UPPSALA
UNIVERSITET

Machine Translation
5LN426 and 5LN711

Sara Stymne
Uppsala University

Slides mainly from Jörg Tiedemann

UPPSALA UNIVERSITET

Outline for Today

Motivation

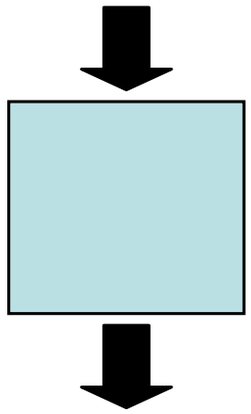
Overview of the course

Classical MT approaches

UPPSALA UNIVERSITET

Machine Translation

美国关岛国际机场及其办公室均接获一名自称沙地阿拉伯富商拉登等发出的电子邮件，威胁将会向机场等公众地方发动生化袭击後，关岛经保持高度戒备。

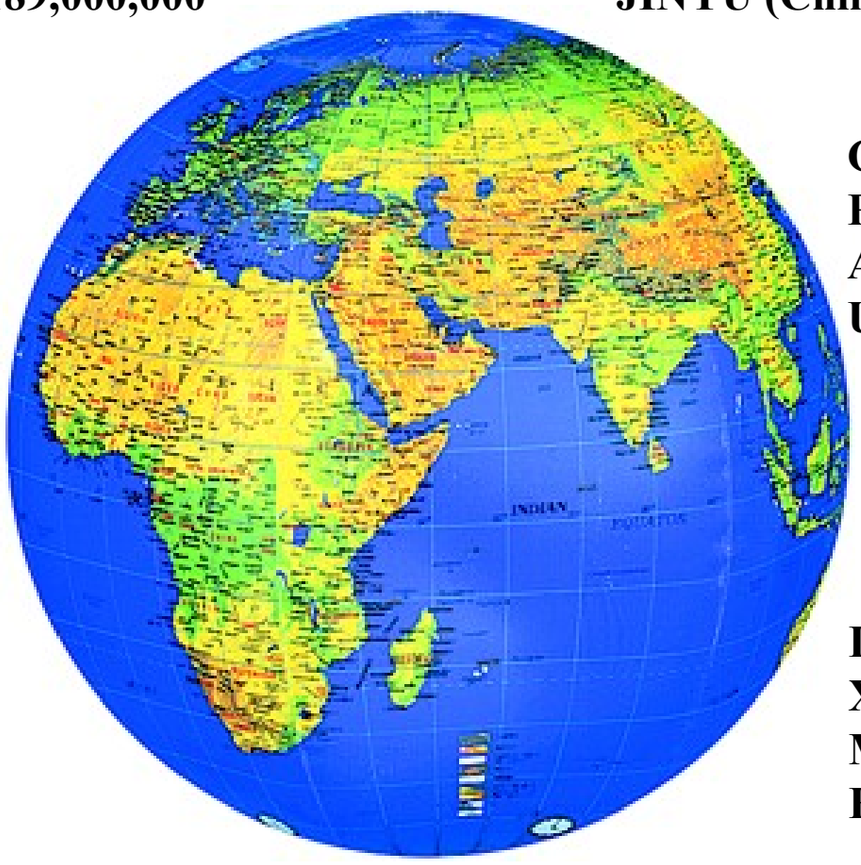


The U.S. island of Guam is maintaining a high state of alert after the Guam airport and its offices both received an e-mail from someone calling himself the Saudi Arabian Osama bin Laden and threatening a biological/chemical attack against public places such as the airport.

UPPSALA UNIVERSITET

Why Machine Translation?

MANDARIN	885,000,000
SPANISH	332,000,000
ENGLISH	322,000,000
BENGALI	189,000,000
HINDI	182,000,000
PORTUGUESE	170,000,000
RUSSIAN	170,000,000
JAPANESE	125,000,000
GERMAN	98,000,000
WU (China)	77,175,000
JAVANESE	75,500,800
KOREAN	75,000,000
FRENCH	72,000,000
VIETNAMESE	67,662,000
TELUGU	66,350,000
YUE (China)	66,000,000
MARATHI	64,783,000
TAMIL	63,075,000

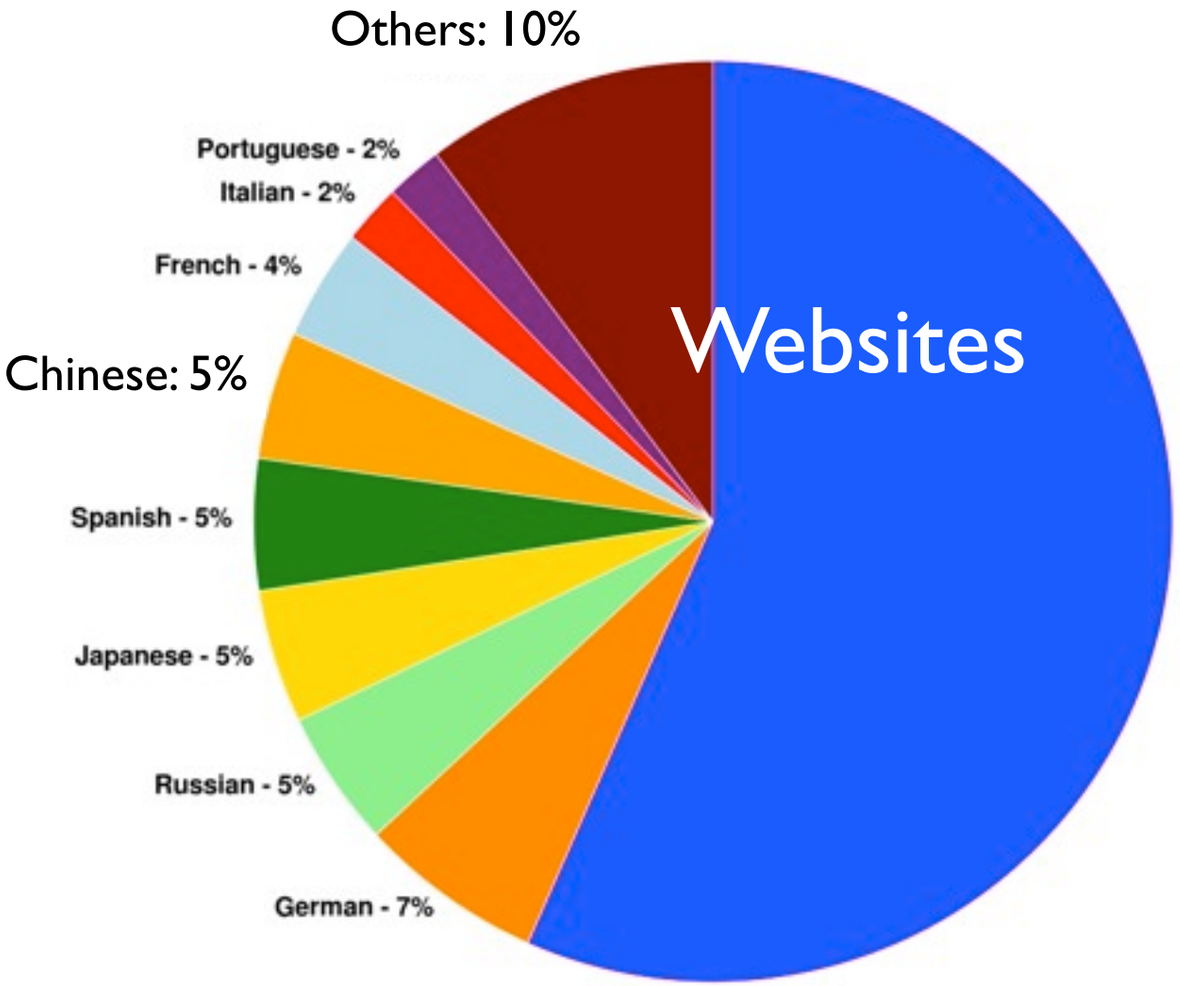


TURKISH	59,000,000
URDU	58,000,000
MIN NAN (China)	49,000,000
JINYU (China)	45,000,000
GUJARATI	44,000,000
POLISH	44,000,000
ARABIC	42,500,000
UKRAINIAN	41,000,000
ITALIAN	37,000,000
XIANG (China)	36,015,000
MALAYALAM	34,022,000
HAKKA (China)	34,000,000
KANNADA	33,663,000
ORIYA	31,000,000
PANJABI	30,000,000
SUNDA	27,000,000

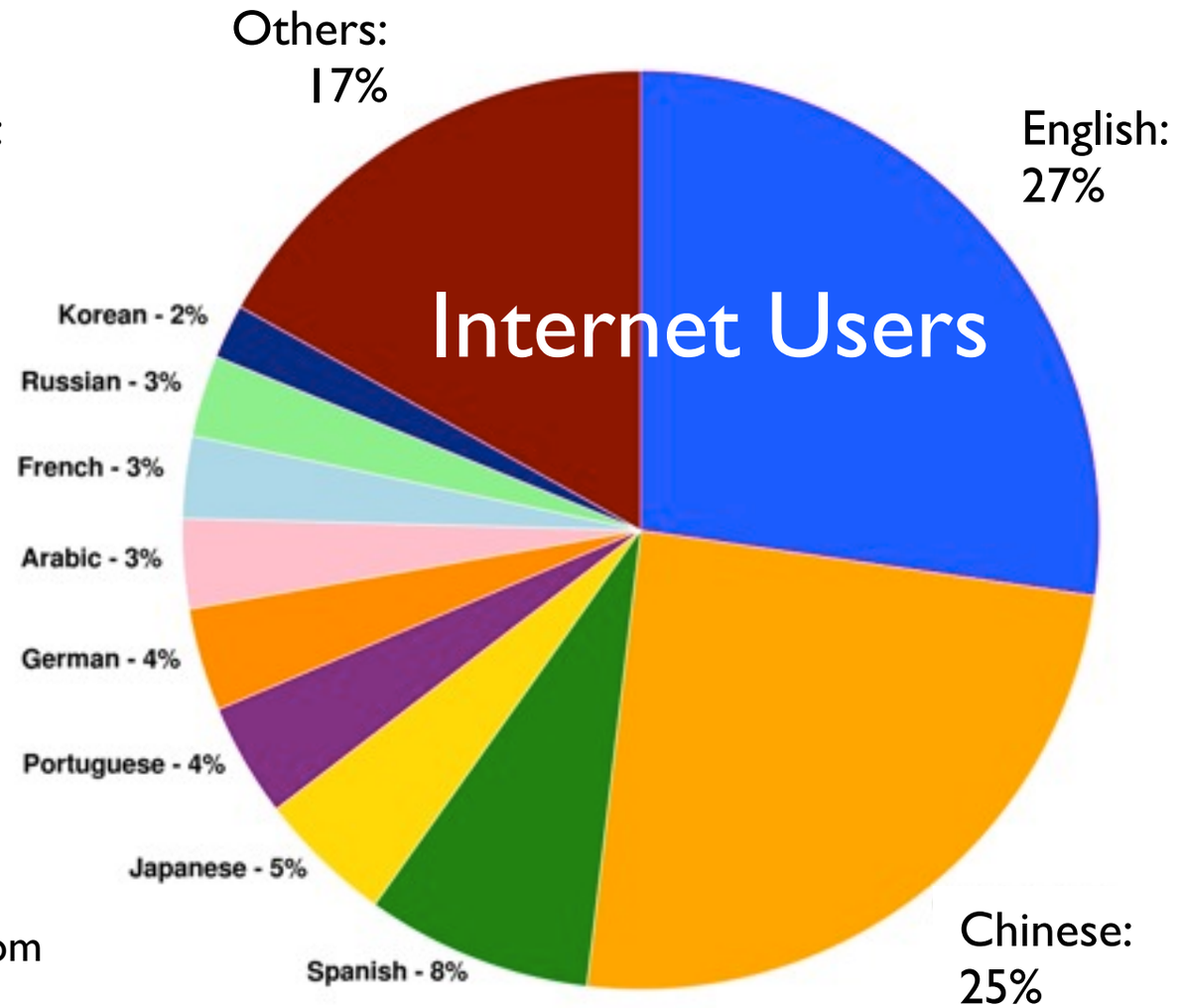
Source: Ethnologue

UPPSALA UNIVERSITET

Why Machine Translation?



> 2 billion Internet users
> 550 million registered domains
> 12 billion indexed web pages



Sources: W3Techs.com, Internet World Stats, WorldWideWebSize.com

UPPSALA UNIVERSITET

Why Machine Translation?

- Translation is expensive
- On-line demand for translation (on-the-fly)
- Globalization, growing export
- Lots of language pairs
- Political issues (UN, EU, minority languages, ...)
- Tourism, movies, news
- ...

UPPSALA UNIVERSITET

MT is a Tough Challenge (and Fun)

Translation errors may be quite severe:

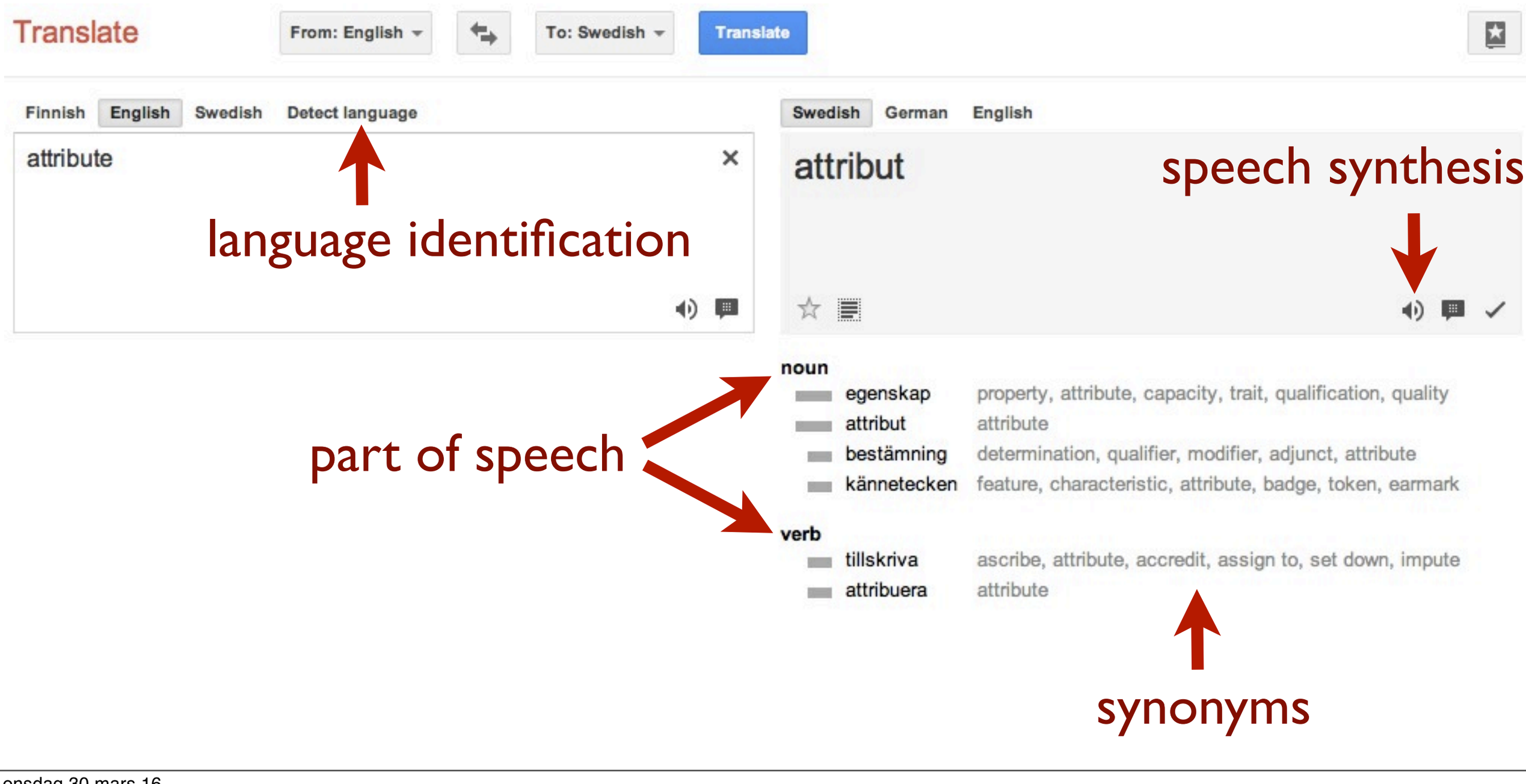
- Doctor's office: *Specialist in women and other diseases*
- Pub: *Ladies are requested not to have children in the bar*
- Hotel: *Please leave your values at the front desk*
- Chinese dining hall: *Translation server error*

MT is not a solved problem .... but constantly improves?

- Input: Vem vann Allsvenskan i fjol?
- Google 2010: Who stole headlines last year?
- Google 2013: Who won the Championship last year?
- Google 2016: Who won the Olympics last year?

UPPSALA UNIVERSITET

MT and Other Language Technology



UPPSALA UNIVERSITET

MT is a Cool Research Topic

How does human language work?

- What are the differences between languages?
- How can we preserve meaning when translating?

Complex but natural task

- MT is not a solved problem
- MT is a useful end-user application

Combines various aspects of computational linguistics

- analyze text or speech
- understand/transfer meaning
- generate text or speech

UPPSALA UNIVERSITET

What is the Problem with MT?

Unrealistic expectations

- "MT is a waste of time because you will never make a machine that can translate Shakespeare"
- MT is useless because it may translate *"The spirit is willing but the flesh is weak"* into the Russian equivalent of *"The vodka is good, but the steak is lousy"*

Unexpected (not humanlike) errors

- German Input: *Fussball ist langweilig.* ***Tore*** *gibt es selten.*
- Google 2012: *Fotboll är tråkigt.* ***Gates*** *är sällsynta.*
- Google 2016: *Fotboll är tråkigt .* ***Mål*** *är sällsynta .*

UPPSALA
UNIVERSITET

What are the problems?

- Source language ambiguity
- Cross-lingual divergences
- Target language variation

UPPSALA UNIVERSITET

Source language ambiguity

Get (English)

- *I'll* ***get*** *a cup of coffee*
- *I didn't* ***get*** *the joke*
- *I* ***get*** *up at 8am*
- *I* ***get*** *nervous*
- *Yeah, I* ***get*** *around*

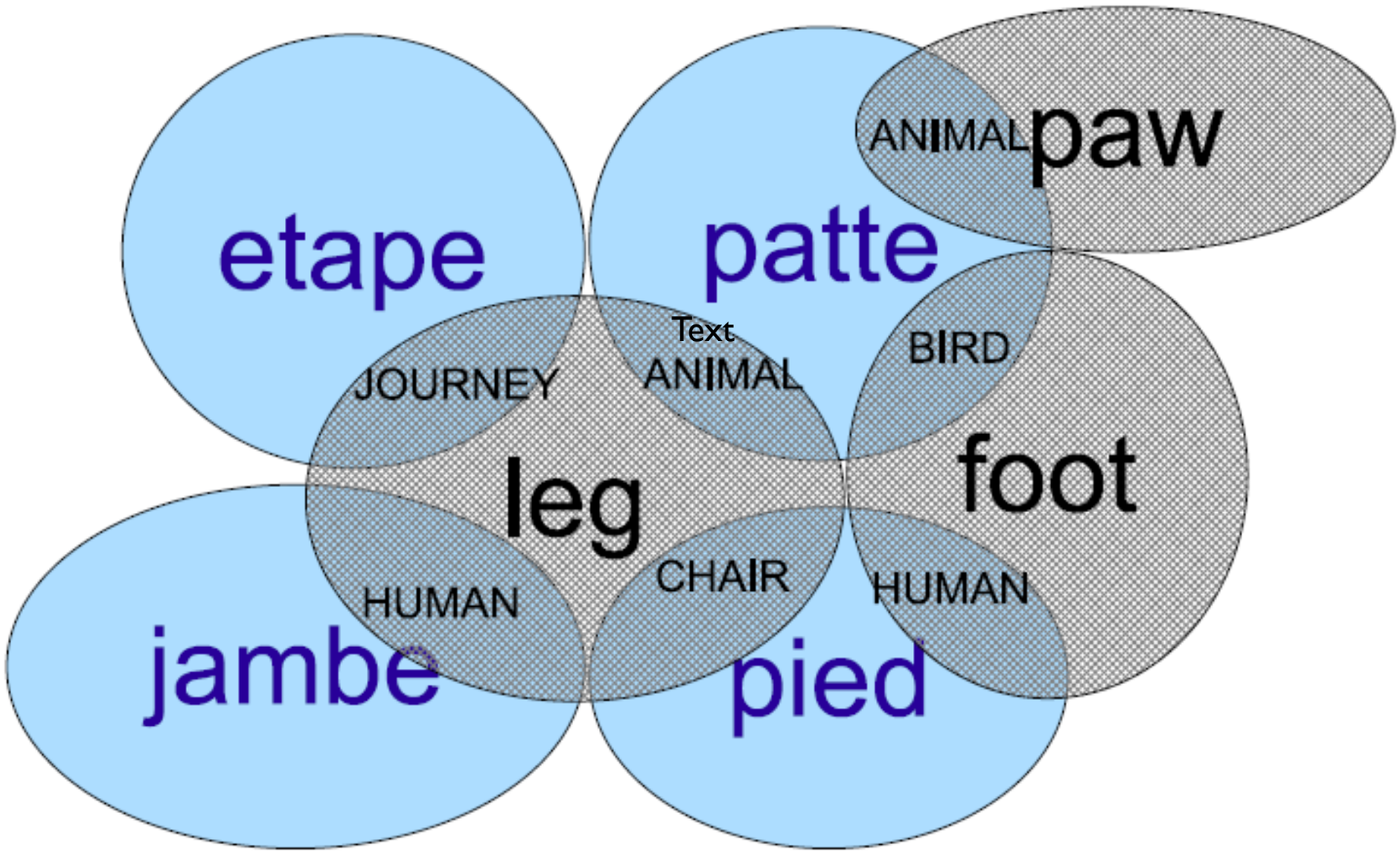
Var (Swedish)

- was, were (verb)
- each, every (pron)
- where, apiece (adv)
- pus (noun)

> Ambiguity is usually solved in context

UPPSALA
UNIVERSITET

Lexical Ambiguities across Languages



From Jurafsky and Martin

UPPSALA UNIVERSITET

Language Divergences and Mismatches

Systematic differences between the 2 languages

- morphology (isolating vs polysynthetic, agglutinative vs fusional)
- syntax (SVO, SOV, VSO, argument structure, pro-drop)

Idiosyncratic and lexical differences

- differences in lexical ambiguity
- lexical gaps
- differences in tempus, aspect, voice
- different idiomatic/fixed expressions
- ...

UPPSALA UNIVERSITET

Verb Frame Divergences

Categorial

- *Kim var förkyld -- Kim had a cold*

Conflation

- *Kim snyter sig -- Kim blows her nose*

Structural

- *Kim sätter sig upp mot Bo -- Kim defies Bo*

Head swapping

- *Kim packar klart -- Kim finishes packing*

Thematic

- *Me gustan uvas -- I like grapes*

UPPSALA UNIVERSITET

Variation in Target Language

Redundancy of natural languages

- *translate "Vid avslutad kurs ..."*
  - *On completion of the course ...*
  - *After completion of the course ...*
  - *Having completed the course ...*
  - *After finishing the course ...*
  - *Once the course has been completed ...*
  - ...

Which one is best? How do we decide that?

UPPSALA UNIVERSITET

In-domain MT with Related Languages

Example from the book:

French input

Nous savons trés bien que les Traités actuels ne susent pas et qu'il sera nécessaire à l'avenir de développer une structure plus ecace et différente pour l'Union, une structure plus constitutionnelle qui indique clairement quelles sont les compétences des Èta ts membres et quelles sont les compétences de l'Union.

Statistical machine translation

We know very well that the current treaties are not enough and that in the future it will be necessary to develop a different and more effective structure for the union, a constitutional structure which clearly indicates what are the responsibilities of the member states and what are the competences of the union.

Human translation

We know all too well that the present Treaties are inadequate and that the Union will need a better and different structure in future, a more constitutional structure which clearly distinguishes the powers of the Member States and those of the Union.

UPPSALA UNIVERSITET

MT Between Less Related Languages

Also from the book:

Chinese input

伦敦每日快报指出,两台记载黛安娜王妃一九九七年巴黎死亡车祸调查资料的手提电脑,被从前大都会警察总长的办公室里偷走.

Statistical machine translation

The London Daily Express pointed out that the death of Princess Diana in 1997 Paris car accident investigation information portable computers, the former city police chief in the offices of stolen.

Human translation

London's Daily Express noted that two laptops with inquiry data on the 1997 Paris car accident that caused the death of Princess Diana were stolen from the office of a former metropolitan police commissioner.

UPPSALA UNIVERSITET

How do Humans do it?

Human translators need

- to **understand** the source language
- to know how to **speak** the target language (well)
- knowledge about the **topic** of the text to be translated
- knowledge about culture, values, traditions and expectations of speakers in both languages

Corresponding NLP challenges

- Natural language understanding
- Language generation
- Topic detection and domain adaptation

UPPSALA UNIVERSITET

Is it Possible at all?

Balance MT quality and input restrictions, depending on task

general purpose	post-editing	sublanguage
browsing quality	editing quality	publishing quality
fully automatic **Gisting**	computer-aided translation (**CAT/ MT**)	fully automatic **FAHQMT**
on-line service	localization, ...	domain-specific tasks

UPPSALA UNIVERSITET

What exactly is MT?

- MT = automatic translation from one language (source language) to another (target language) using computers
- MT ≠ translation memories and bilingual dictionaries
- MT - usually sentence-by-sentence translation
- MT often refers to translation of written text (cf speech-to-speech translation)
- Semi-automatic: CAT = computer aided translation

UPPSALA UNIVERSITET

Computer-Assisted Translation Tools

A range of tools to support translators

Translation memories

- A database that stores previously translated sentences/ segments
- When translating a new segment, it searches for a matching segment, to display
- Fuzzy matching, it finds similar segments if no full match, and highlights the differences
  - The translator edits this segment, if good enough
  - A score is shown that indicates how similar the matched segment is
- Some TM software has integration with MT

UPPSALA
UNIVERSITET

Course overview

UPPSALA UNIVERSITET

Course Overview (5LN426, 5LN711)

Lectures

- Introduction of main MT approaches
- MT Evaluation
- Basics of Statistical MT and Word-Based Models
- Phrase-Based SMT
- Tree-based SMT; Document-Level Models
- Seminars: Advanced topics in SMT
  given by master students

Labs

- Practical sessions and assignments
- 4 written reports, 1 oral
- Performed in pairs, signup by email to Sara and Aaron

UPPSALA UNIVERSITET

Course Overview (5LN426, 5LN711)

Bachelor students

- 3 individual assignments
  - 2 assignments on MT/SMT, will be given at least 2 weeks before each deadline
  - 1 literature assignment, choose and summarize a research article

Master students

- Group project, 3-4 students
- Groups will be created based on wishes for topics to work on
- Give a seminar on your topic
- Perform a practical project and write a report
- Individual reflection report

UPPSALA UNIVERSITET

Examination

Bachelor program

- Lab sessions with assignments, no grades
- Three graded individual reports
- VG:VG on at least 2 individual assignments

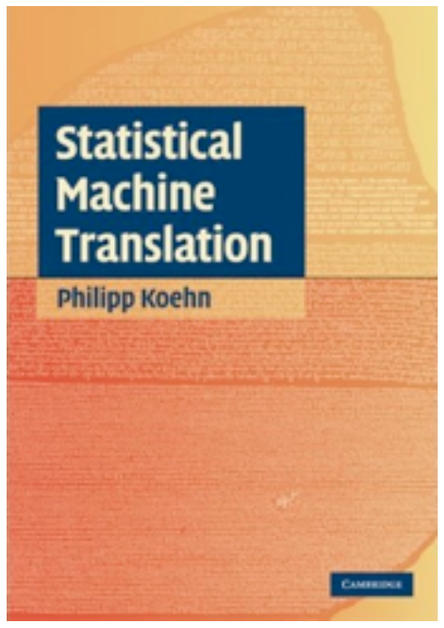
Master program

- Lab sessions with assignments, no grades
- Group work, graded (in total)
- Individual reflection report, graded
- Final grades based on both

UPPSALA UNIVERSITET

Course Information

Website: *http://stp.lingfil.uu.se/~sara/kurser/MT16/*



Literature:

- Philipp Koehn: Statistical Machine Translation
- Daniel Jurafsky and James H. Martin: Speech and language processing
- Other (on-line) material including research articles

Lab sessions:

- STP Linux account

UPPSALA
UNIVERSITET

Teachers

Teachers:

- Sara Stymne (Course coordinator, lectures, lab, supervision)
- Aaron Smith (Lecture, labs)
- Christian Hardmeier (Lectures, supervision)
- Fabienne Cap (Lectures, supervision)
- Mats Dahllöf (Examiner)

Guest lectures:

- Anna Sågvall-Hein, Convertus
  - Machine translation provider
- Nils-Erik Lindström and others, Semantix
  - Translation and interpreting agency

UPPSALA UNIVERSITET

Road Map

Type	Date	Time	Place	Topic	Reading / Assignments
F	2016-03-30	10-12	6-K1031	Introduction (SS)	Koehn 1; JM 25.1-2; Hutchins; CFMF
F	2016-03-30	14-16	6-K1031	MT evaluation (SS)	Koehn 8; JM 25.9
F	2016-04-04	10-12	2-0076	MT in practice (Convertus) - guest lecture	
L	2016-04-06	10-12	Chomsky	MT in practice (AS)	lab report 1
F	2016-04-11	10-12	2-0076	Introduction to SMT (FC)	Koehn Ch 4, Ch 7, KK97
L	2016-04-13	10-12	Chomsky	Word-based SMT (SS)	lab report 2
L	2016-04-18	10-12	Chomsky	Word-based SMT (SS)	lab report 2
F	2016-04-18	14-16	?	Machine translation at Semantix, a translation provider - guest lecture	
F	2016-04-20	10-12	6-K1031	Parallel Corpora, Alignment (AS)	Koehn 2-4, JT 3-4, KK97, KK99
L	2016-04-25	10-12	Chomsky	Parallel corpora & alignment (AS)	lab report 3
F	2016-04-27	10-12	2-0076	Phrase-based SMT (FC)	Koehn Ch 5
L	2016-05-02	10-12	Chomsky	Phrase-based SMT (AS)	lab report 4
F	2016-05-04	10-12	6-K1031	Decoding (CH)	Koehn Ch 6
L	2016-05-09	10-12	Chomsky	Phrase-based SMT (AS)	lab report 4
F	2016-05-11	10-12	2-0076	Tree-based SMT & MT for morphologically rich languages (SS, FC)	Koehn 10.2, 11
F	2016-05-16	10-12	2-0076	Document-wide decoding & Neural MT (CH)	
L	2016-05-18	10-12	Chomsky	Document-wide decoding lab (AS)	oral lab report 5
S	2016-05-23	10-12	2-0076	Seminar - master student presentations	
S	2016-05-25	10-12	6-K1031	Seminar - master student presentations	

UPPSALA UNIVERSITET

Deadlines

Deadlines (All, Master, Bachelor)

- April 13, lab 1 (A)
- April 15, topic selection (M)
- April 26, lab 2 (A)
- May 3, lab 3 (A)
- May 10, ass. 1 (B)
- May 16, ass 2a (B)
- May 17, lab 4 (A)
- May 18/25, lab 5 (A)
- May 23/25, seminar presentation (M)
- June 3, ass. 2b+3 (B)
- June 3, project report (M)
- June 3, reflection report (M)
- June 3, backup lab deadline (A)

Note that there are also many deadlines in the information extraction course! You need to plan your time carefully!

UPPSALA UNIVERSITET

Changes from last year

Changes

- New course coordinator and one new teacher: practical reasons
- Additional guest lecture: course improvement
- Labs in pairs: practical reasons, and better interaction
- Bachelors:
  - Assignments instead of exam: based on evaluation and better fit with syllabus
- Masters
  - Group project instead of individual project: practical reasons, better chance for interaction between students, chance to do more advanced projects

UPPSALA
UNIVERSITET

Classical Translation Models

UPPSALA UNIVERSITET

MT as Decoding (”Engineer’s MT”)

When I look at an article in Russian, I say:
This is really written in English,
but it has been coded in some strange symbols.
I will now proceed to decode.

[Weaver, 1947, 1949]

UPPSALA UNIVERSITET

Linguistically Motivated MT

Assume that human translators work like this:

- Understand the original message
- Transfer meaning to the target language context
- Produce a grammatical message in the target language

Transfer-based Machine Translation

- Source language analysis
- Transfer abstract representation
- Generation of target language text

UPPSALA UNIVERSITET

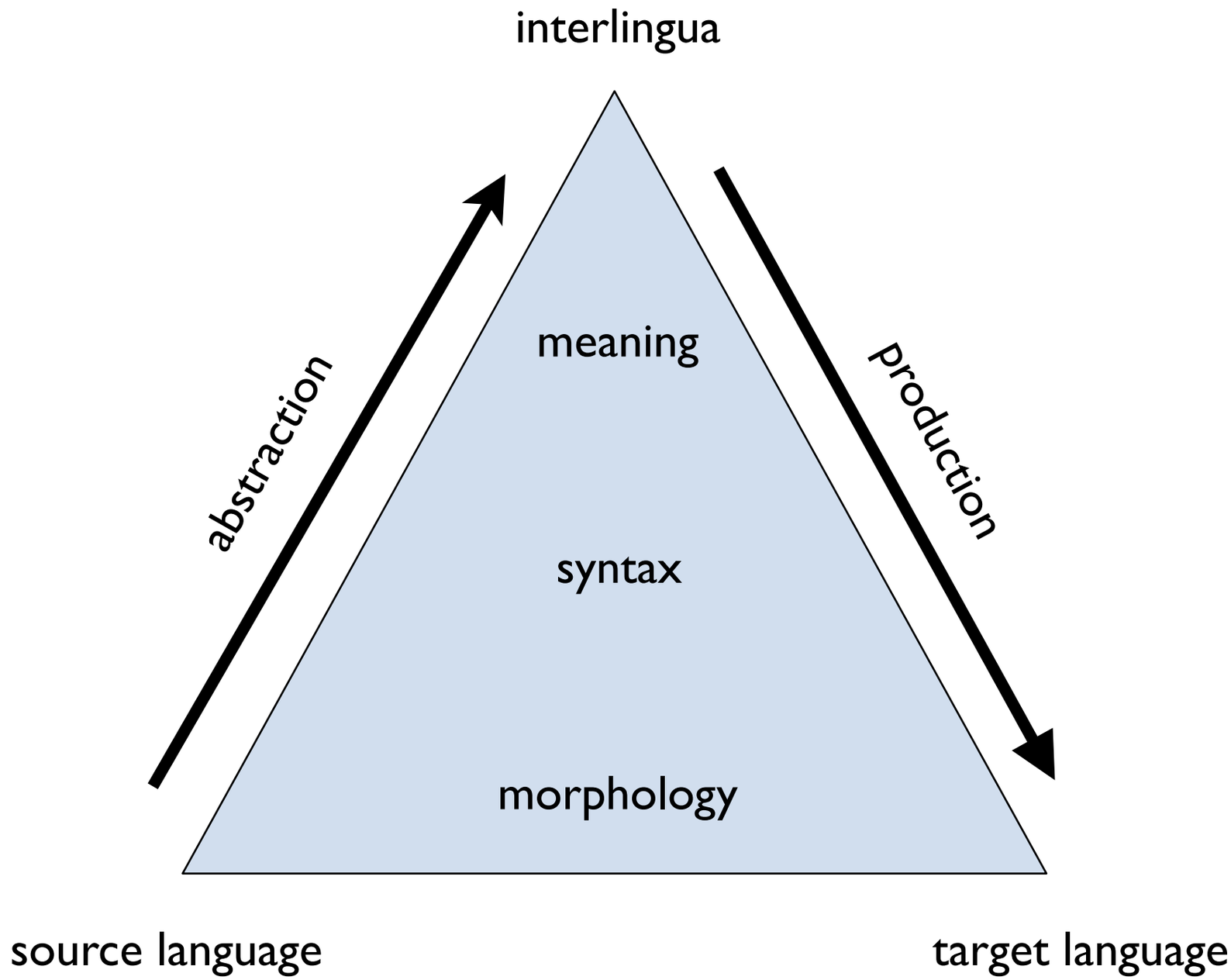
Early optimism

- Show video!

- Georgetown-IBM demo, 1954
- English-Russian MT
  - 250 words
  - 6 grammar rules
- Much media coverage!

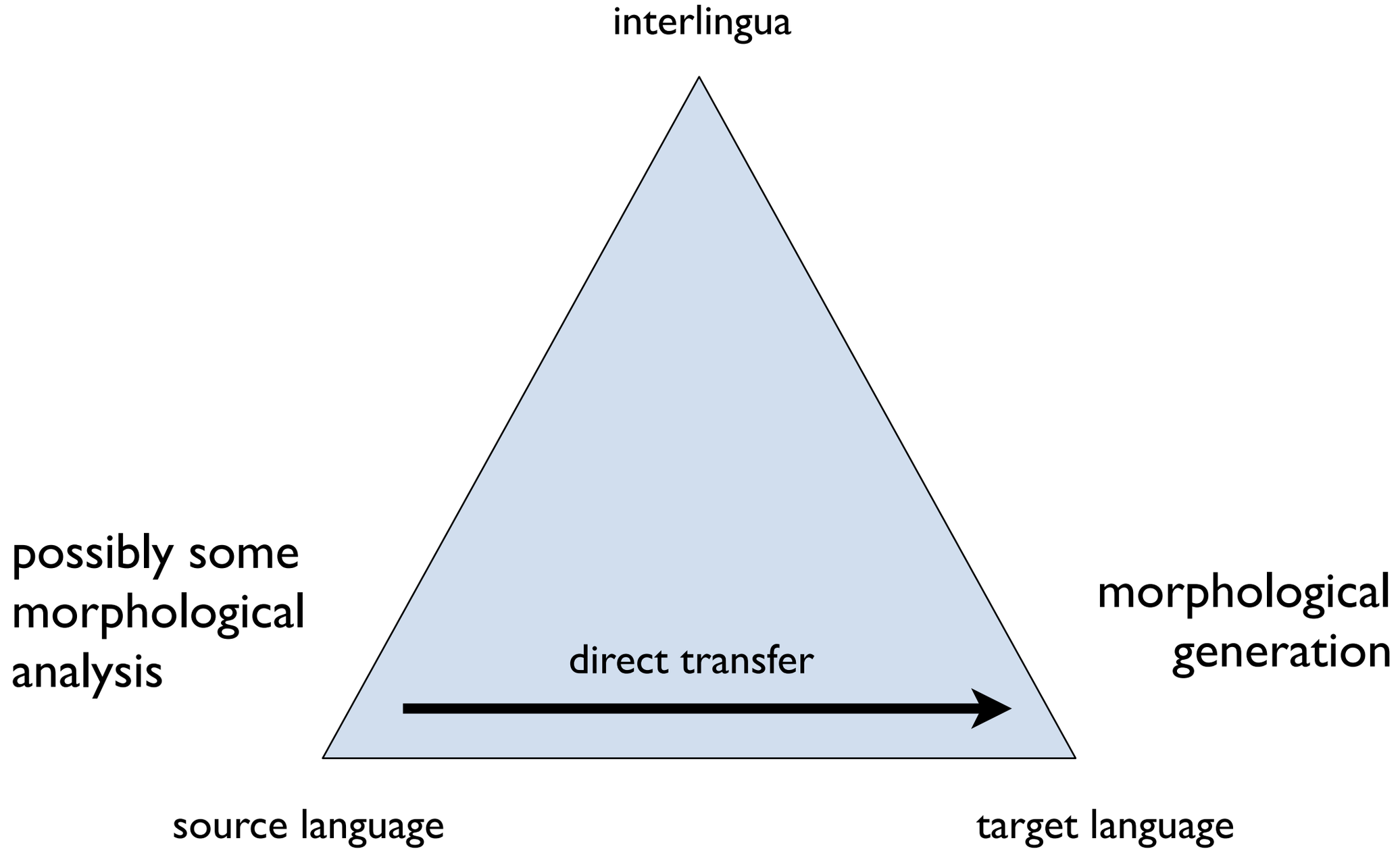
UPPSALA
UNIVERSITET

The Vauquois Triangle



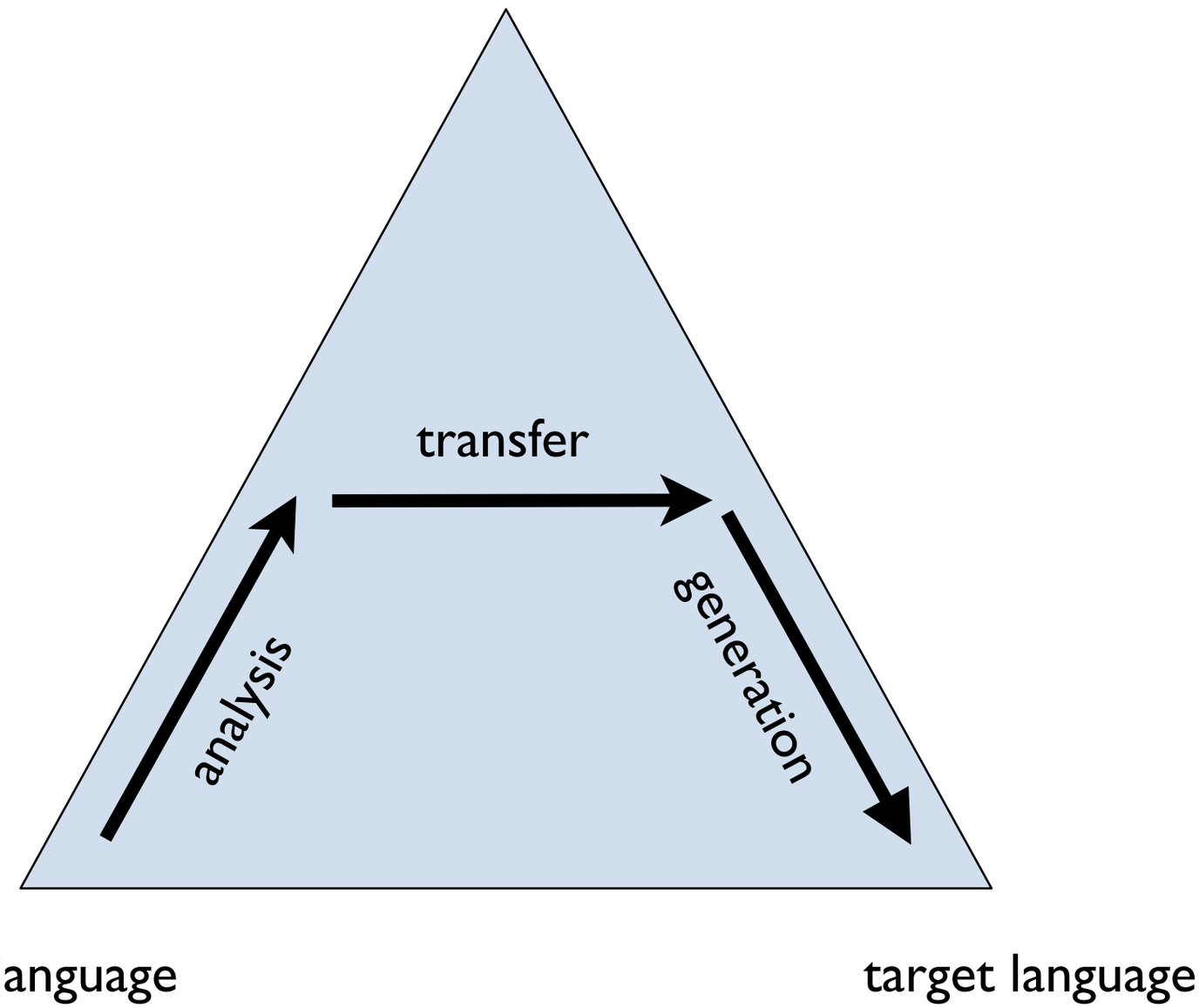
UPPSALA
UNIVERSITET

“Direct Machine Translation”



UPPSALA UNIVERSITET

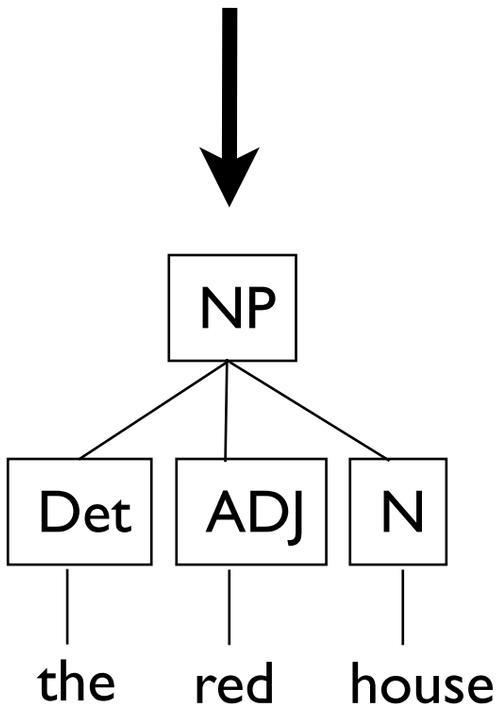
Transfer-Based Systems: 3 Steps



UPPSALA UNIVERSITET

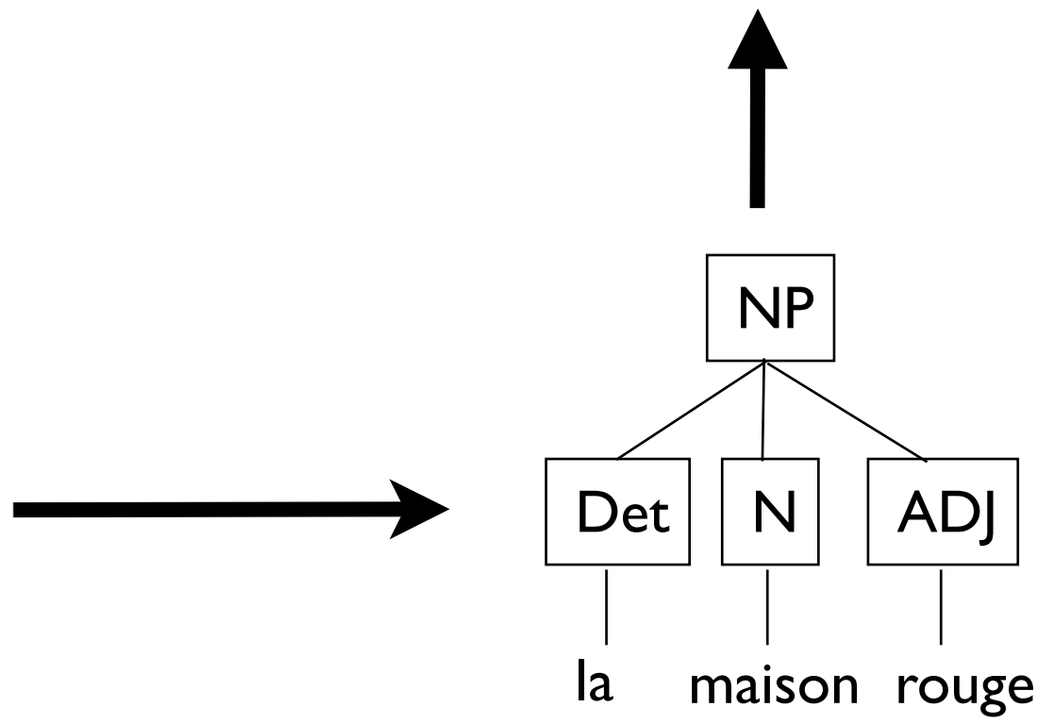
Transfer-Based Machine Translation

The red house .



N → house
ADJ → red
Det → the
NP → Det ADJ N

La maison rouge.



N → maison
ADJ → rouge
Det → la
NP → Det N ADJ

UPPSALA UNIVERSITET

Transfer-Based Machine Translation

What do we need?

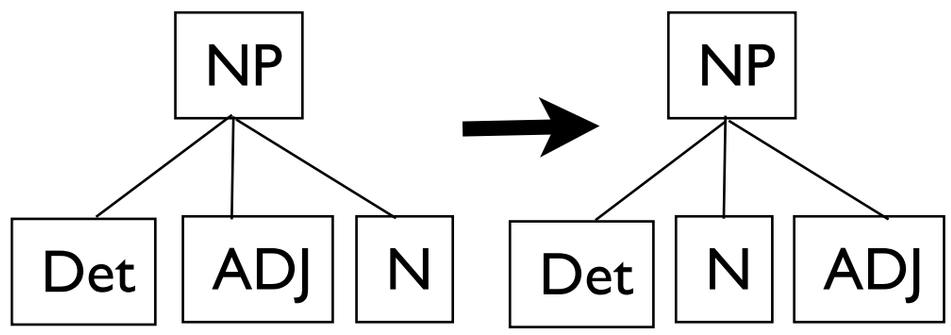
- source language parser
- transfer rules
- target language generator

Assumption

- Languages are very similar on a high level of abstraction

Transfer

- large bilingual dictionaries
- structural transfer rules



UPPSALA UNIVERSITET

Transfer-Based Machine Translation

Assumptions

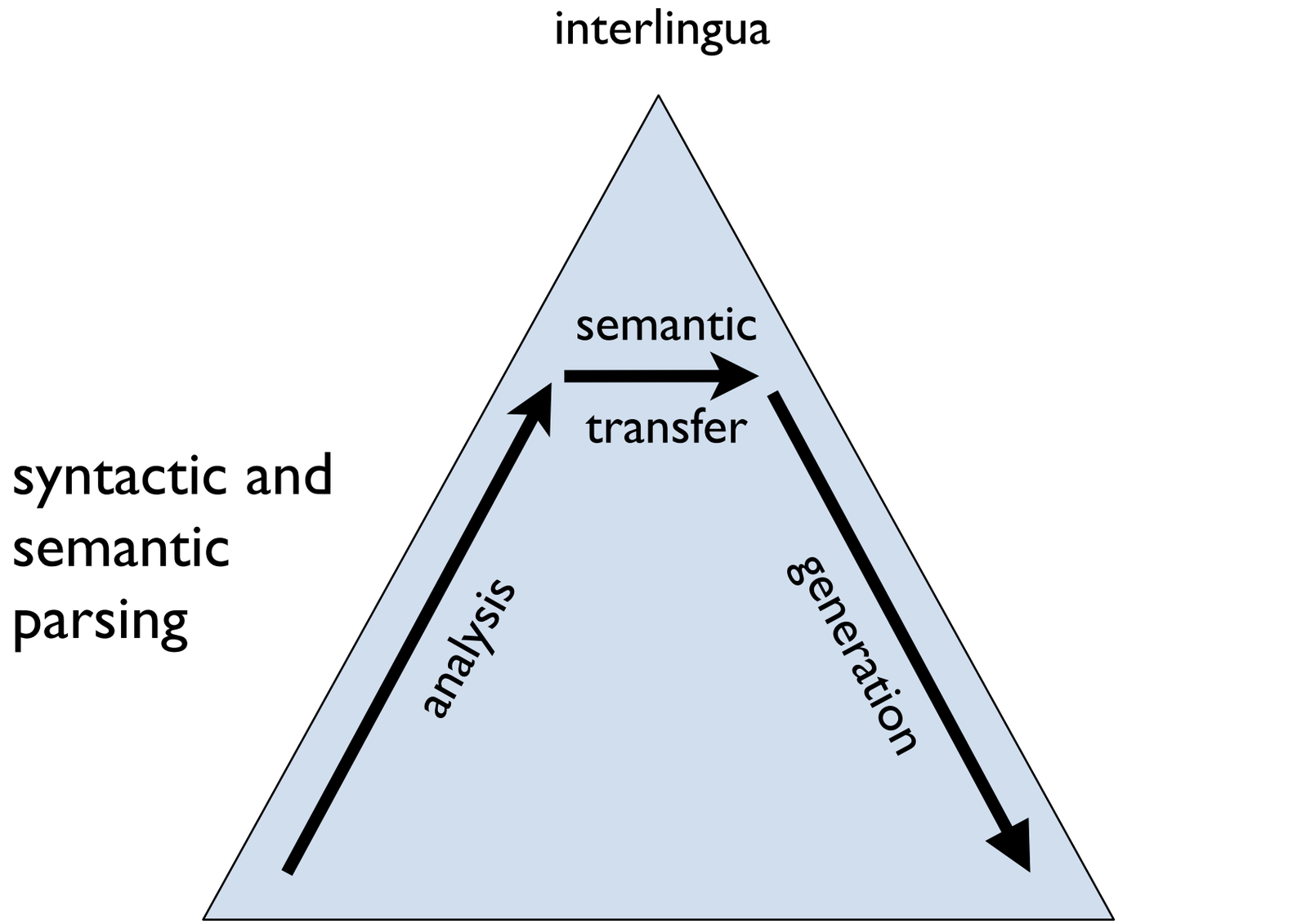
- strong language regularities
- only a few exceptions
- languages = compact grammars + lexicon
- robust and accurate parsing and generation

Grammar development (3 independent modules)

- rules defined by experts (traditional approach)
- induced from data

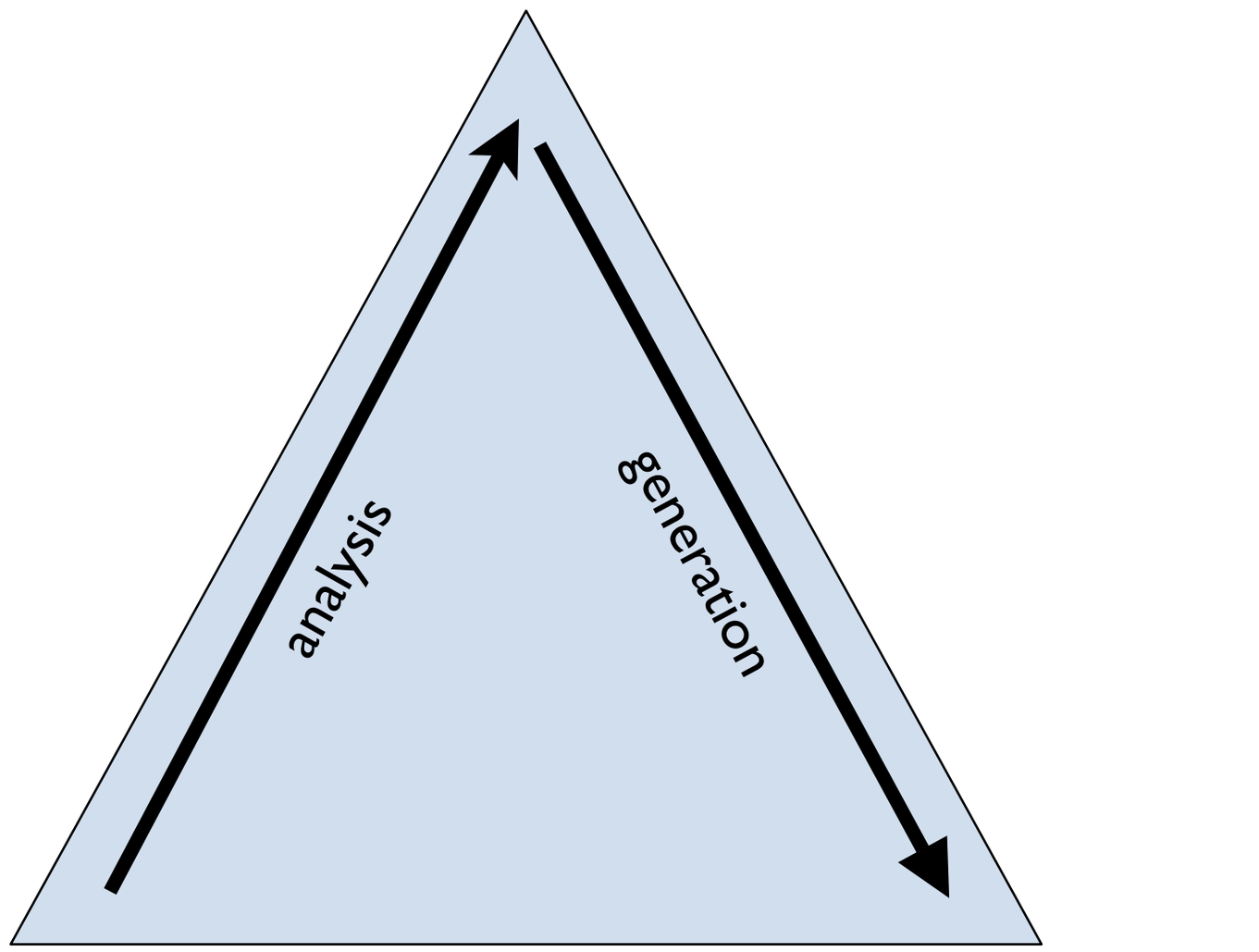
UPPSALA UNIVERSITET

Transfer-Based Machine Translation



UPPSALA UNIVERSITET

Interlingua-Based Models



UPPSALA UNIVERSITET

Interlingua-Based Models

Assumptions

- All languages can be generated from the same abstract meaning representation
- All aspects of language can be captured by an interlingua

Advantage:

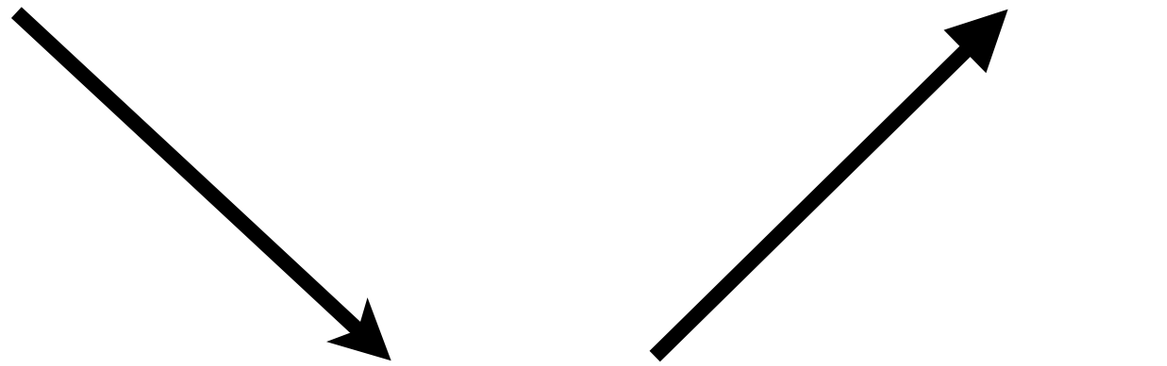
- no transfer
- new language = new analysis and generation modules, no transfer modules

UPPSALA UNIVERSITET

Interlingua-Based Models

Mary did not slap the green witch.

Maria no dió una bofetada a la bruja verde



EVENT	SLAPPING		
AGENT	MARY		
TENSE	PAST		
POLARITY	NEGATIVE		
THEME	WITCH		
	DEFINITENESS	DEF	
	ATTRIBUTES	HAS-COLOR	GREEN

UPPSALA UNIVERSITET

Expert-Driven Systems

Linguistic grammar formalisms

Handcrafted rules

- VP → PP[+Goal] V ⇒ VP → V PP[+Goal]
- ...

Preference mechanisms

- more specific rules first (covering exceptions over more general abstract rules)
- prefer phrases over single word entries in dictionaries
- prefer domain-specific terminology over general vocab.

UPPSALA
UNIVERSITET

Expert-Driven Systems

ALPAC report (1966)

- MT quality is too low
- No advantage of MT over human translation
- Almost all funding was stopped in the U.S.

Problems:

- Robustness and translation speed
- Expensive development
- Not very flexible (new domains, languages ...)
- Static, categorical models, but languages are dynamic and ambiguous
- Slow development

UPPSALA
UNIVERSITET

Learning to Translate

Induce translation knowledge from data

- existing translations
- existing language resources

Learning Transfer Systems

- data-driven parsing
- induced transfer rules
- language modeling for generation

Global Data-Driven MT

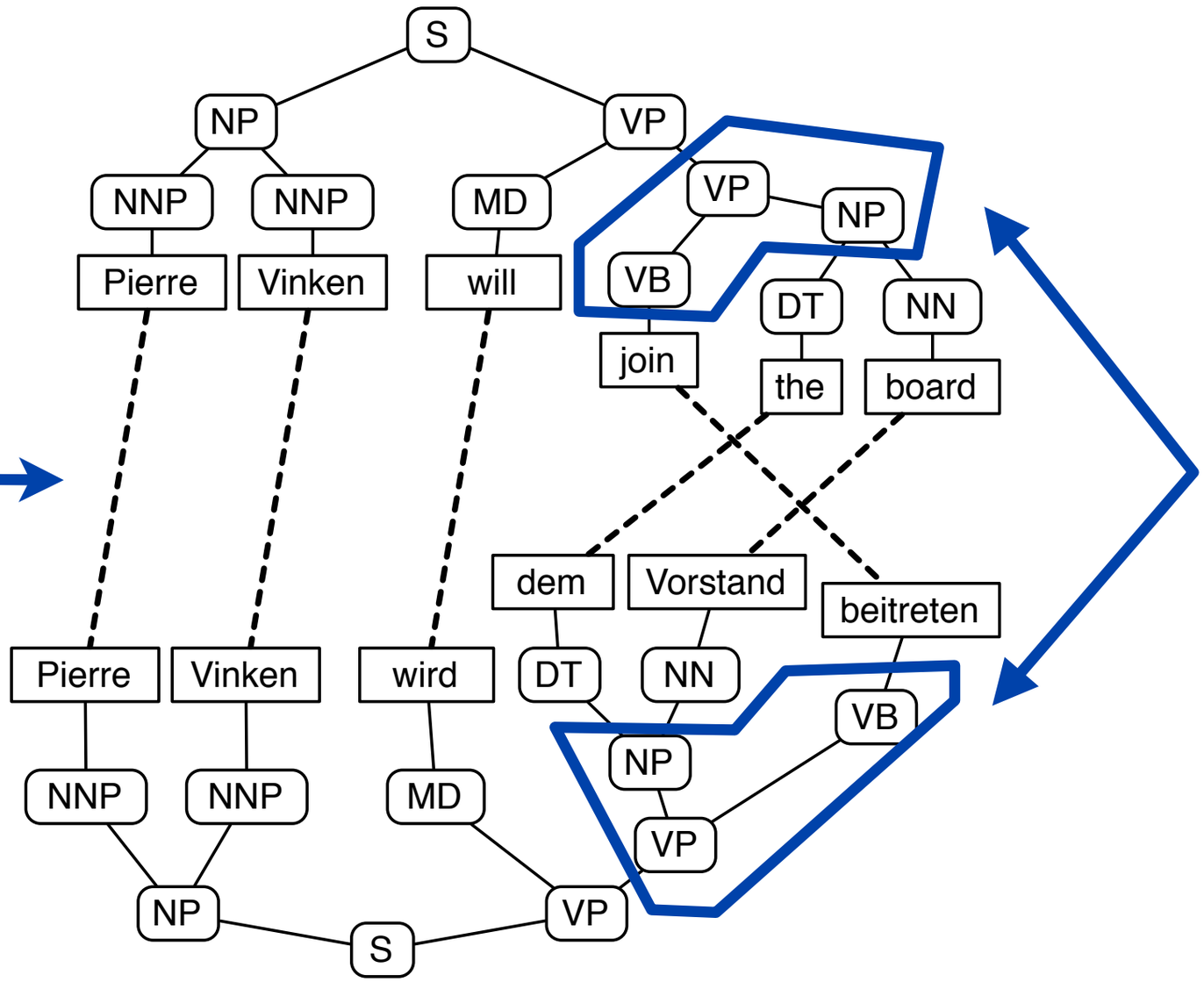
- surface-to-surface MT
- abstraction within a global translation model

UPPSALA UNIVERSITET

Learning Transfer-Based MT

Possibility: Induce transfer rules from aligned data

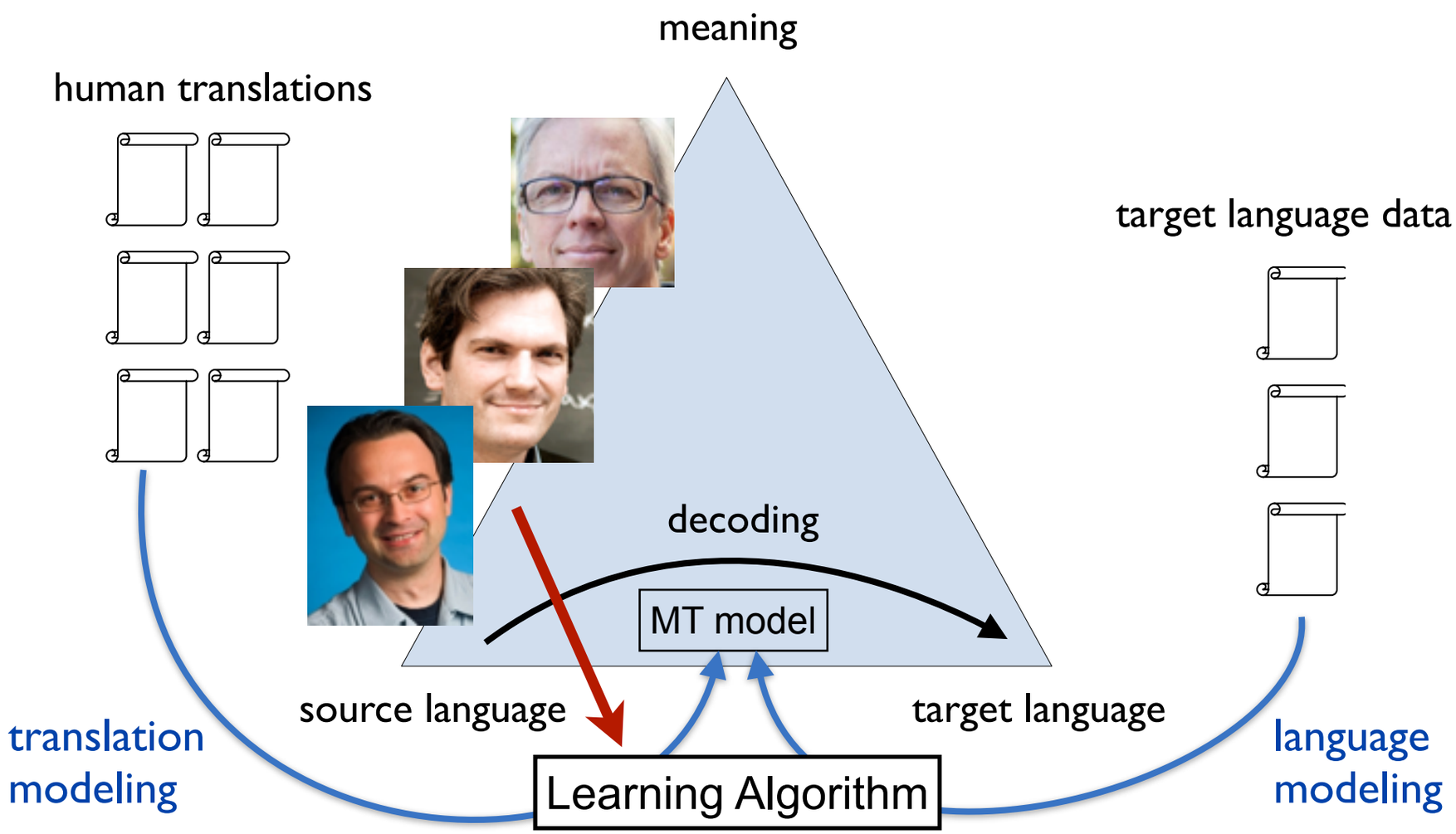
Extract bilingual dictionaries



Extract transfer rules

UPPSALA UNIVERSITET

Data-Driven Machine Translation



UPPSALA
UNIVERSITET

Statistical Machine Translation

1947: MT as decoding (Warren Weaver)

1988: Word-based models

1999: Public implementation of alignment models (GIZA)

2003: Phrase-based SMT

2004: Public phrase-based decoder (Pharaoh)

2005: Hierarchical models

2007: Moses (end-to-end SMT toolbox)

2014: Neural machine translation

along with many tools, much more data and better computers

UPPSALA UNIVERSITET

Advantages of Data-Driven MT

Human Translations Naturally Appear

- no need for artificial annotation
- can be provided by non-experts

Implicit Linguistics

- translation knowledge is in the data
- distributional relations within and across languages

Constant Learning is Possible

- feed with new data as they appear
- quickly adapt to new domains and language pairs

UPPSALA UNIVERSITET

Take-Home Messages

Machine Translation is important

Machine Translation is difficult

Main Rule-Based Approaches

- Transfer-Based MT
- Interlingua-Based MT
- Direct Translation Systems

Expert-Driven Systems with Hand-Crafted Rules

Data-Driven Systems based on Example Data

UPPSALA UNIVERSITET

Outlook

This afternoon: MT evaluation

Next week: MT in practice

- Guest lecture, Convertus (Commercial MT solutions in Uppsala)
- Lab 1: evaluation

Then:

- Introduction to SMT
- Lab 2: word-based models
- Guest lecture, Semantix