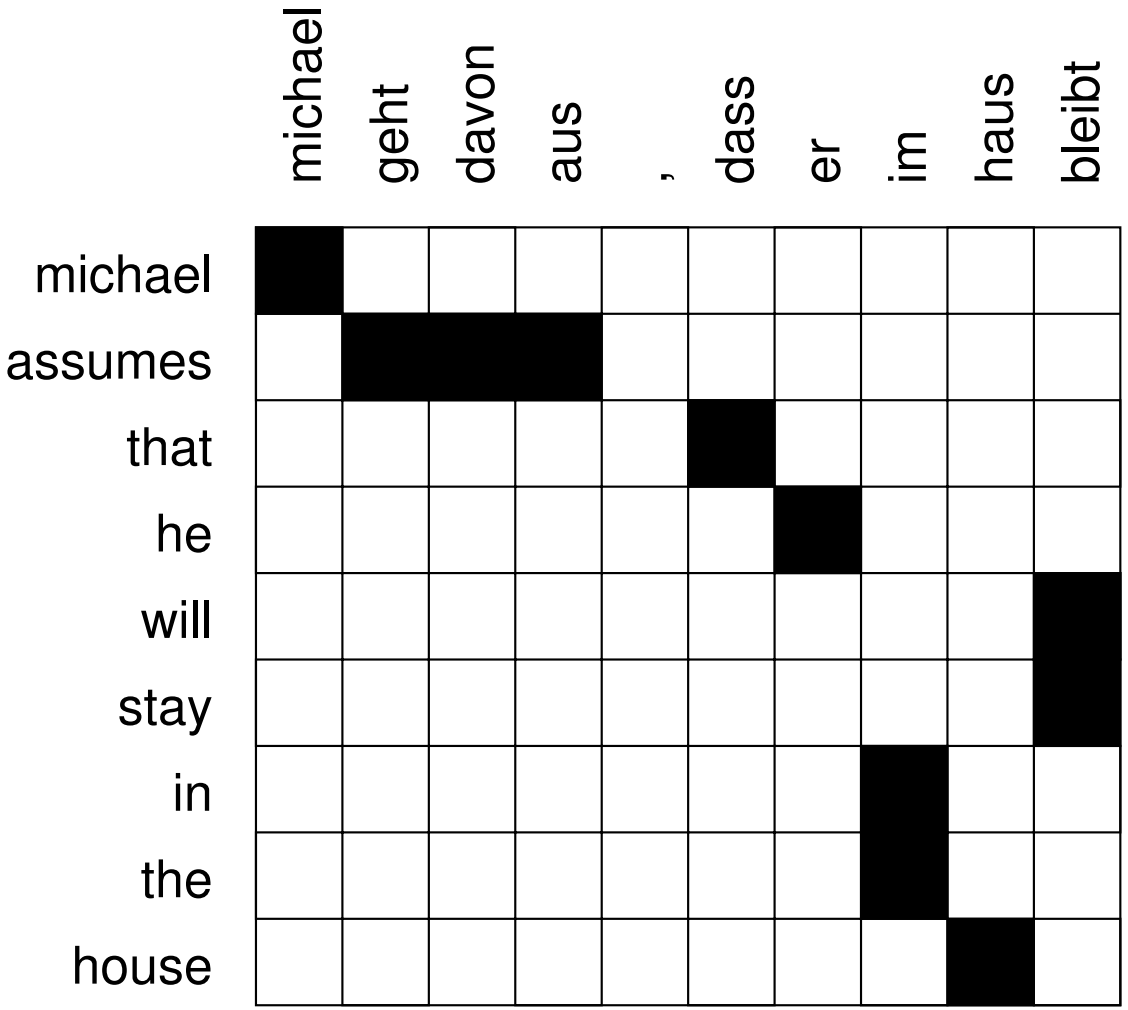
police stopped the demonstration 722,000

police suppressed the demonstration 1,400

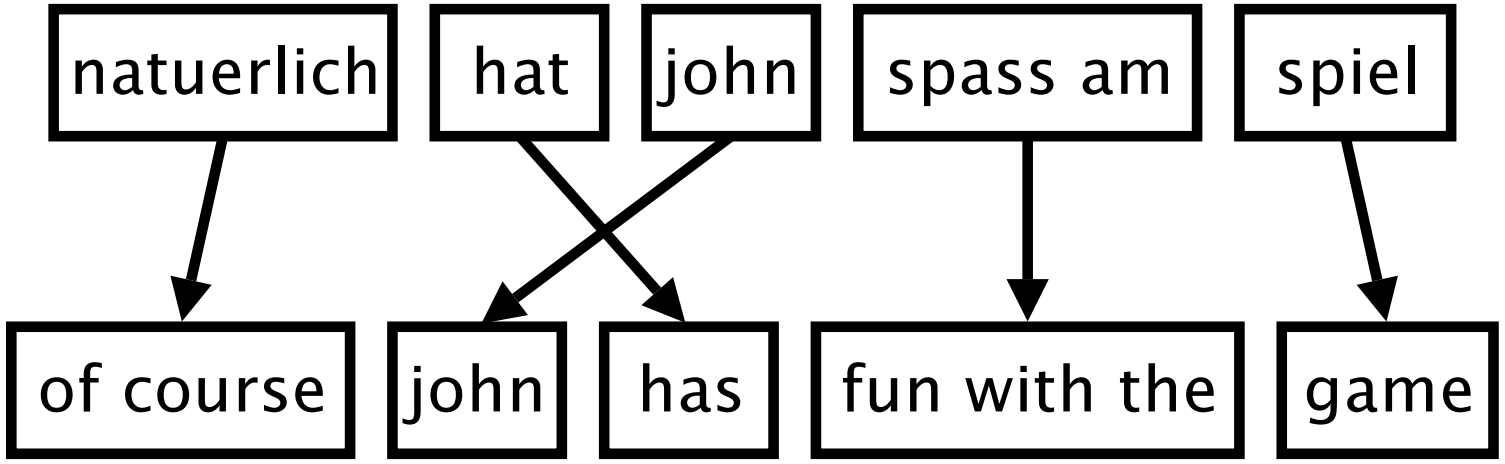
police shut down the demonstration 2,040

where are we now?

Word Alignment



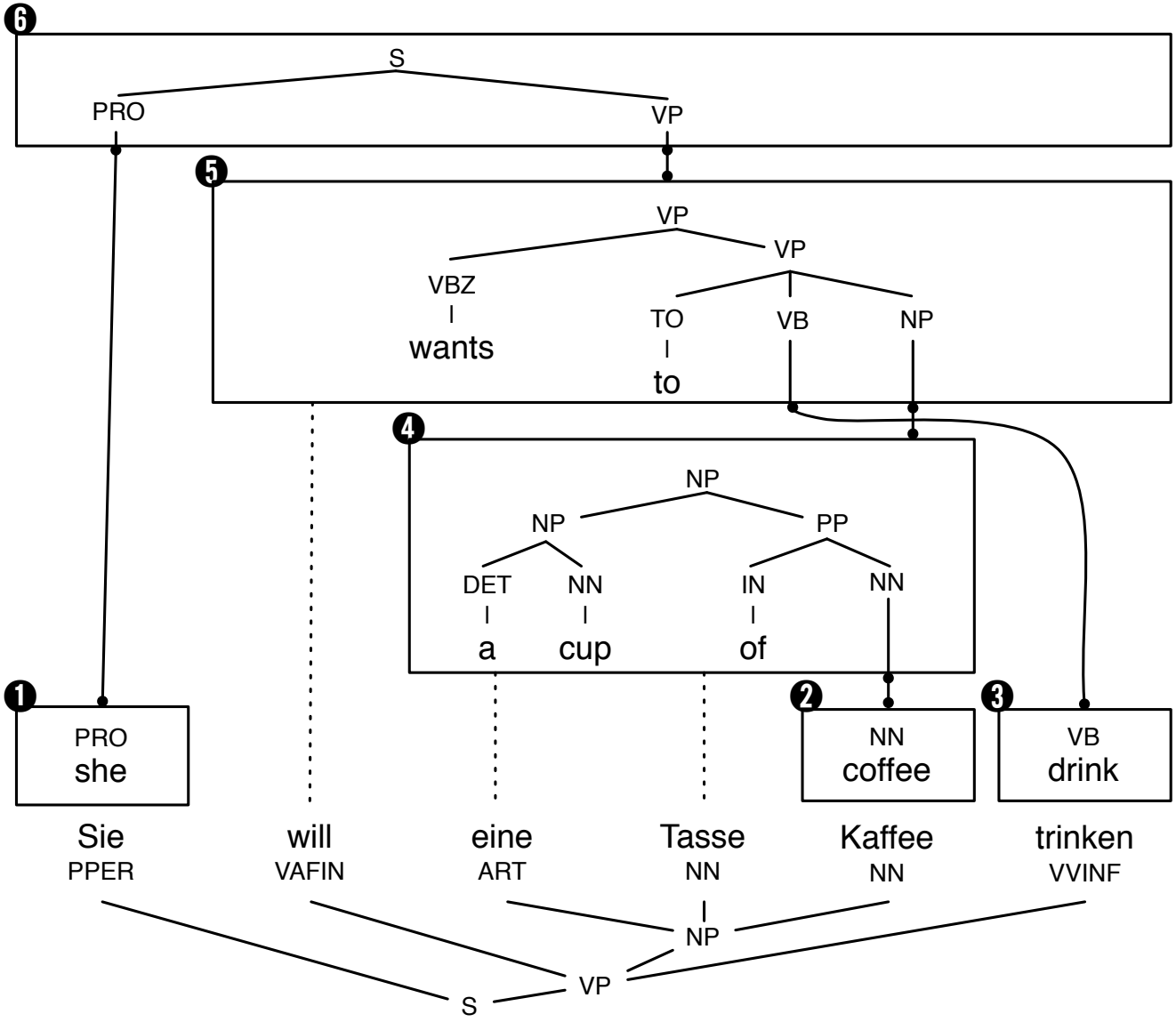
Phrase-Based Model



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

- Workhorse of today's statistical machine translation

Syntax-Based Translation



Semantic Translation

- Abstract meaning representation [Knight et al., ongoing]

```
(w / want-01
   :agent (b / boy)
   :theme (l / love
        :agent (g / girl)
        :patient b))
```

- Generalizes over equivalent syntactic constructs
  (e.g., active and passive)

- Defines semantic relationships
  - semantic roles
  - co-reference
  - discourse relations

- In a very preliminary stage

what is it good for?

what is it good *enough* for?

Why Machine Translation?

Assimilation — reader initiates translation, wants to know content

- user is tolerant of inferior quality
- focus of majority of research (GALE program, etc.)

Communication — participants don't speak same language, rely on translation

- users can ask questions, when something is unclear
- chat room translations, hand-held devices
- often combined with speech recognition, IWSLT campaign

Dissemination — publisher wants to make content available in other languages

- high demands for quality
- currently almost exclusively done by human translators

Problem: No Single Right Answer

这个 机场 的 安全 工作 由 以色列 方面 负责 .

Israeli officials are responsible for airport security.
Israel is in charge of the security at this airport.
The security work for this airport is the responsibility of the Israel government.
Israeli side was in charge of the security of this airport.
Israel is responsible for the airport's security.
Israel is responsible for safety work at this airport.
Israel presides over the security of the airport.
Israel took charge of the airport security.
The safety of this airport is taken charge of by Israel.
This airport's security is the responsibility of the Israeli security officials.

Quality

HTER	**assessment**
0%	
	publishable
10%	
	editable
20%	
30%	gistable
40%	triagable
50%	

(scale developed in preparation of DARPA GALE programme)

Applications

HTER	assessment	application examples
0%		Seamless bridging of language divide
	publishable	Automatic publication of official announcements
10%		
	editable	Increased productivity of human translators
20%		Access to official publications
		Multi-lingual communication (chat, social networks)
30%	gistable	Information gathering
		Trend spotting
40%	triagable	Identifying relevant documents
50%		

Current State of the Art

HTER	assessment	language pairs and domains
0%		
	publishable	French-English restricted domain
10%		French-English technical document localization
	editable	French-English news stories
20%		
		English-German news stories
30%	gistable	English-Czech open domain
40%	triagable	
50%		

(informal rough estimates by presenter)

Thank You

questions?