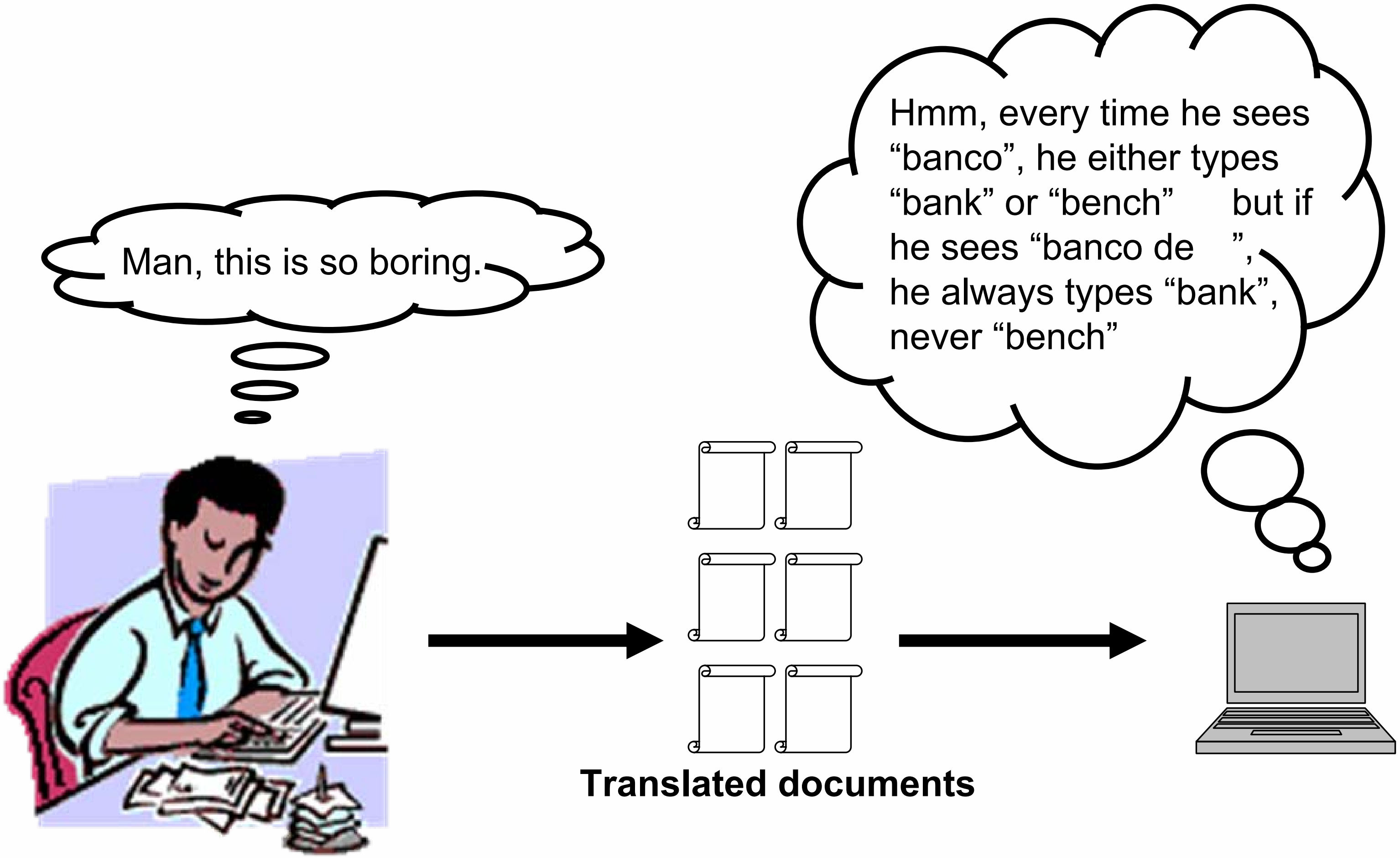
Automated Translation How does it work?

Finnish ELRC Workshop

Jörg Tiedemann
University of Helsinki

Why Machine Translation?



(illustration by Kevin Knight)

Task-Dependent Machine Translation

Balance MT quality and input restrictions, depending on task

Task-Dependent Machine Translation

Balance MT quality and input restrictions, depending on task

general purpose **browsing quality**
fully automatic **gisting**
on-line service

Task-Dependent Machine Translation

Balance MT quality and input restrictions, depending on task

general purpose browsing quality	post-editing editing quality
fully automatic gisting	computer-aided translation (CAT)
on-line service	localisation & more

Task-Dependent Machine Translation

Balance MT quality and input restrictions, depending on task

general purpose browsing quality	post-editing editing quality	sublanguage publishing quality
fully automatic gisting	computer-aided translation (CAT)	fully automatic FAHQMT
on-line service	localisation & more	restricted domain

Task-Dependent Machine Translation

Balance MT quality and input restrictions, depending on task

general purpose browsing quality	post-editing editing quality	sublanguage publishing quality
fully automatic gisting	computer-aided translation (CAT)	fully automatic FAHQMT
on-line service	localisation & more	restricted domain

But - how does it work?

Becoming a Translator

Learn to understand languages

- recognise patterns
- find relations between linguistic units and the real world
- generalise and make abstractions

Learn to speak a language

- remember and repeat
- produce new utterances
- become fluent

Learn to translate

- identify mappings between languages

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Becoming an MT System

Natural language understanding

- recognise patterns
- find relations between linguistic units (and the real world)
- generalise and make abstractions

Language modeling and generation

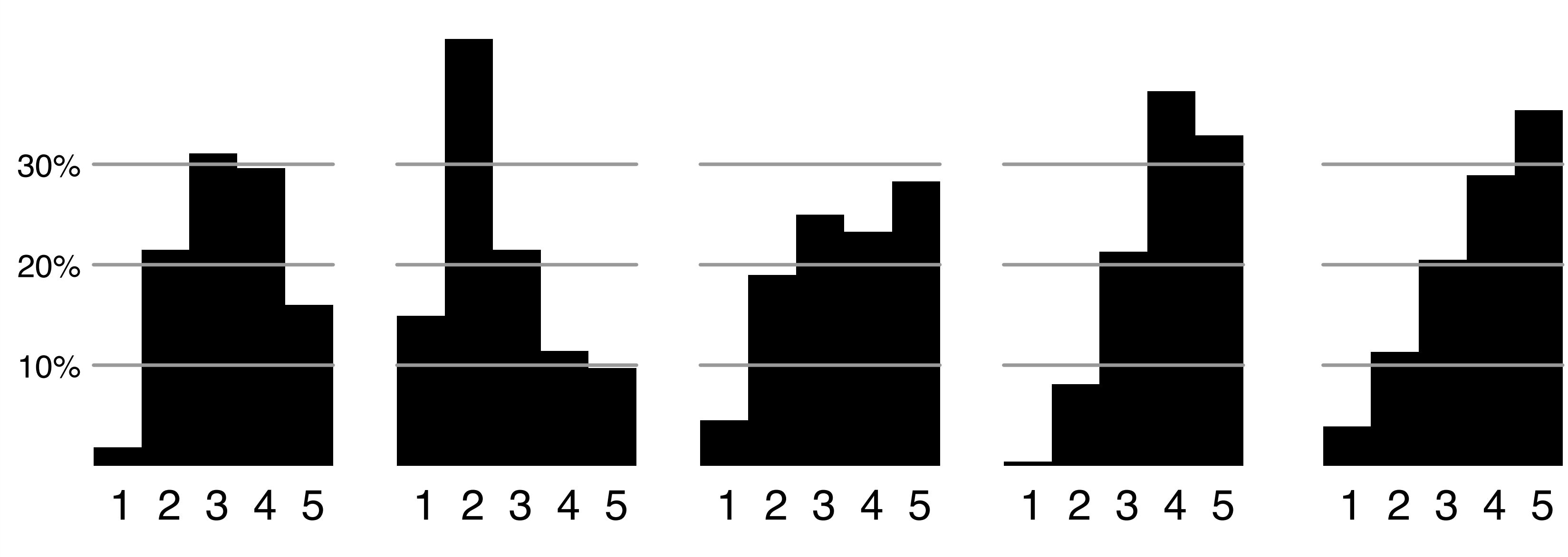
- remember and repeat
- produce new utterances
- become fluent

Alignment and transfer

- identify mappings between languages

Why is Translation so Difficult?

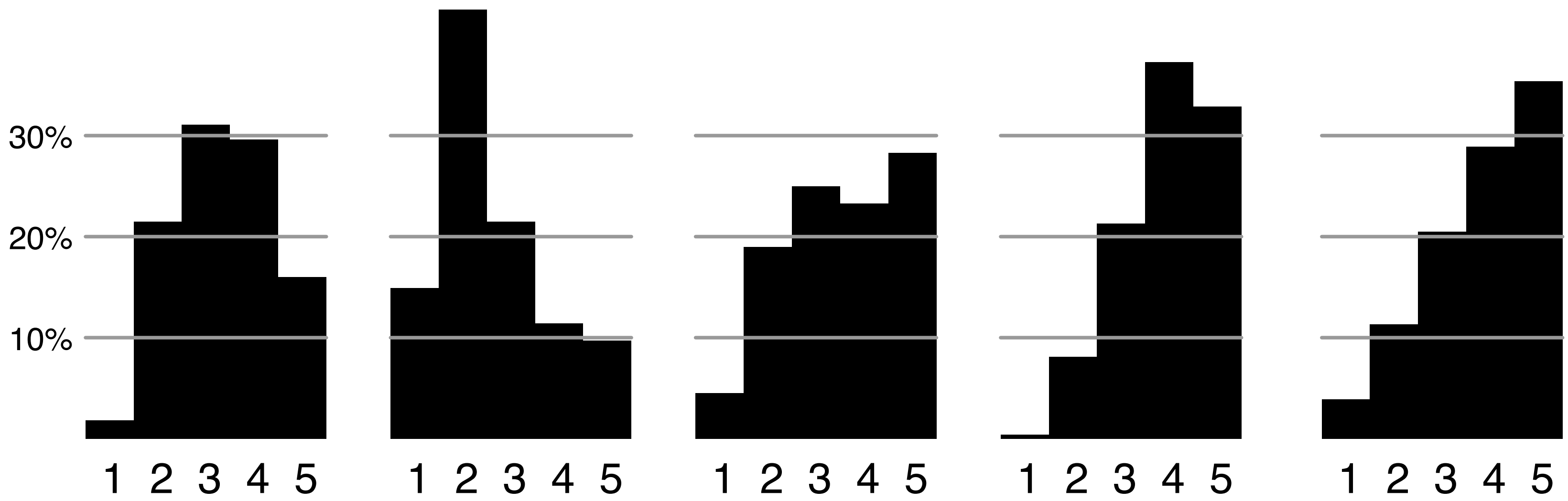
Histogram of adequacy judgments by different evaluators:



(from WMT 2006 evaluation)

Why is Translation so Difficult?

Histogram of adequacy judgments by different evaluators:



(from WMT 2006 evaluation)

ambiguity

fuzzy concepts

pragmatics

subjectivity

language divergences

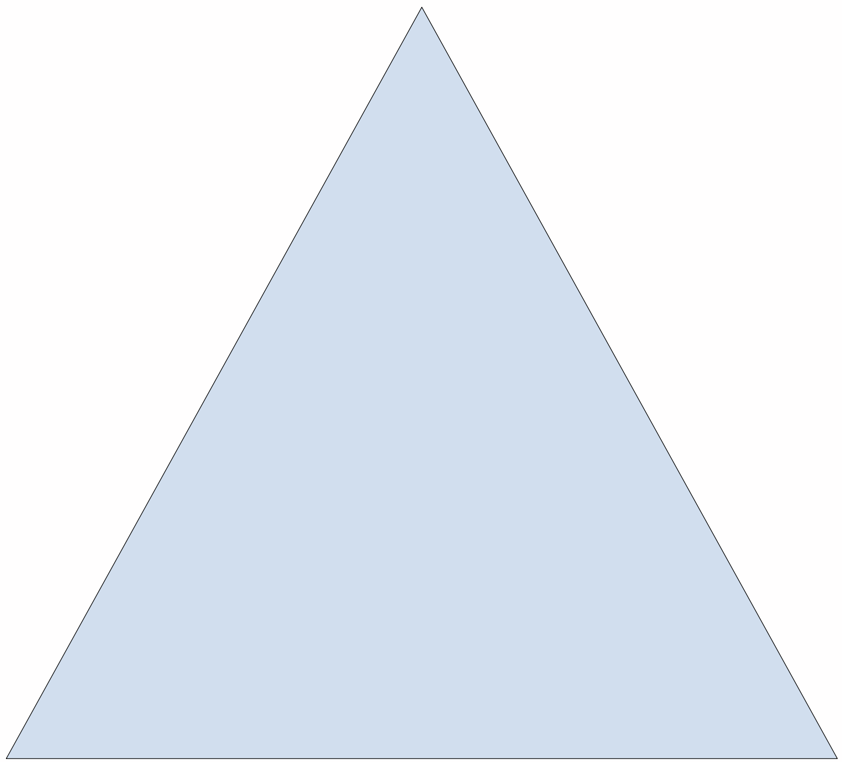
redundancy

world knowledge

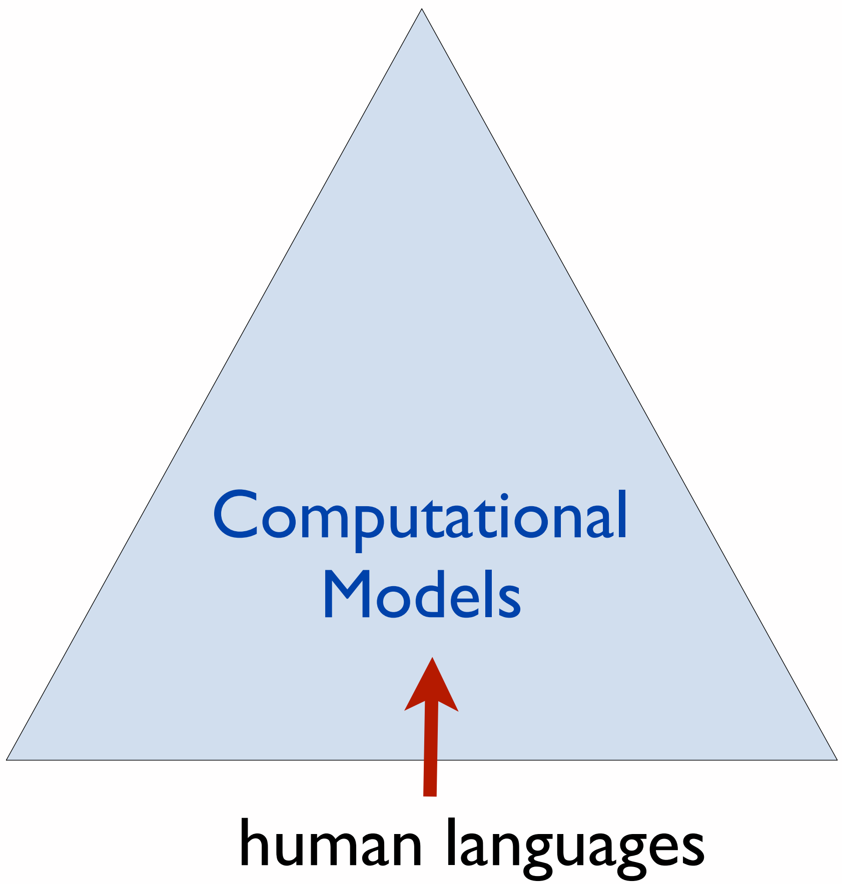
creativity

natural and stylistic variation

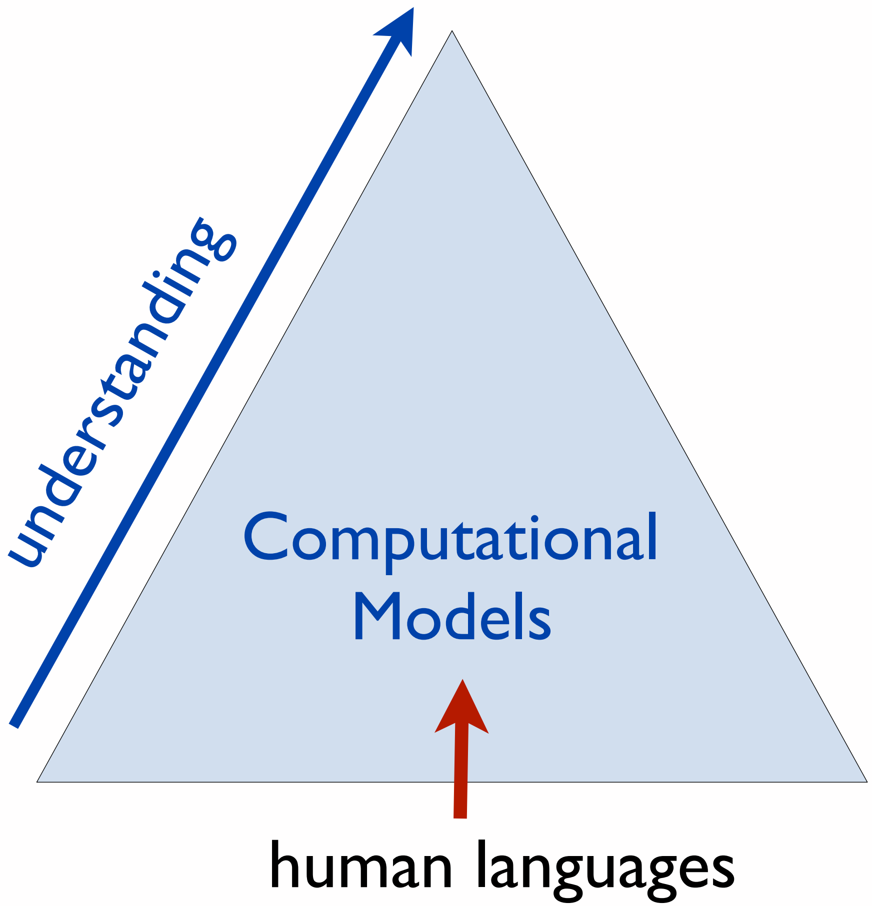
Language Technology and MT



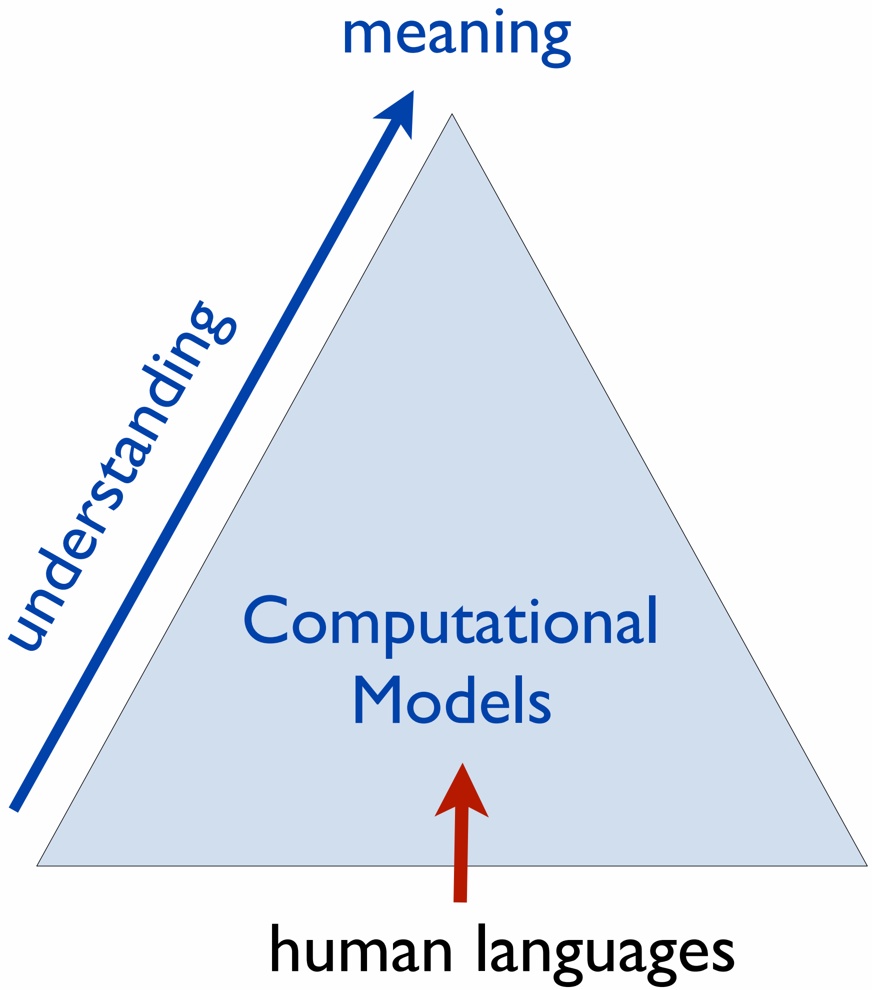
Language Technology and MT



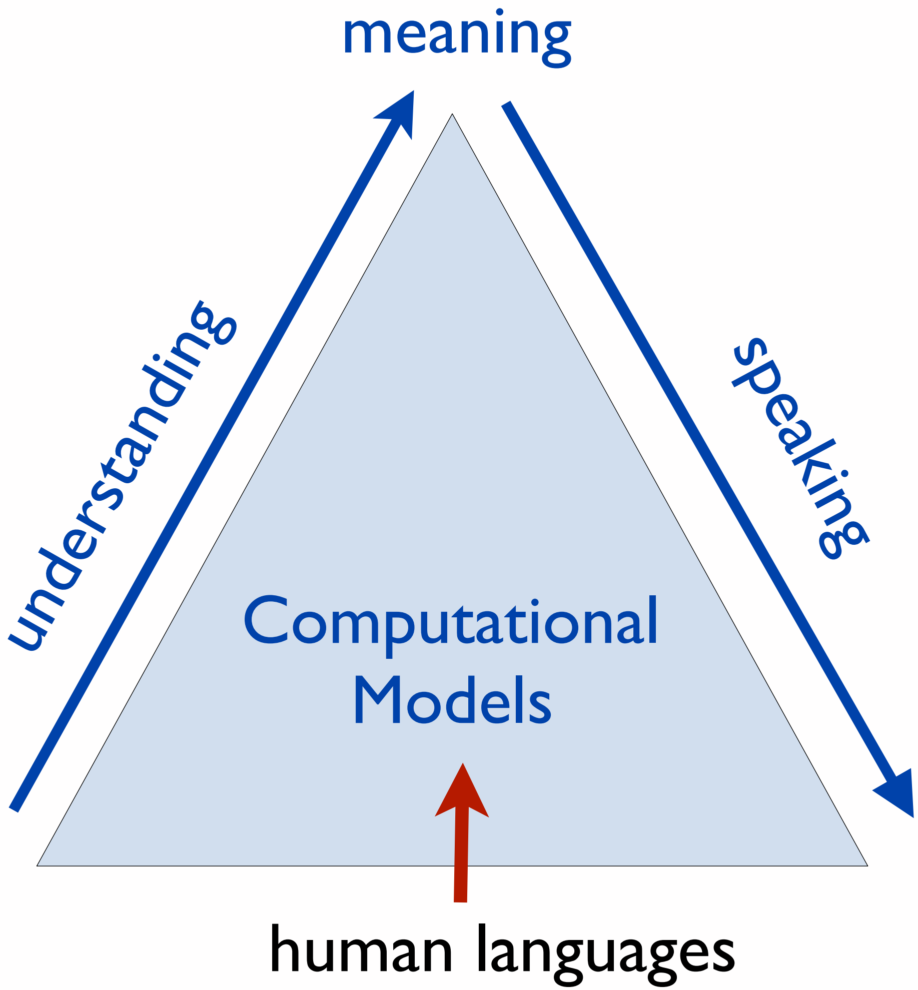
Language Technology and MT



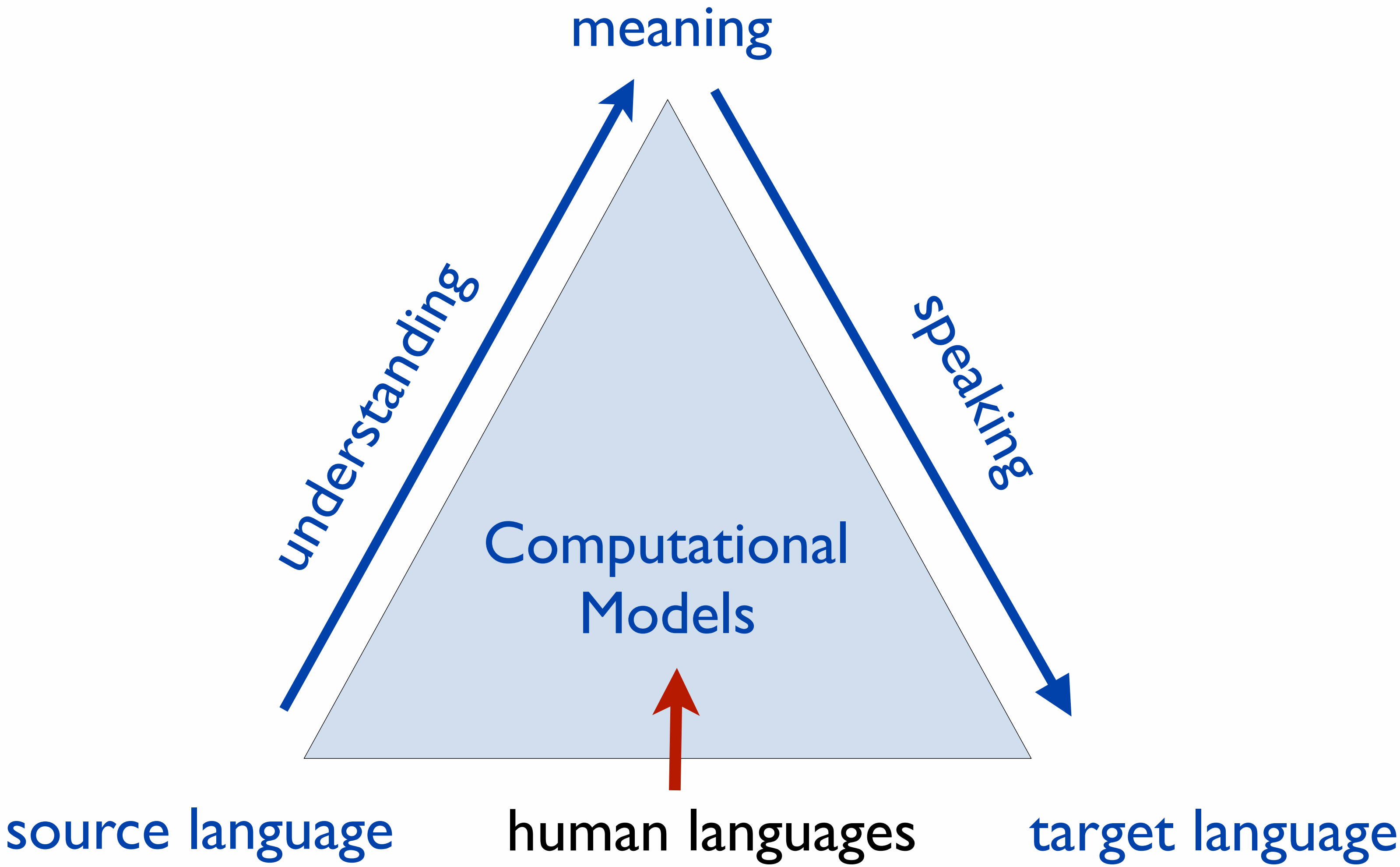
Language Technology and MT



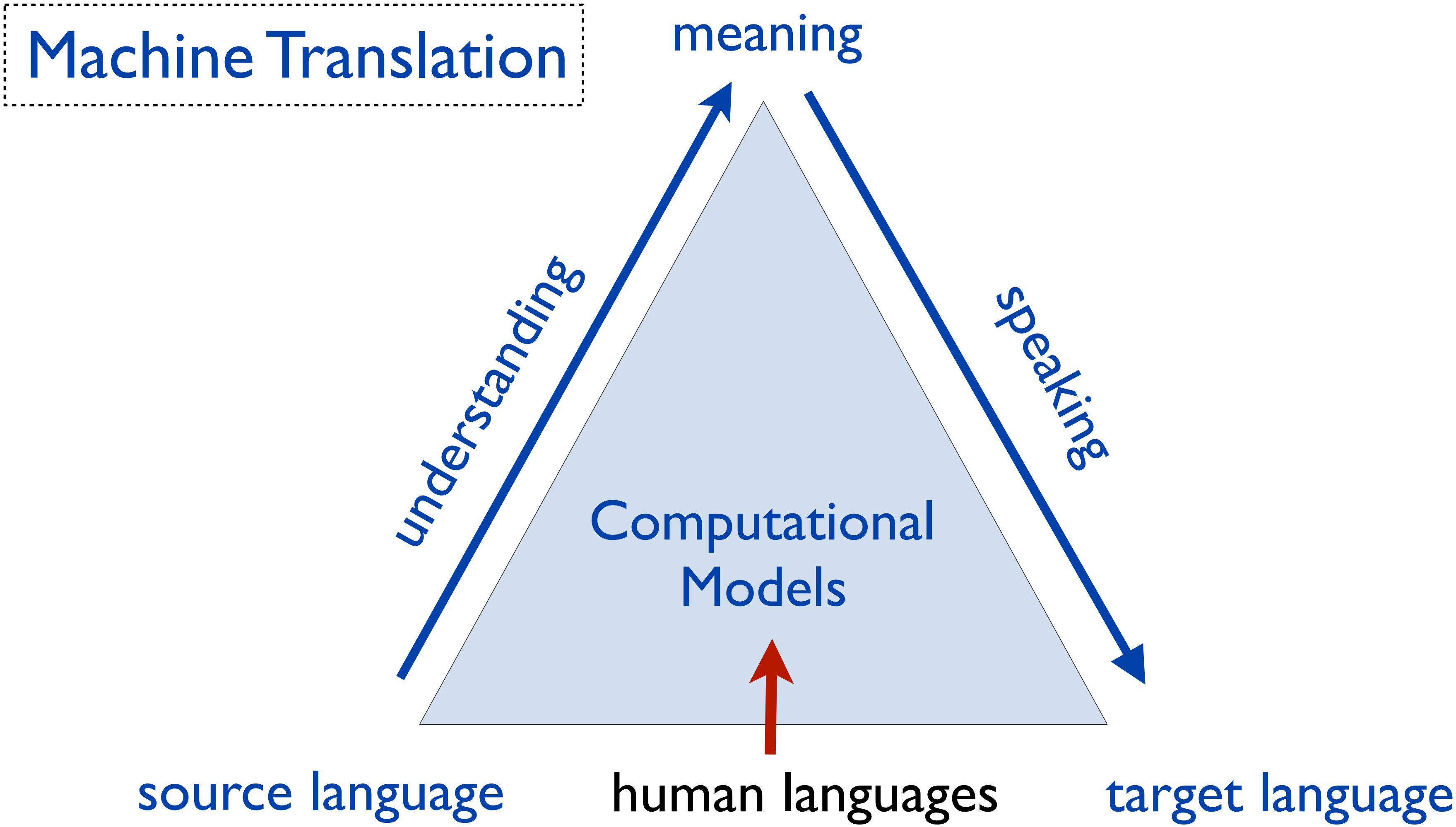
Language Technology and MT



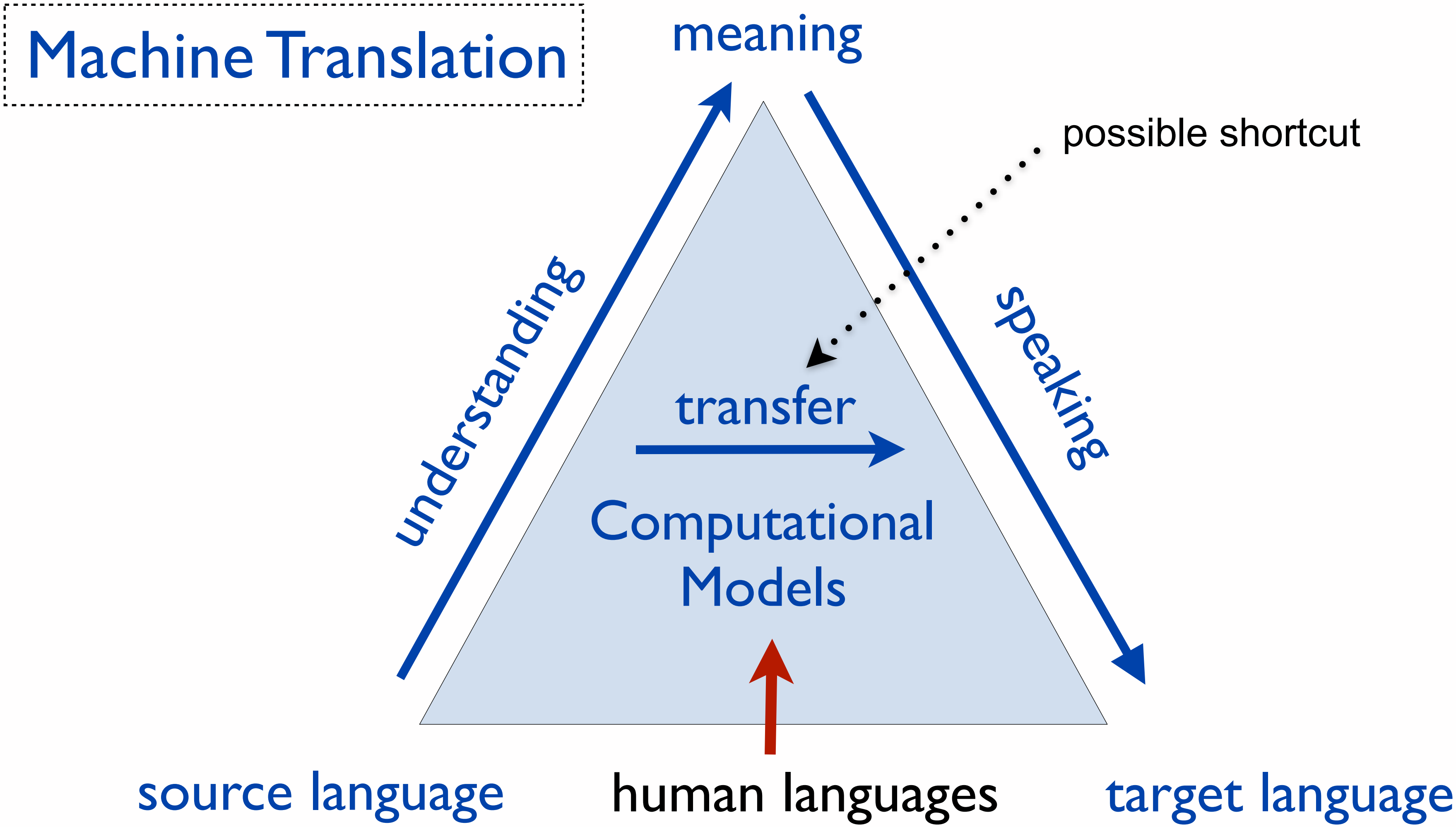
Language Technology and MT



Language Technology and MT



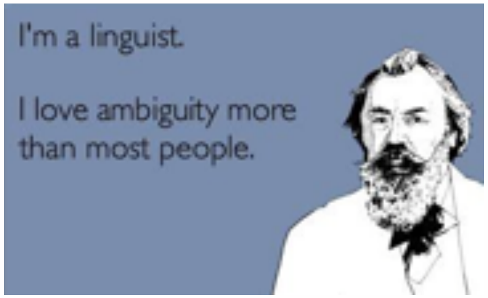
Language Technology and MT



Language Technology and MT

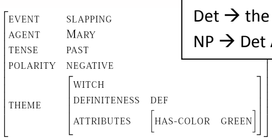
Machine Translation

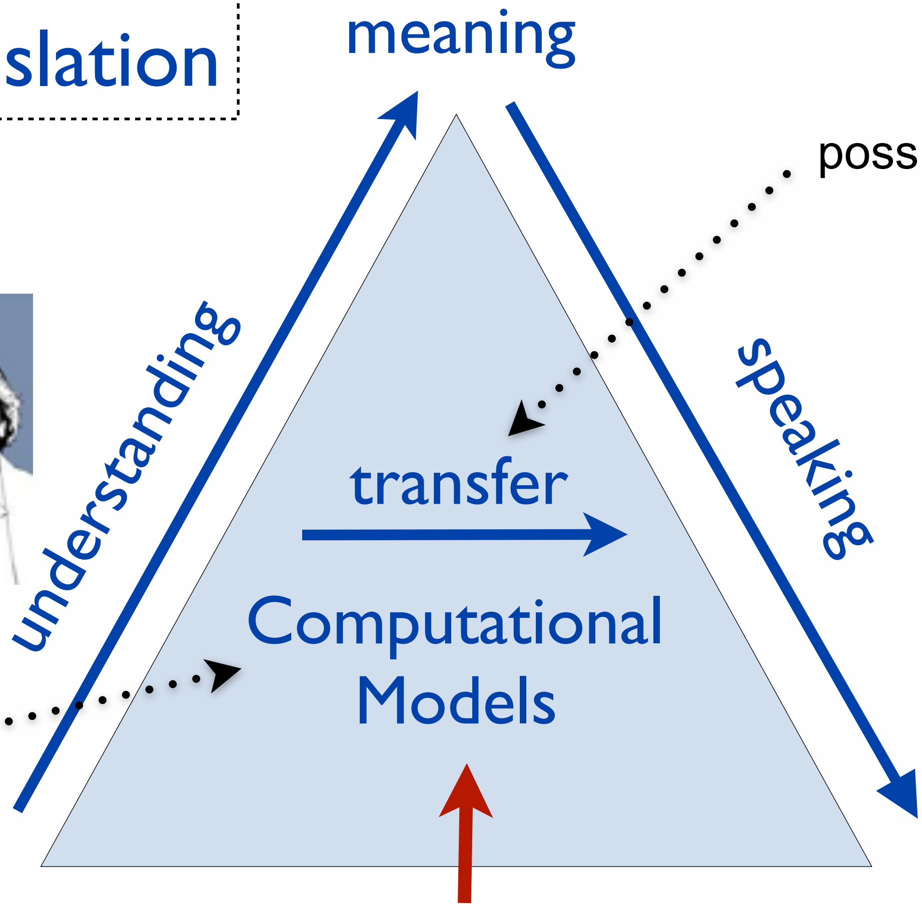
expert-driven



VP → PP[+Goal] V ⇒ VP → V

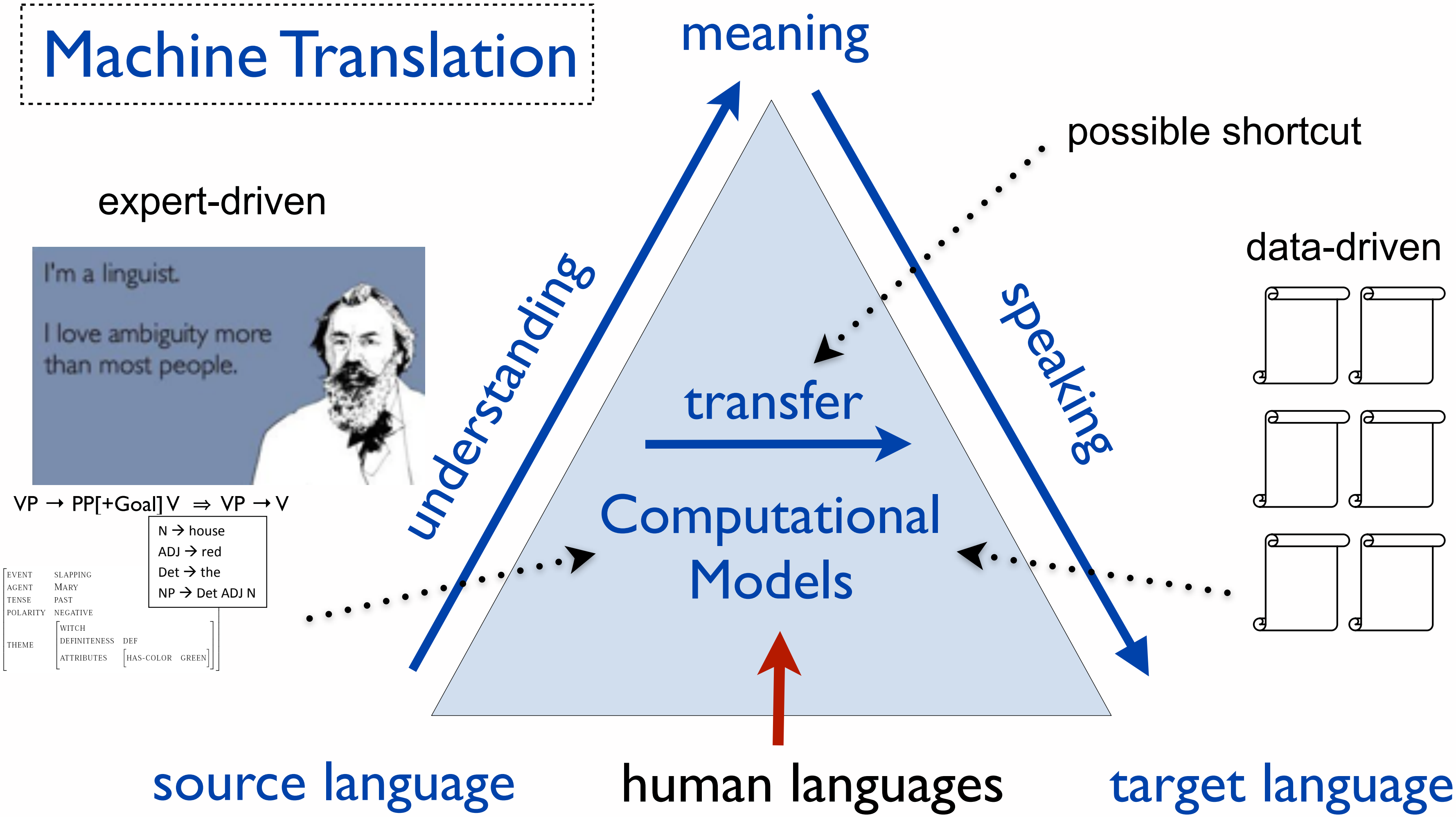
N → house
ADJ → red
Det → the
NP → Det ADJ N





HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Language Technology and MT



The Advantage of Data-Driven MT

The Advantage of Data-Driven MT

Human translations naturally appear

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

The Advantage 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

The Advantage of Data-Driven MT

Human translations naturally appear

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

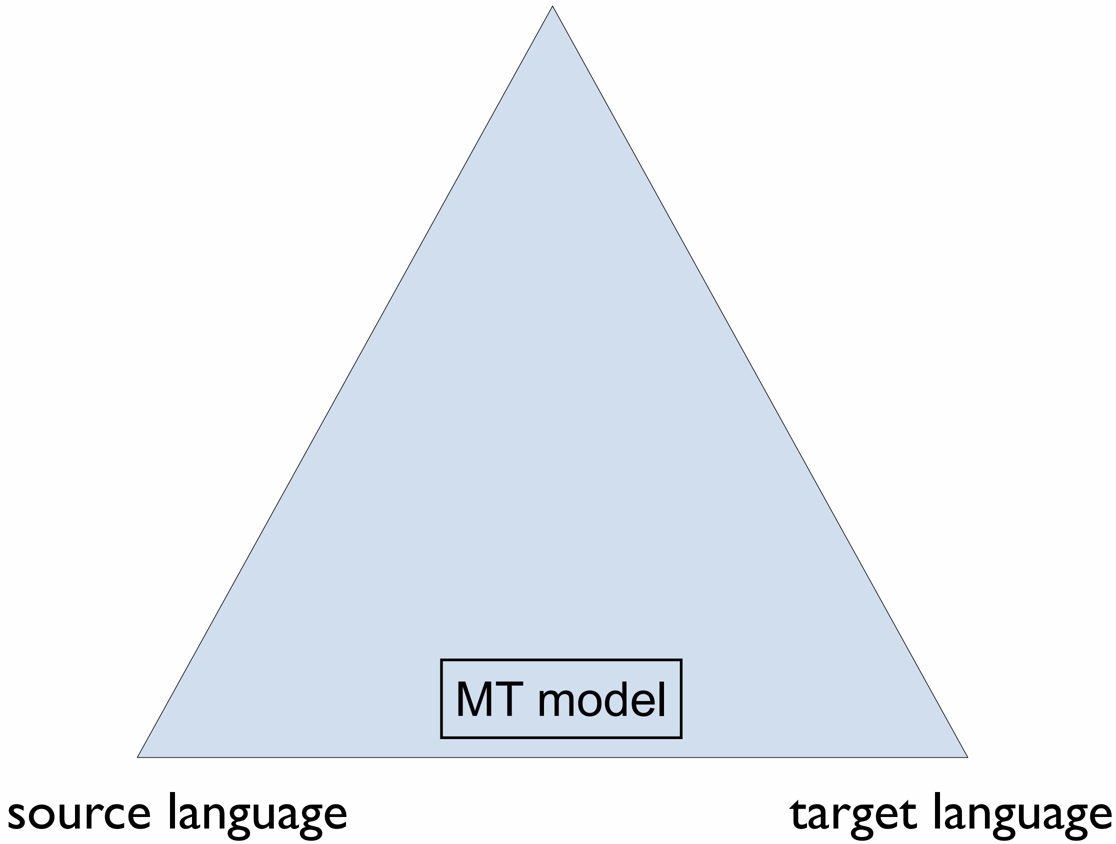
Implicit linguistics

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

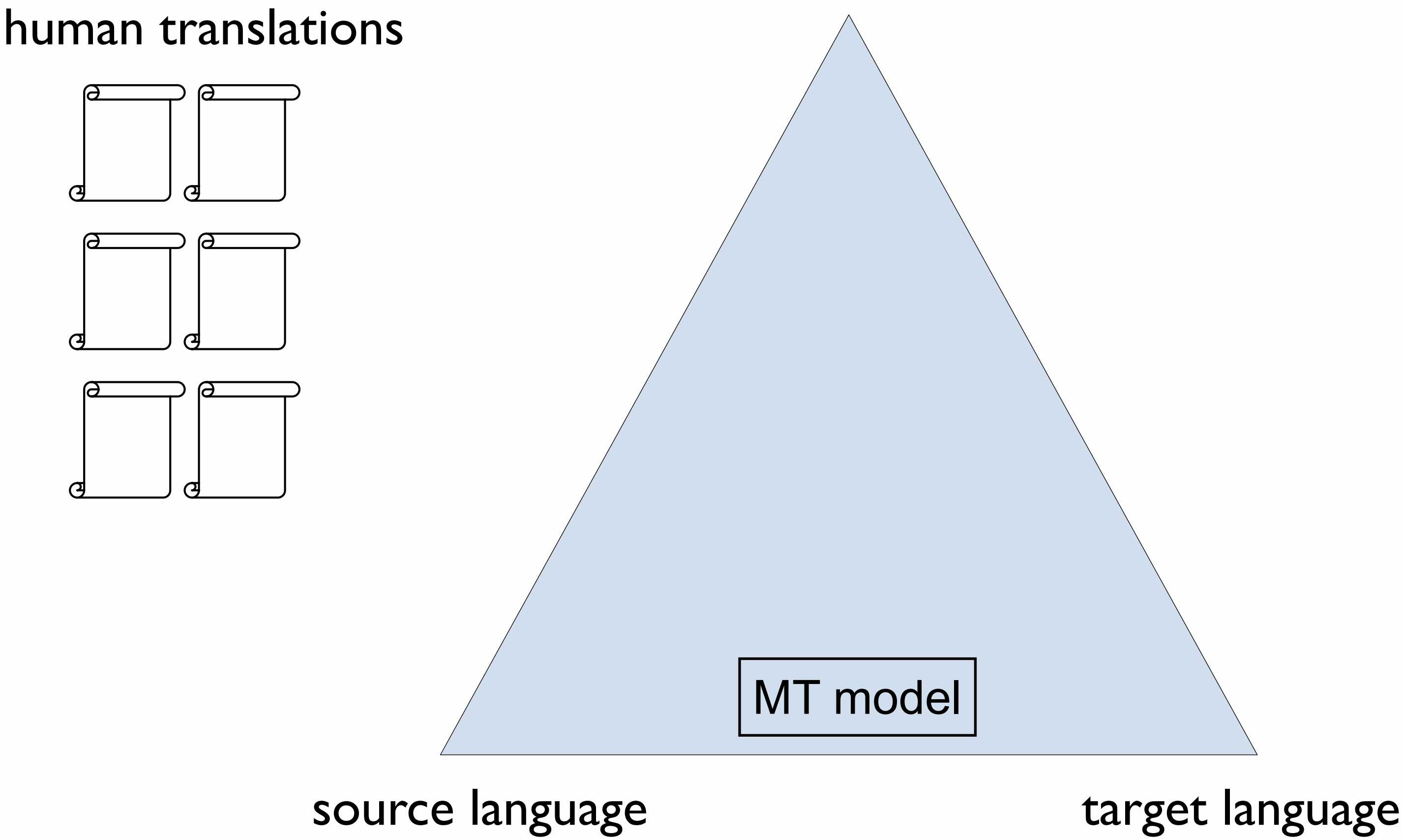
Constant learning is possible

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

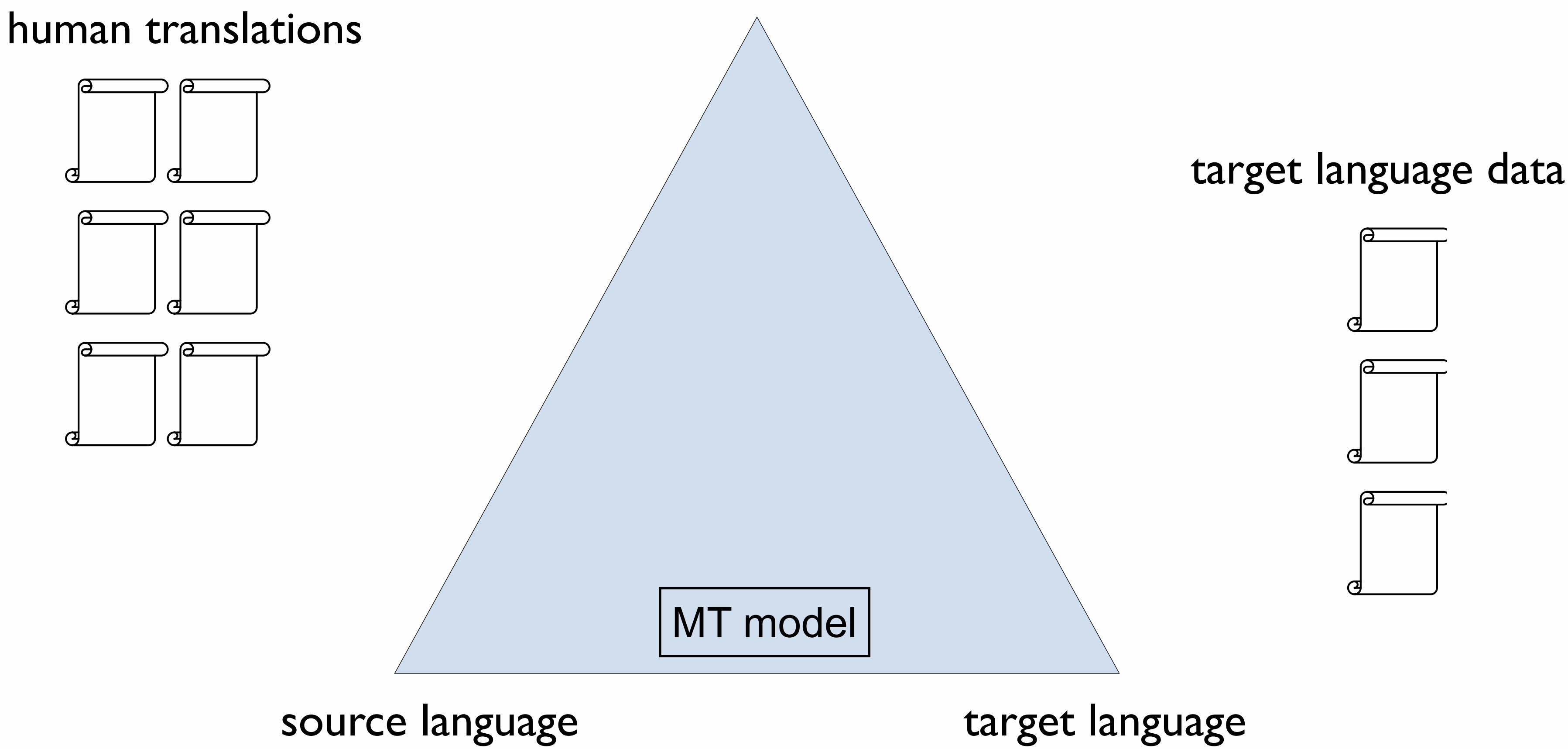
Data-Driven Machine Translation



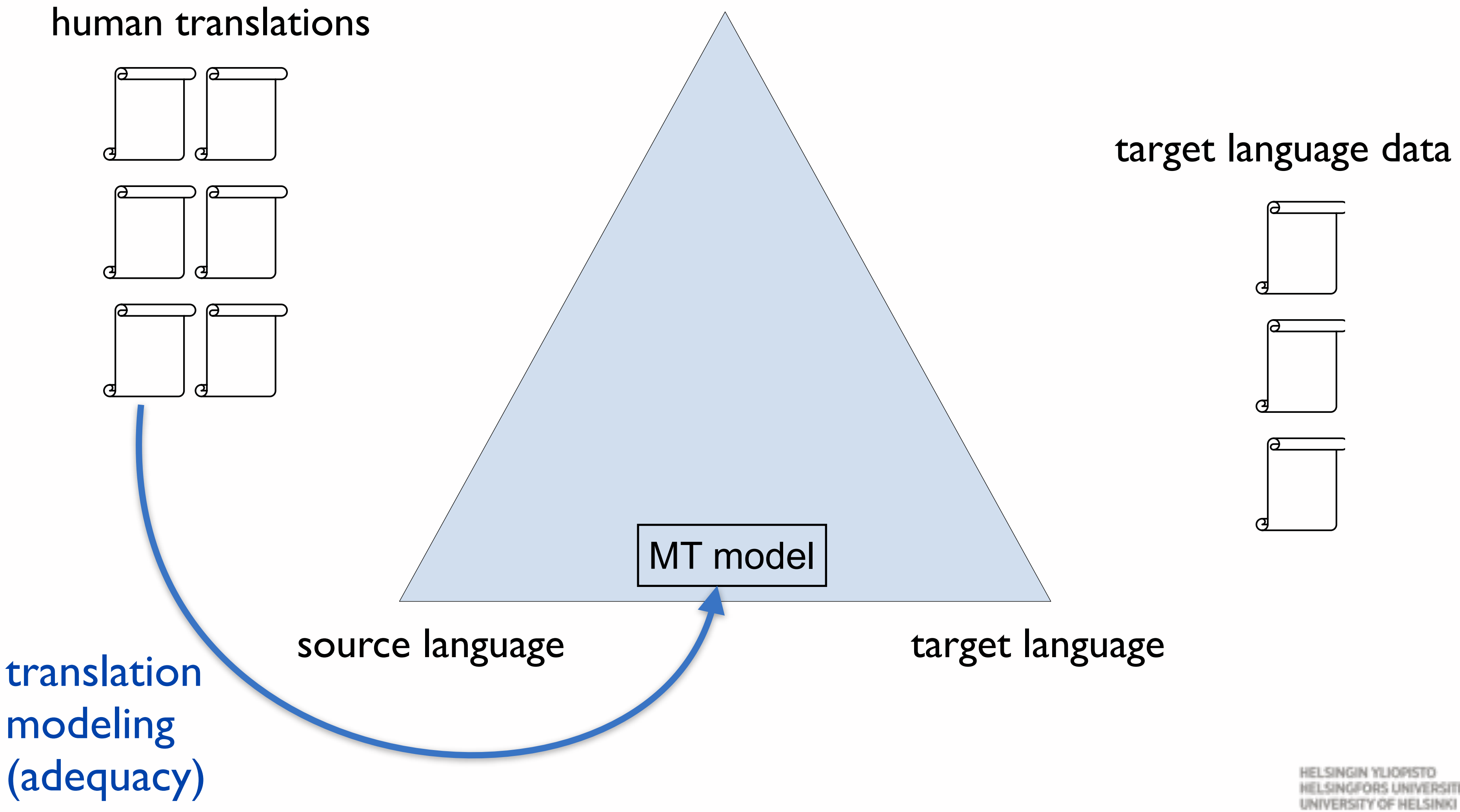
Data-Driven Machine Translation



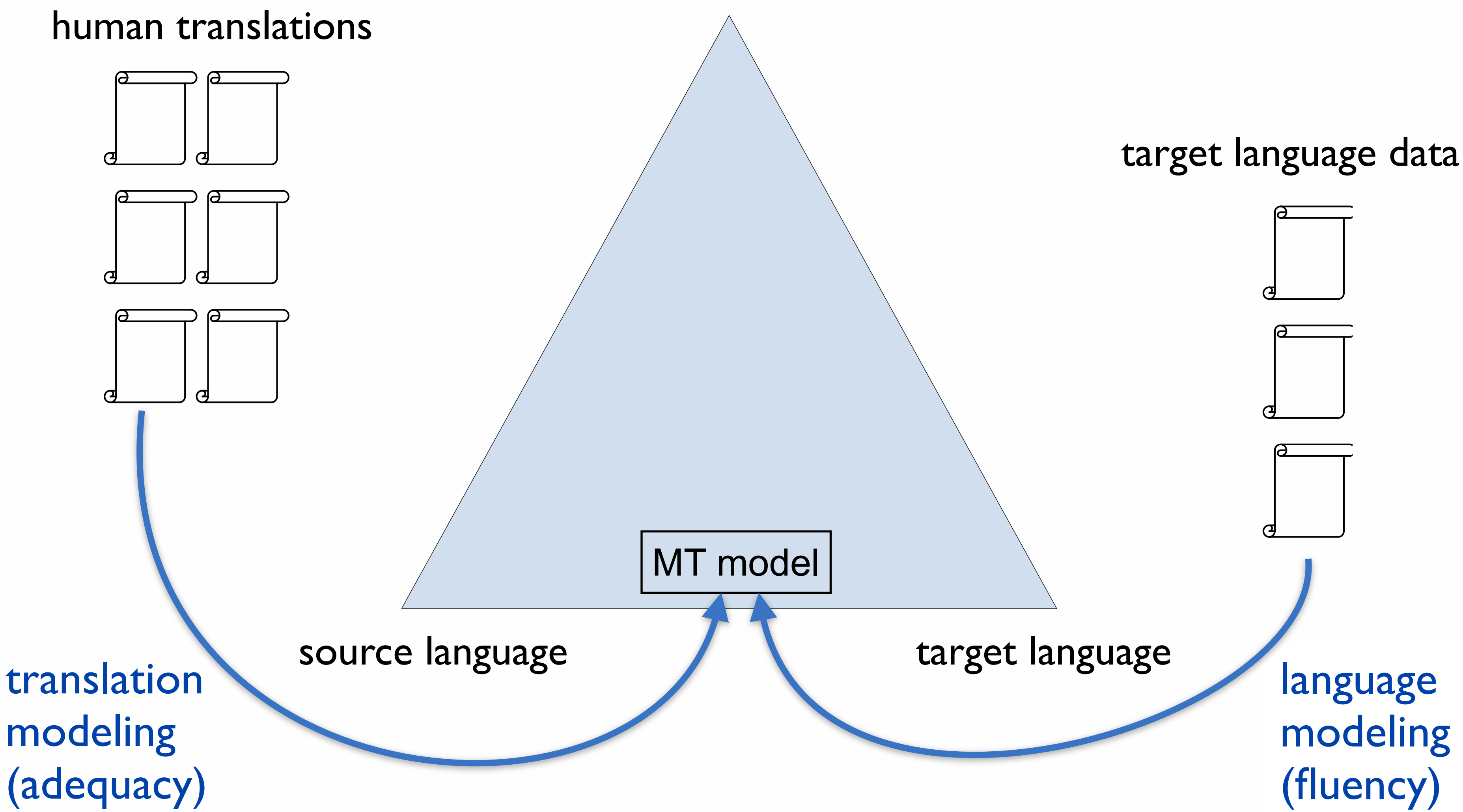
Data-Driven Machine Translation



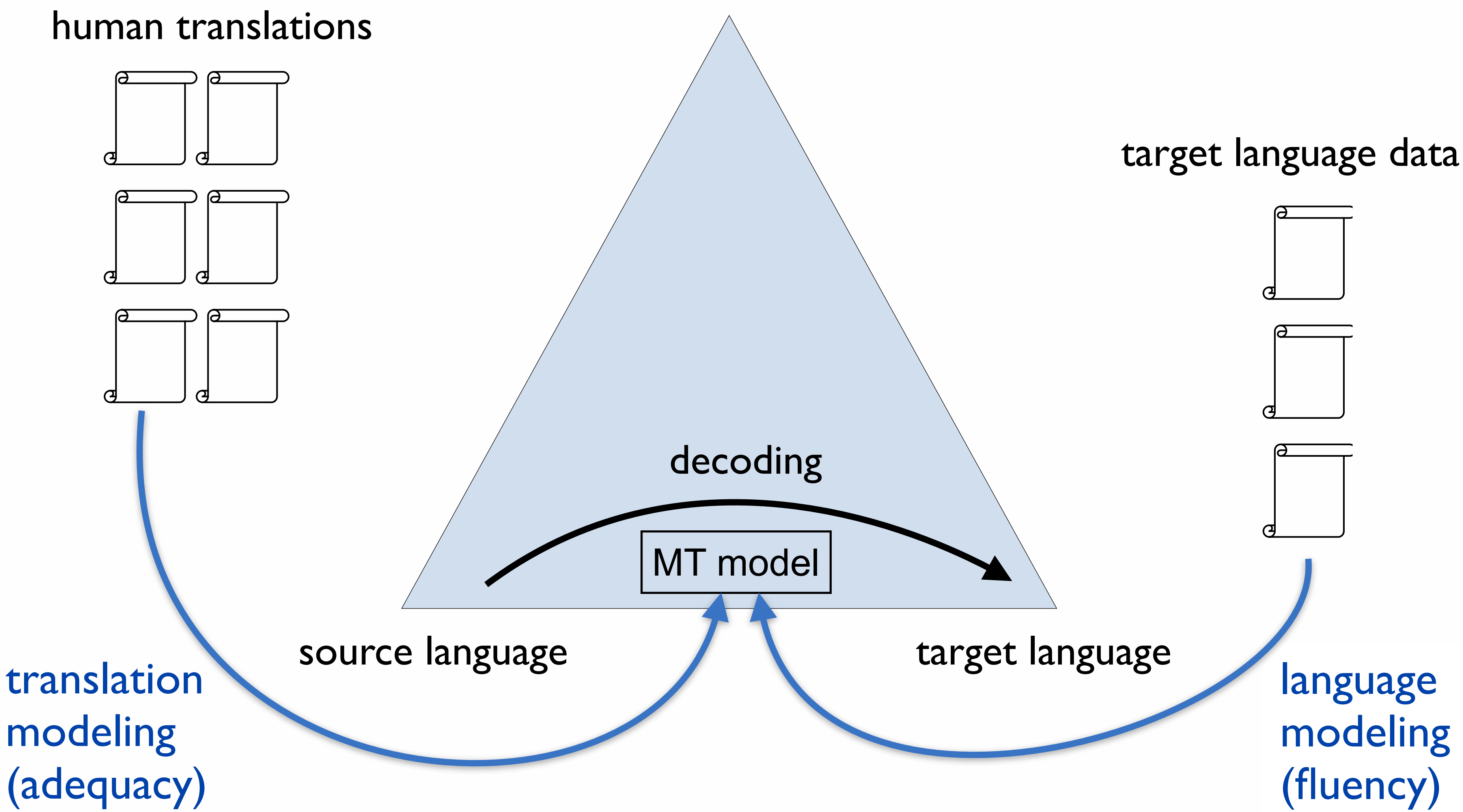
Data-Driven Machine Translation



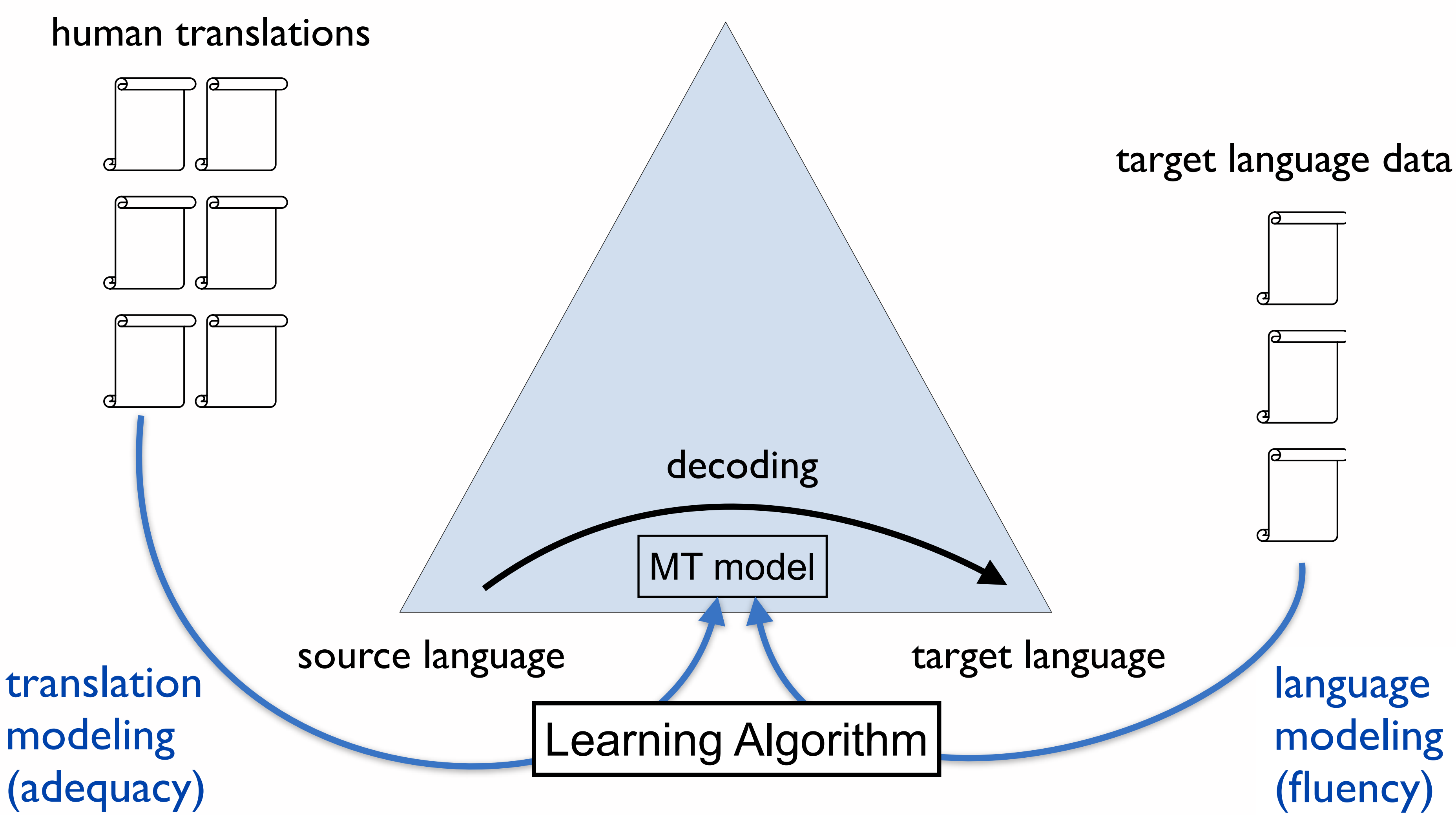
Data-Driven Machine Translation



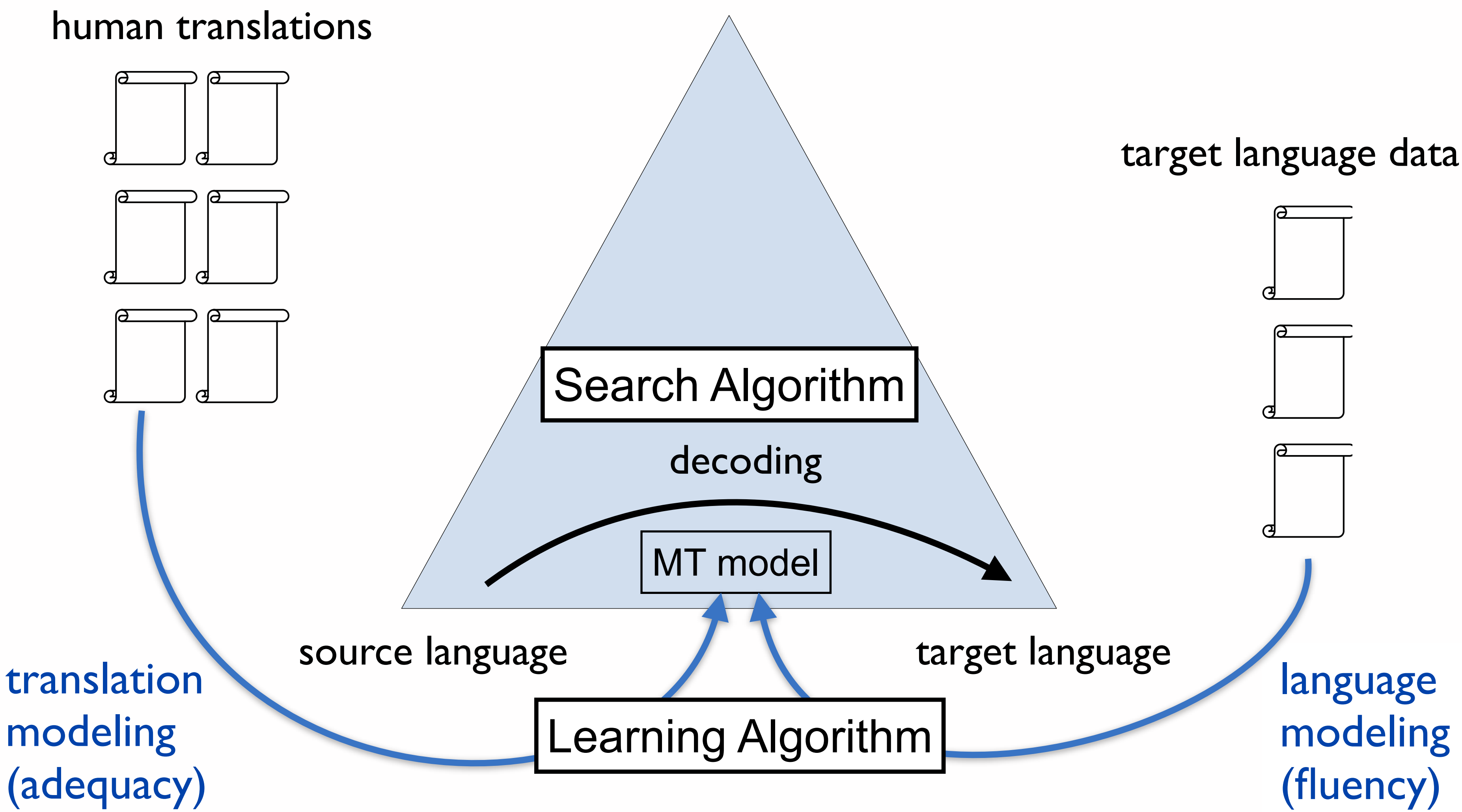
Data-Driven Machine Translation



Data-Driven Machine Translation



Data-Driven Machine Translation



How Does It Work?

Probabilistic Translation Models

- likelihood of a target language sentence **t** to be a good translation of the source language sentence **s**
- decompose into smaller components
- define how components may be combined

How Does It Work?

Probabilistic Translation Models

- likelihood of a target language sentence **t** to be a good translation of the source language sentence **s**
- decompose into smaller components
- define how components may be combined

Probabilistic Language Models

- likelihood of **t** to be a fluent target language sentence
- decompose into smaller components
- define how components may be combined

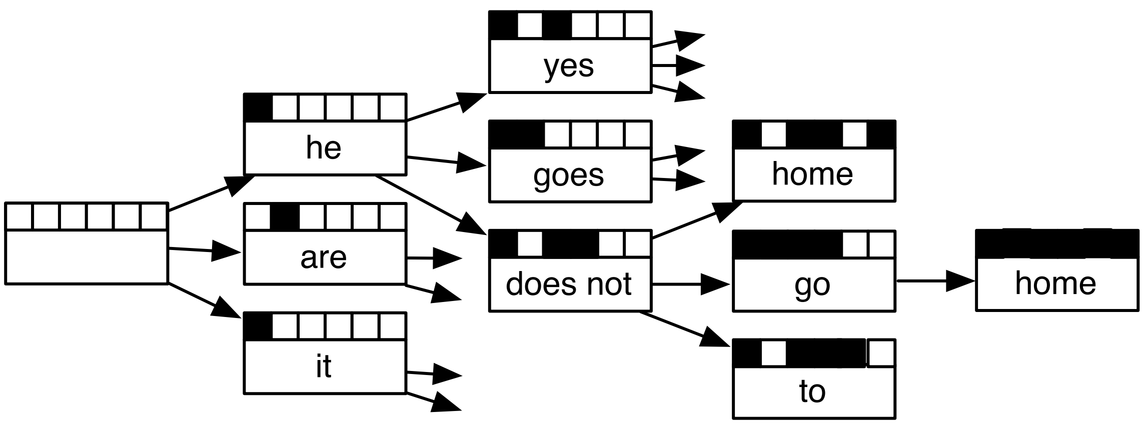
How Does It Work?

Translation = Decoding

- given a probabilistic model of translation
- find the most likely translation of a given sentence **s**

Search Problem

- many possible translation options
- many combinations of the various components



Statistical Machine Translation

1947: MT as decoding (Warren Weaver)

1988: Word-based models

1999: Public implementation of word-based 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

Finding Patterns (Knight, 1997)

Your assignment, translate this to Arcturan: farok crrrok hihok yorok clok kantok ok-yurp

1a. ok-voon ororok sprok . translated sentence 1b. at-voon bichat dat .	7a. lalok farok ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok hihok yorok zanzanok . 11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok hihok mok . 12b. wat nnat forat arrat vat gat .

Database of example translations

Finding Patterns (Knight, 1997)

Your assignment, translate this to Arcturan: farok crrrok hihok yorok clok kantok ok-yurp

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok **farok** ororok lalok sprok izok enemok . 7b. wat **jjat** bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok **farok** izok stok . 5b. totat **jjat** quat cat .	11a. lalok nok crrrok hihok yorok zanzanok . 11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok hihok mok . 12b. wat nnat forat arrat vat gat .

Finding Patterns (Knight, 1997)

Your assignment, translate this to Arcturan: **farok** crrrok **hihok yorok** **clok** kantok ok-yurp

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok farok ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok hihok yorok zanzanok . 11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok hihok mok . 12b. wat nnat forat arrat vat gat .

process of elimination

Finding Patterns (Knight, 1997)

Your assignment, put these words in order: **{ jjat, arrat, mat, bat, oloat, at-yurp }**

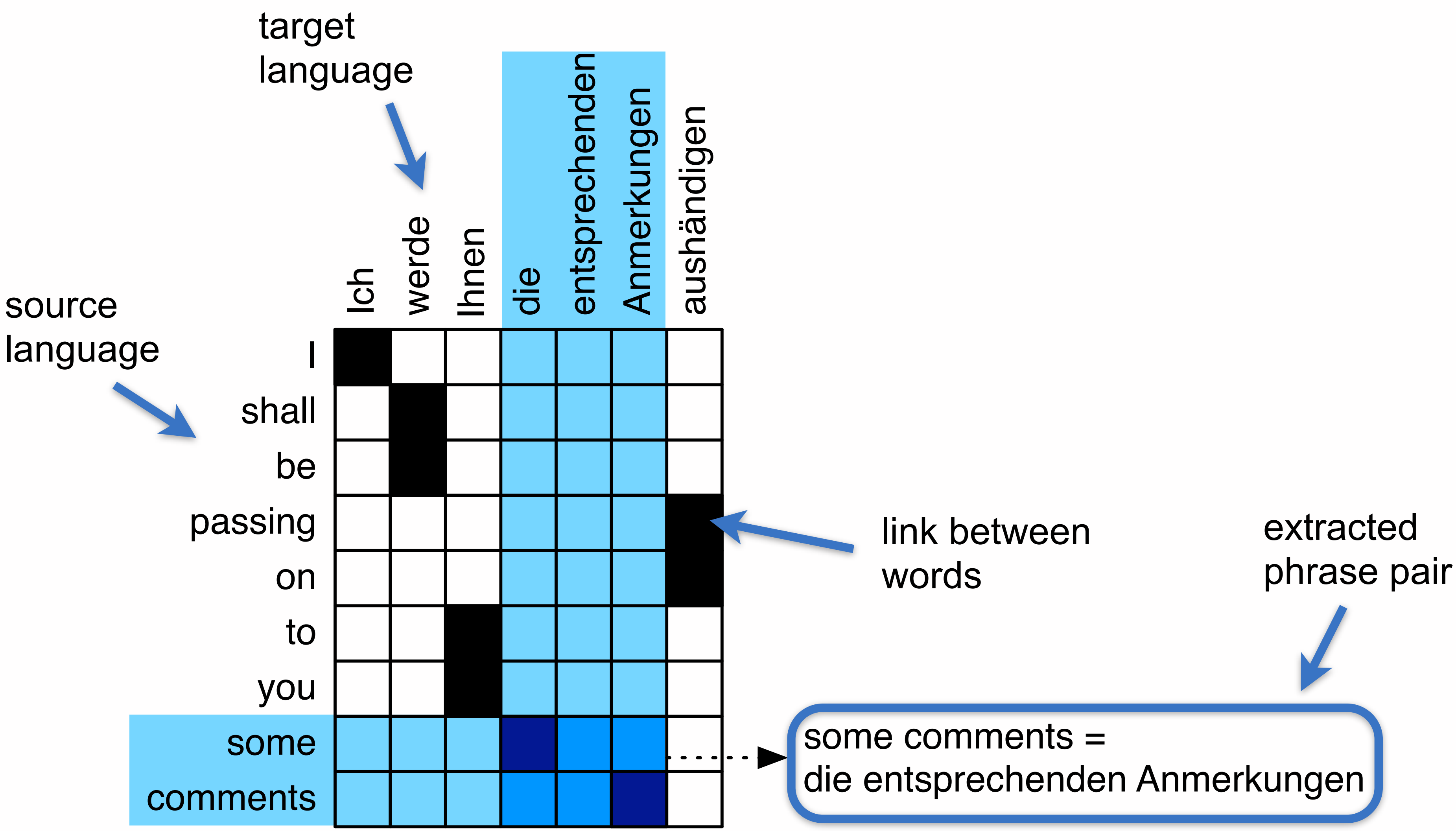
1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok farok ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok hihok yorok zanzanok . 11b. wat nnat arrat mat zanzanat . zero fertility
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok hihok mok . 12b. wat nnat forat arrat vat gat .

Finding Patterns (Knight, 1997)

Clients do not sell pharmaceuticals in Europe => Clientes no venden medicinas en Europa

1a. Garcia and associates . 1b. Garcia y asociados .	7a. the clients and the associates are enemies . 7b. los clients y los asociados son enemigos .
2a. Carlos Garcia has three associates . 2b. Carlos Garcia tiene tres asociados .	8a. the company has three groups . 8b. la empresa tiene tres grupos .
3a. his associates are not strong . 3b. sus asociados no son fuertes .	9a. its groups are in Europe . 9b. sus grupos estan en Europa .
4a. Garcia has a company also . 4b. Garcia tambien tiene una empresa .	10a. the modern groups sell strong pharmaceuticals . 10b. los grupos modernos venden medicinas fuertes .
5a. its clients are angry . 5b. sus clientes estan enfadados .	11a. the groups do not sell zenzanine . 11b. los grupos no venden zanzanina .
6a. the associates are also angry . 6b. los asociados tambien estan enfadados .	12a. the small groups are not modern . 12b. los grupos pequenos no son modernos .

Phrase Translation Extraction

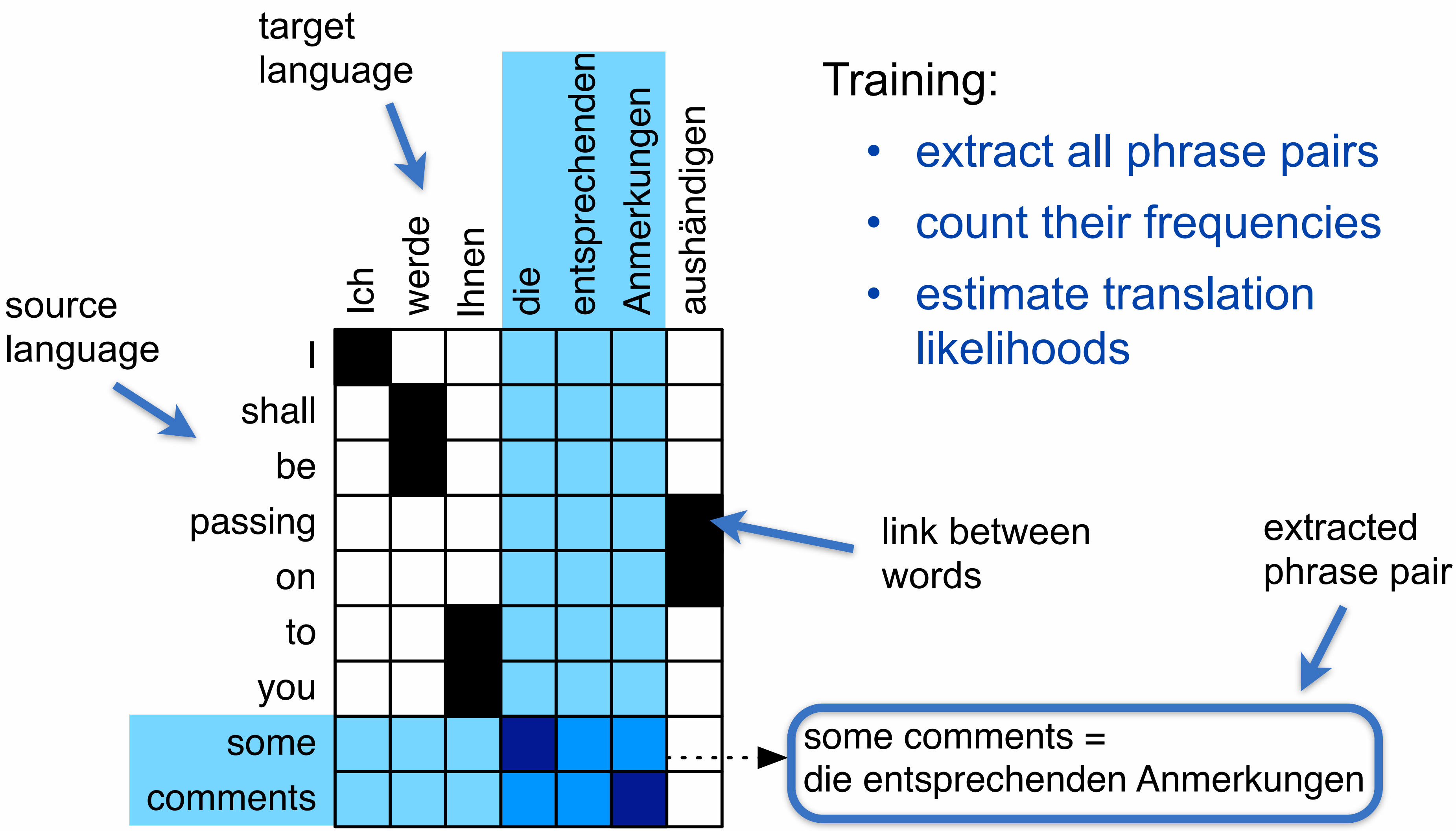


(illustration by Philip Williams and Philipp Koehn)

Phrase Translation Extraction

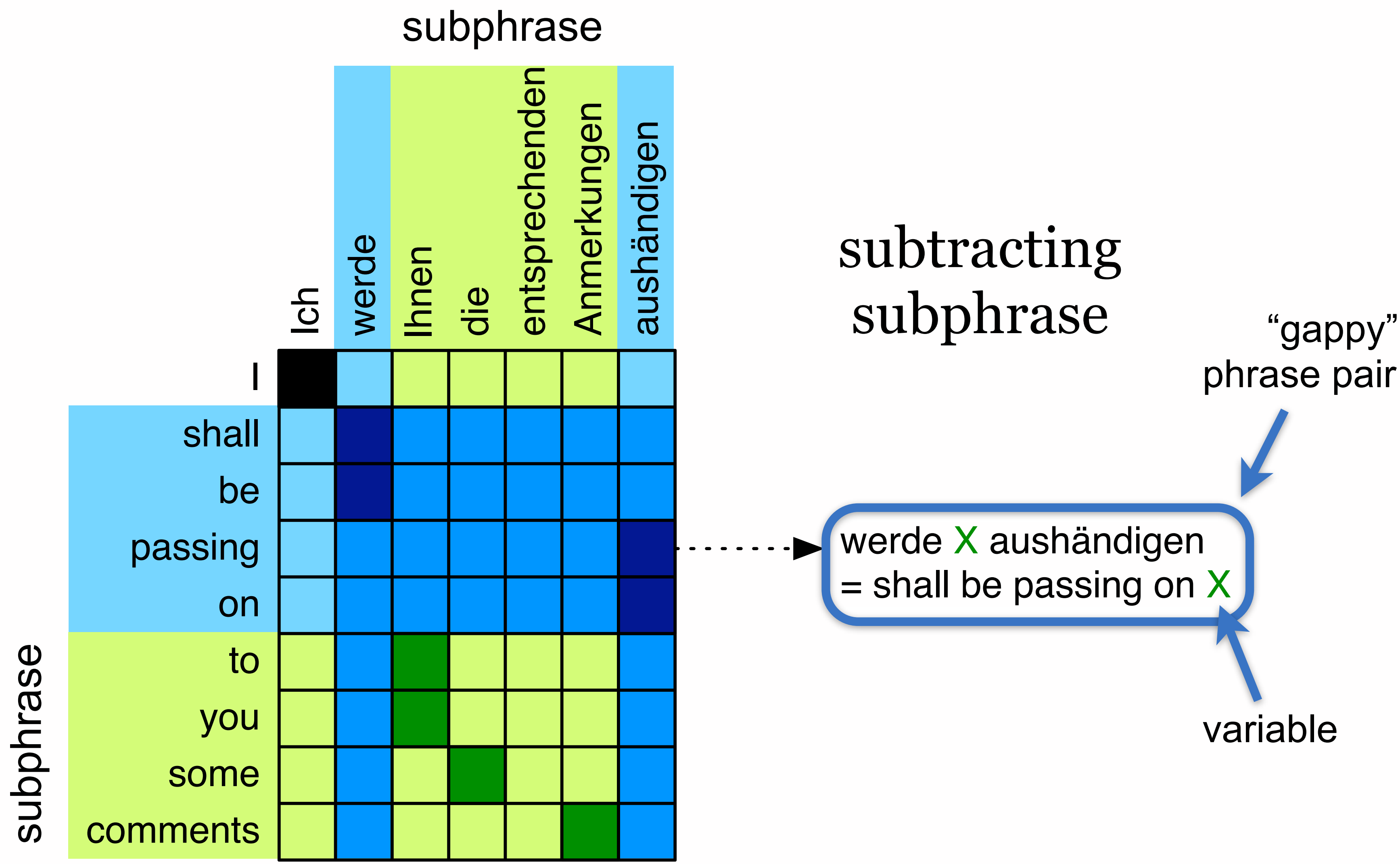
Training:

- extract all phrase pairs
- count their frequencies
- estimate translation likelihoods



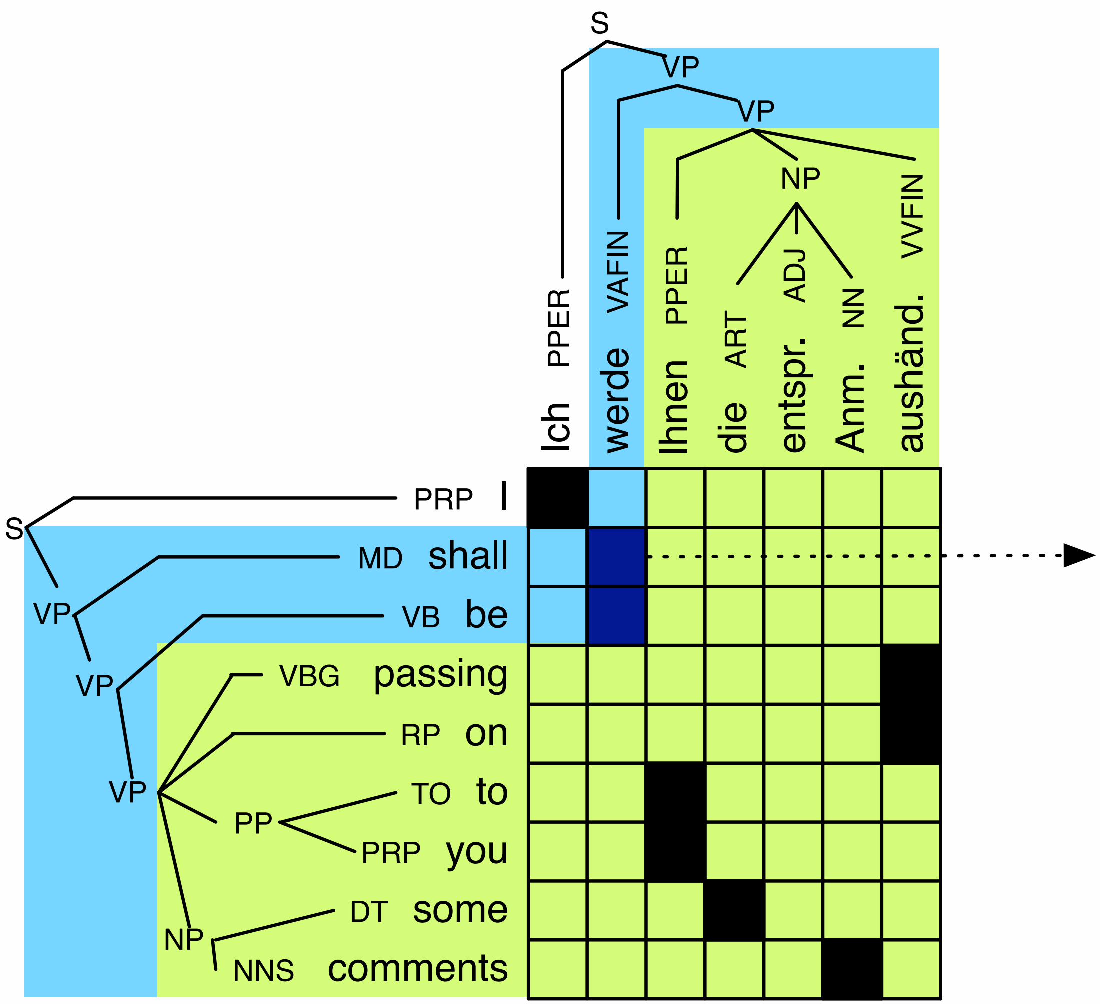
(illustration by Philip Williams and Philipp Koehn)

Hierarchical Phrase Translations



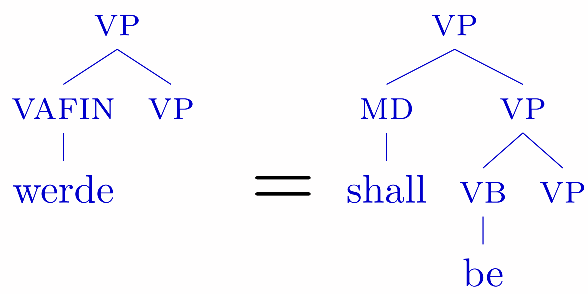
(illustration by Philip Williams and Philipp Koehn)

Syntactic Translation Rules



(illustration by Philip Williams and Philipp Koehn)

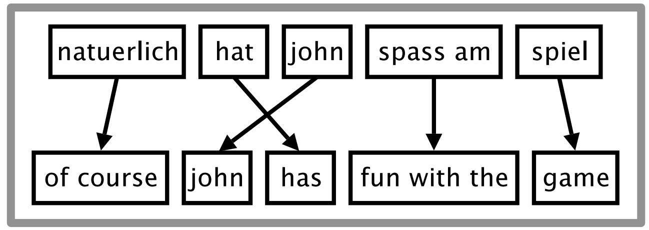
probabilistic synchronous grammars



Statistical MT Models

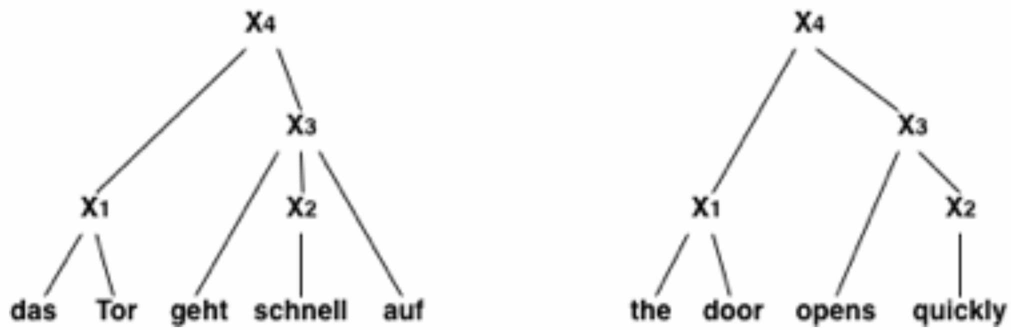
Phrase-Based SMT

- translation of fragments
- left-to-right beam search



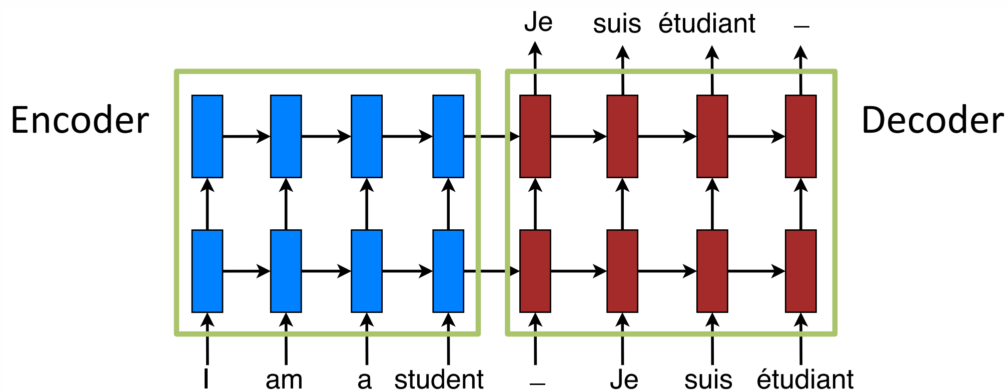
Syntax-Based SMT

- synchronous grammars
- translate = parsing



Neural MT

- continuous vector representations
- recurrent networks



The Situation for Finnish

The Situation for Finnish

Quoting from “The Finnish Language in the Digital Age”:

- “There are also a few applications for automatically translating language, even though these often fail to produce linguistically and idiomatically correct translations, especially when Finnish is the target language. This is partly due to the specific linguistic characteristics of the Finnish language.” (p.37)

- “Google and Microsoft provide statistical MT for Finnish, but the quality remains poor, due to the complexity of Finnish morphology and the free word order which current statistical MT is poorly equipped for.” (p.60)

The Situation for Finnish

Workshop on SMT 2015

- our winning contribution:

The Situation for Finnish

Workshop on SMT 2015

- our winning contribution:

Finnish - English

system	$BLEU$	TER
unconstrained		
baseline	18.9	0.737
primary	19.3	0.728
constrained		
baseline	15.5	0.780
factored	17.9	0.749

French - English: 33.1
German - English: 29.3

The Situation for Finnish

Workshop on SMT 2015

- our winning contribution:

Finnish - English

system	$BLEU$	TER
unconstrained		
baseline	18.9	0.737
primary	19.3	0.728
constrained		
baseline	15.5	0.780
factored	17.9	0.749

French - English: 33.1
German - English: 29.3

English - Finnish

system	$BLEU_{dev}$	$BLEU$	TER
constrained	12.7	10.7	0.842
unconstrained	15.7	14.8	0.796

English - French: 33.6
English - German: 24.9

What’s Wrong With Finnish?

What's Wrong With Finnish?

Nothing!

- The problem is in the models we use!

What's Wrong With Finnish?

Nothing!

- The problem is in the models we use!

What current SMT cannot cope with well:

- rich inflectional systems (marking case, person, …)
- case and number agreement (over long distances)
- derivation and composition
- morphophonological alternations
- (relatively) free word order

What's Wrong With Finnish?

Nothing!

- The problem is in the models we use!

What current SMT cannot cope with well:

- rich inflectional systems (marking case, person, …)
- case and number agreement (over long distances)
- derivation and composition
- morphophonological alternations
- (relatively) free word order

Well, that's Finnish ...

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

What's Wrong With Finnish?

Nothing!

- The problem is in the models we use!

15 grammatical cases
nouns: ca. 2,000 forms
verbs: > 12,000 forms

What current SMT cannot cope with well:

- rich inflectional systems (marking case, person, …)
- case and number agreement (over long distances)
- derivation and composition
- morphophonological alternations
- (relatively) free word order

Well, that's Finnish ...

What's Wrong With Finnish?

Nothing!

- The problem is in the models we use!

15 grammatical cases
nouns: ca. 2,000 forms
verbs: > 12,000 forms

What current SMT cannot cope with well:

- rich inflectional systems (marking case, person, …)
- case and number agreement (over long distances)
- derivation and composition
- morphophonological alternations
- (relatively) free word order

word types in Finnish:
20-30% derivatives
60-70% compounds

Well, that's Finnish ...

What's Wrong With Finnish?

Nothing!

- The problem is in the models we use!

15 grammatical cases
nouns: ca. 2,000 forms
verbs: > 12,000 forms

What current SMT cannot cope with well:

- rich inflectional systems (marking case, person, …)
- case and number agreement (over long distances)
- derivation and composition
- morphophonological alternations
- (relatively) free word order

word types in Finnish:
20-30% derivatives
60-70% compounds

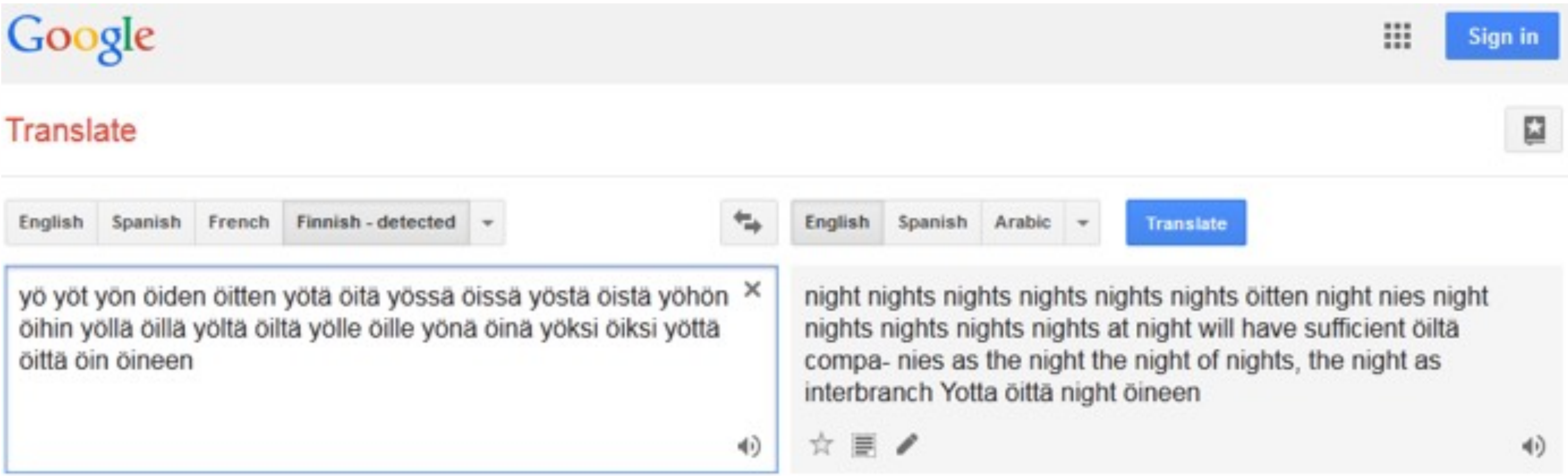
consonant gradation
vowel harmony
...

Well, that's Finnish ...

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

What's Wrong With Our Models?

Research and development is focused on English …



- *"In a selection of leading conferences and scientific journals published between 2008 and 2010, there were 971 publications on language technology for English and only 10 for Finnish."*

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

What Do We Need To Change?

(The Finnish Language in the Digital Age)

What Do We Need To Change?

Open resources and tools

- *"Due to early commercial successes for Finnish language technology, the availability of basic tools such as parsers and lexicons in the research community for processing Finnish became limited. As an odd consequence, technology specifically adapted to the Finnish language was only marginally involved in Finnish research projects and therefore most of the research and development prototypes used English."*

(The Finnish Language in the Digital Age)

What Do We Need To Change?

Open resources and tools

- *"Due to early commercial successes for Finnish language technology, the availability of basic tools such as parsers and lexicons in the research community for processing Finnish became limited. As an odd consequence, technology specifically adapted to the Finnish language was only marginally involved in Finnish research projects and therefore most of the research and development prototypes used English."*

Funding for basic research

- *"After this decline in language technology basic research funding in Finland, many experts migrated to diverse small companies."*

(The Finnish Language in the Digital Age)

What Do We Need To Change?

Open resources and tools

- *"Due to early commercial successes for Finnish language technology, the availability of basic tools such as parsers and lexicons in the research community for processing Finnish became limited. As an odd consequence, technology specifically adapted to the Finnish language was only marginally involved in Finnish research projects and therefore most of the research and development prototypes used English."*

Funding for basic research

- *"After this decline in language technology basic research funding in Finland, many experts migrated to diverse small companies."*

Funding for workflow integration

- *"After the period of basic research funding only small scale industrial project funding has been provided by Tekes" .... "As a result, Finland" ... "lost some very promising high-tech innovations to the US"*

(The Finnish Language in the Digital Age)

Summary

Data-driven machine translation

- learn from data (without supervision)
- various models and production systems
- Neural MT offers better abstraction

MT in translation workflows

- speed and sufficient quality
- urgent need for more (in-domain) data
- better support for non-English languages!

Questions?

Derivation and Composition in Finnish

Agglutination: halu+tu+imm+i+lla+mme+ko
("desire, something that is, most, on, our, question")

20-30% of Word Types are Derivatives:

kirja [book] - kirjasto [library], kirjaamo [registry], kirjallisuus [literature], kirjoittaa [to write], kirjanen [booklet], kirjallinen [literary] etc.

60-70% of Word Types are Compounds:

maahanmuutto [immigration], kansaneläkelaitos [Social Insurance Institution], yleisurheilumaaottelu [international event in athletics].

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Many Open Challenges

Sufficient in-domain training data

- only for a small fraction of the World's languages
- only for a few textual domains

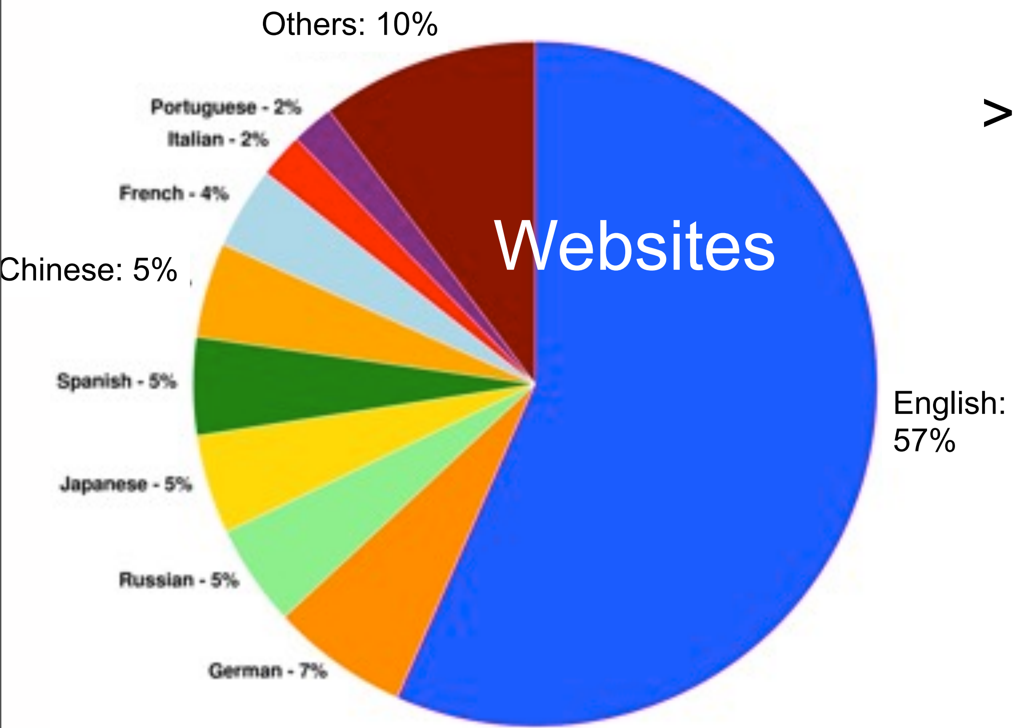
Morphology and syntax

- productive morphology requires special treatment
- word order differences and syntactic flexibility

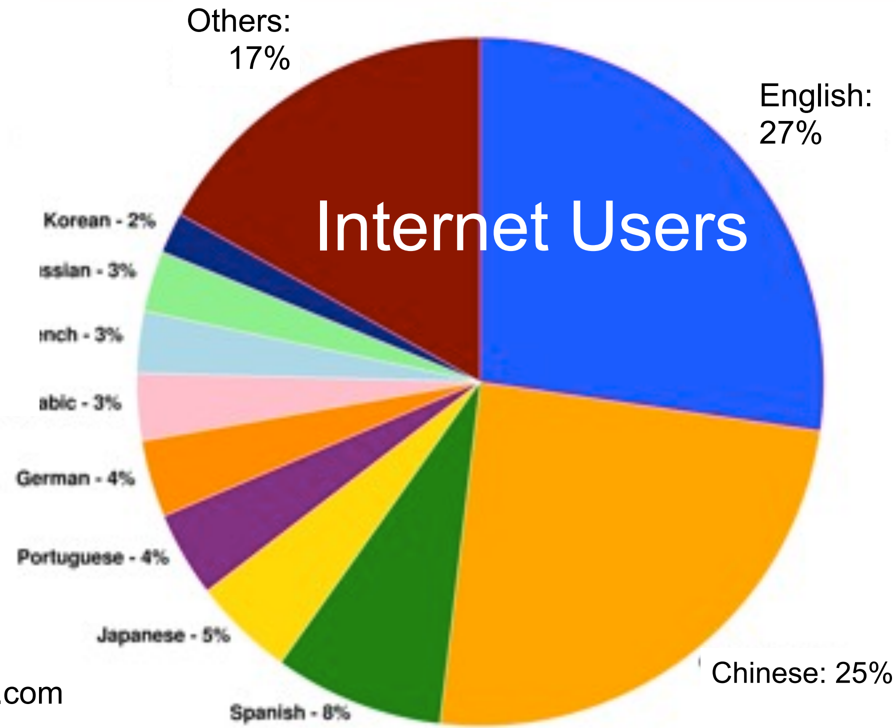
Noisy data

- quality of the training data
- noisy translation input (social media, speech, ...)

Why Machine Translation?



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



Cross-Lingual Information Access!

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

UNIVERSITY OF HELSINKI

Why is Translation so Difficult?

Why is Translation so Difficult?

Natural languages are ambiguous

- lexical, morphological, syntactic ambiguities
- no well-defined categorical concepts

Why is Translation so Difficult?

Natural languages are ambiguous

- lexical, morphological, syntactic ambiguities
- no well-defined categorical concepts

Natural languages are different

- lexical semantics, grammar
- style, culture, etymology

Why is Translation so Difficult?

Natural languages are ambiguous

- lexical, morphological, syntactic ambiguities
- no well-defined categorical concepts

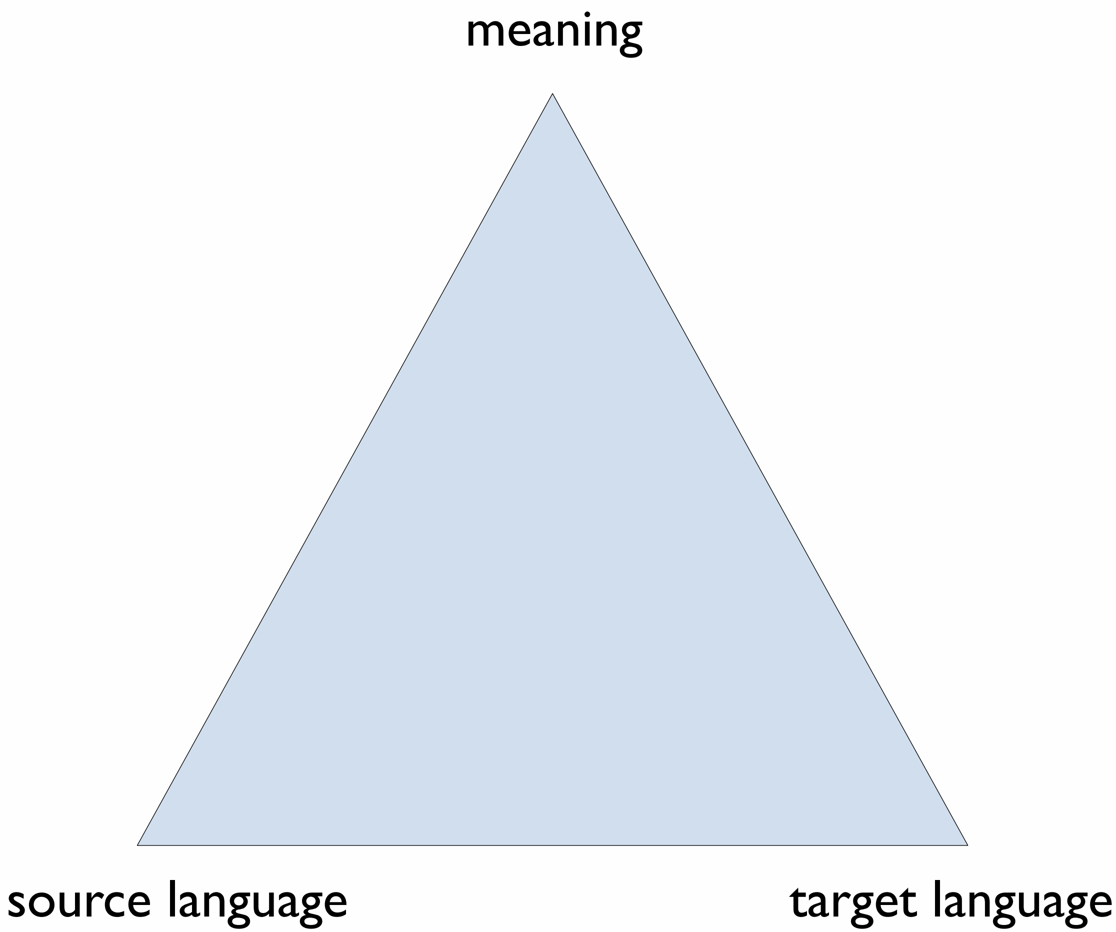
Natural languages are different

- lexical semantics, grammar
- style, culture, etymology

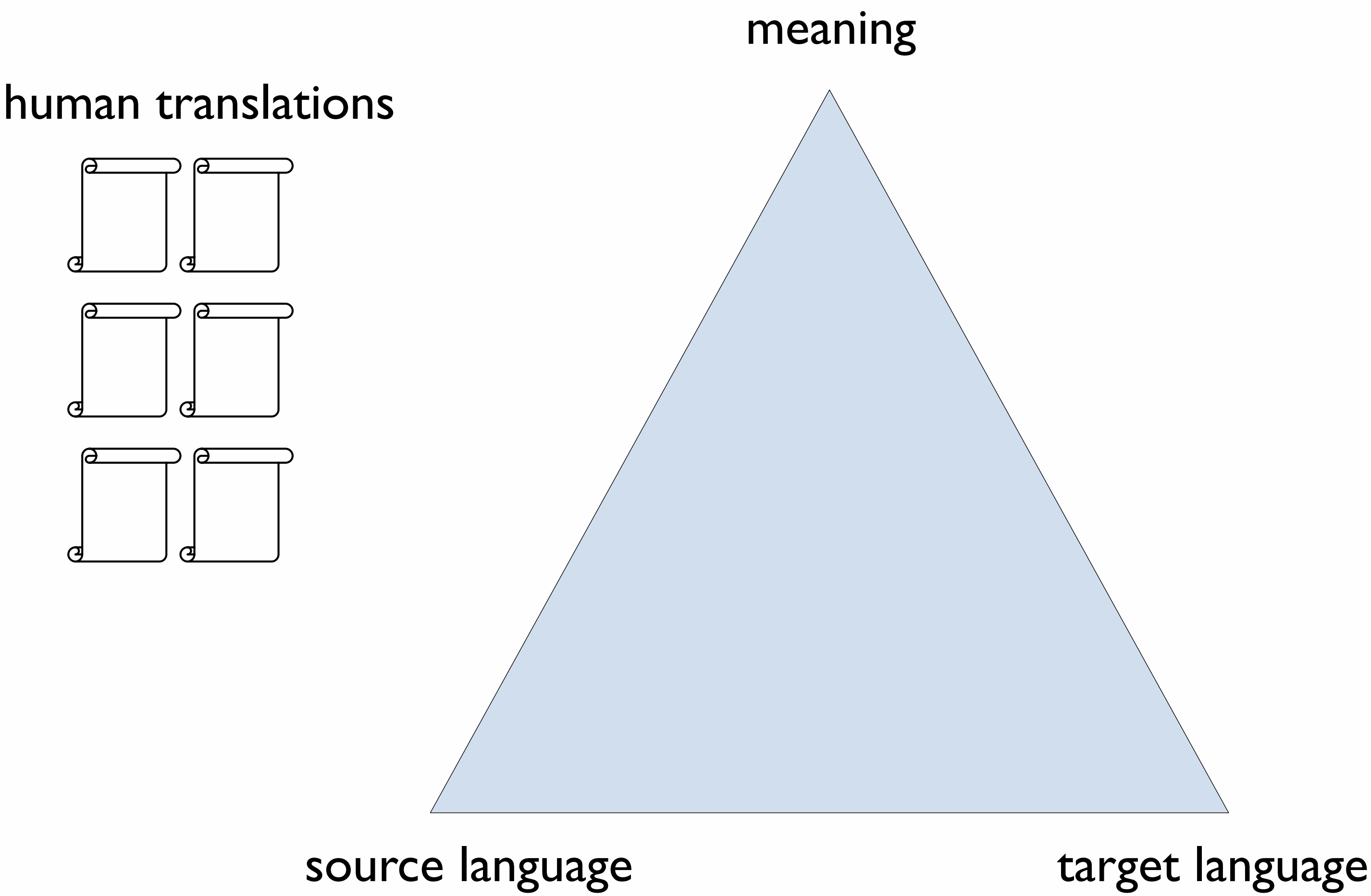
Natural redundancy and variation in languages

- stylistic and rhetorical variation
- dynamic, productive, language change

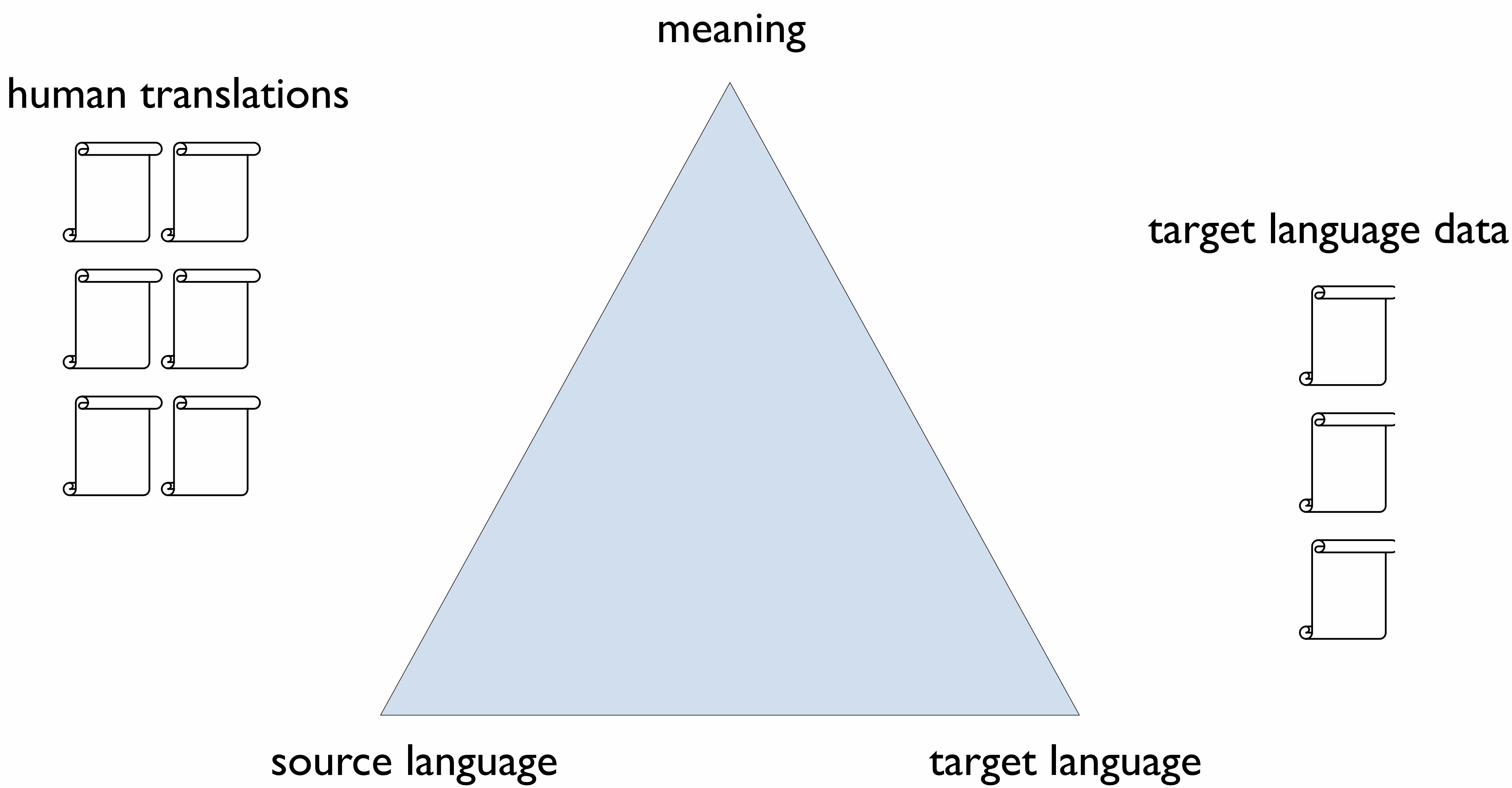
Data-Driven Machine Translation



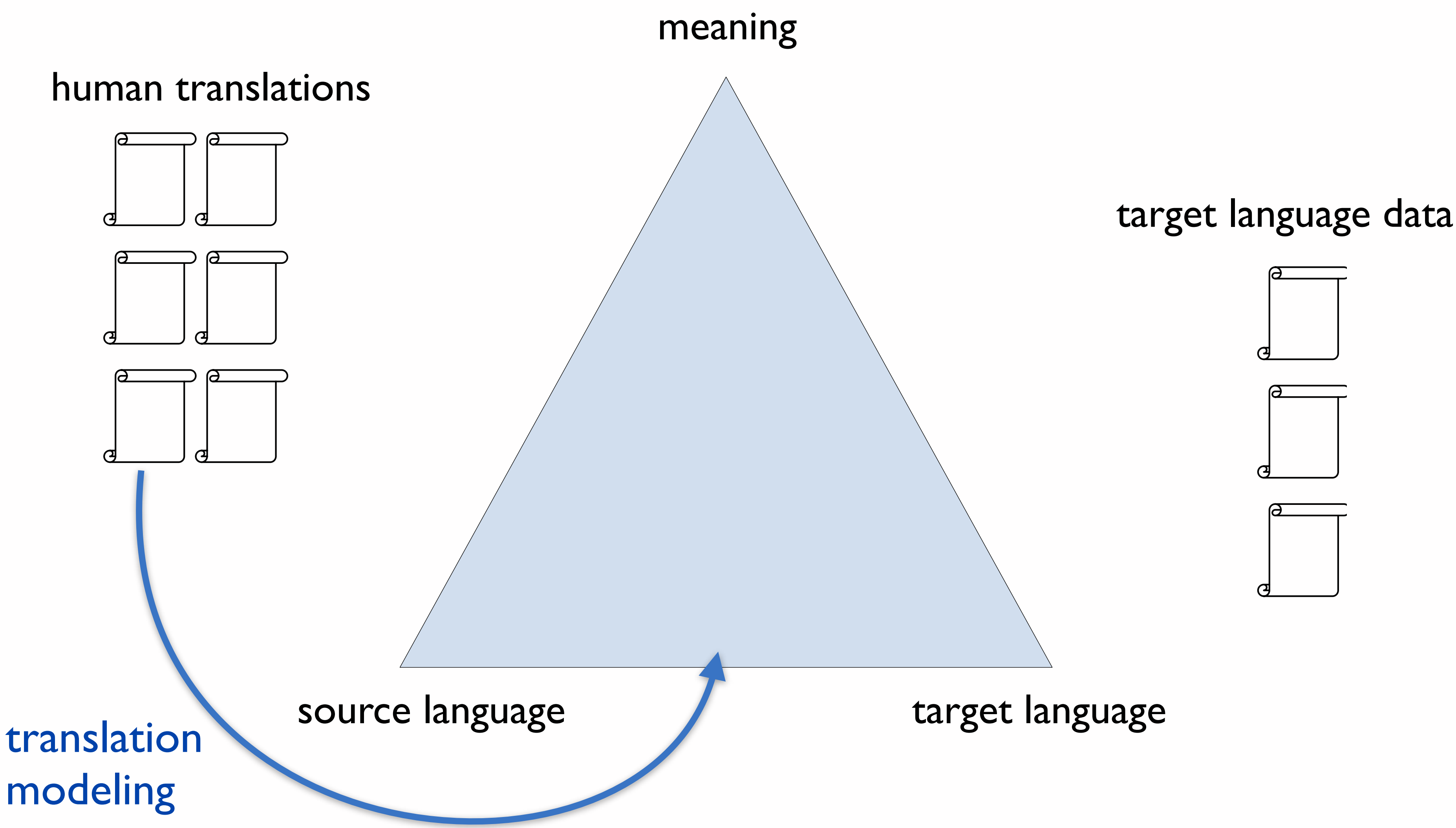
Data-Driven Machine Translation



Data-Driven Machine Translation

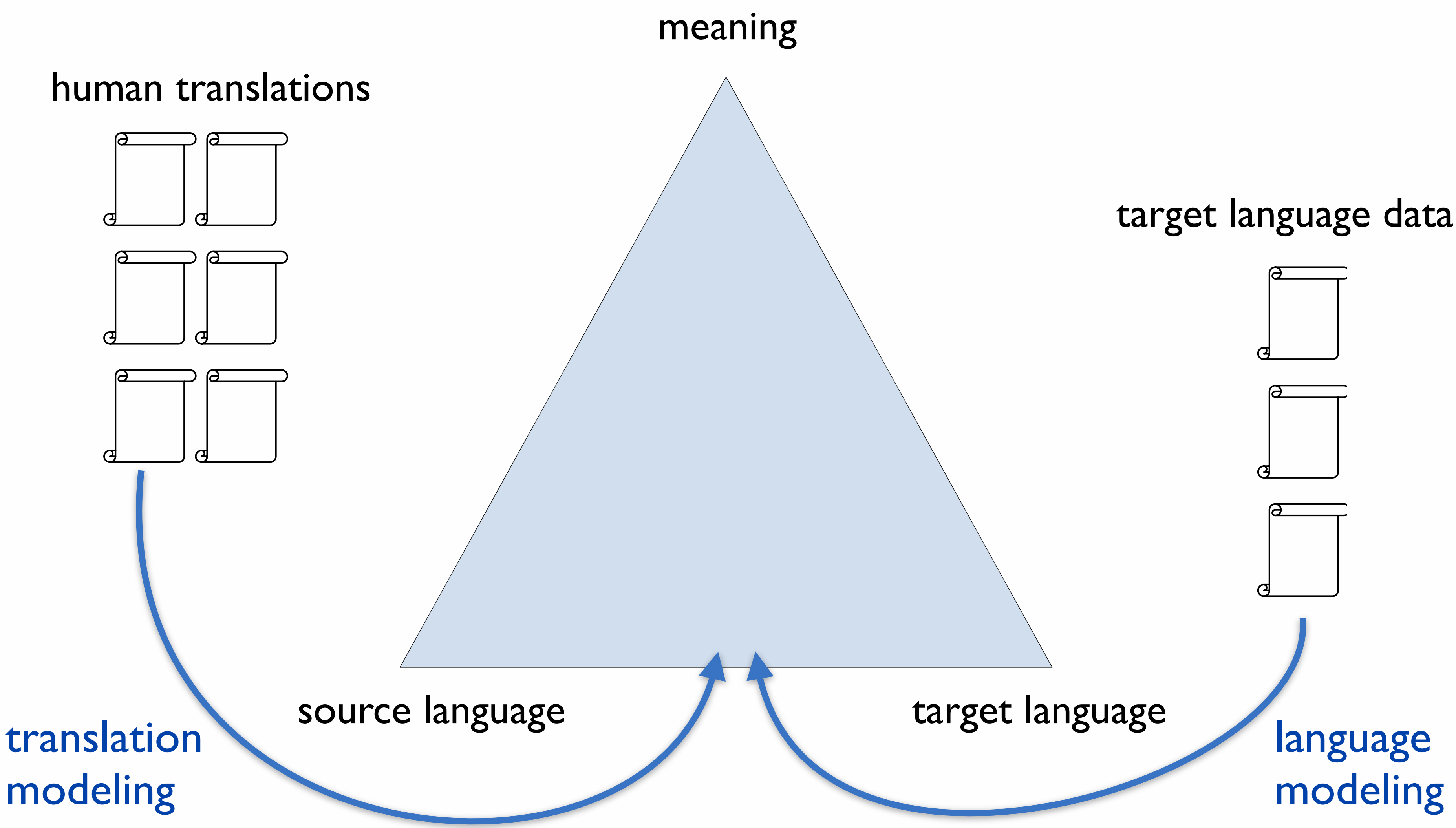


Data-Driven Machine Translation

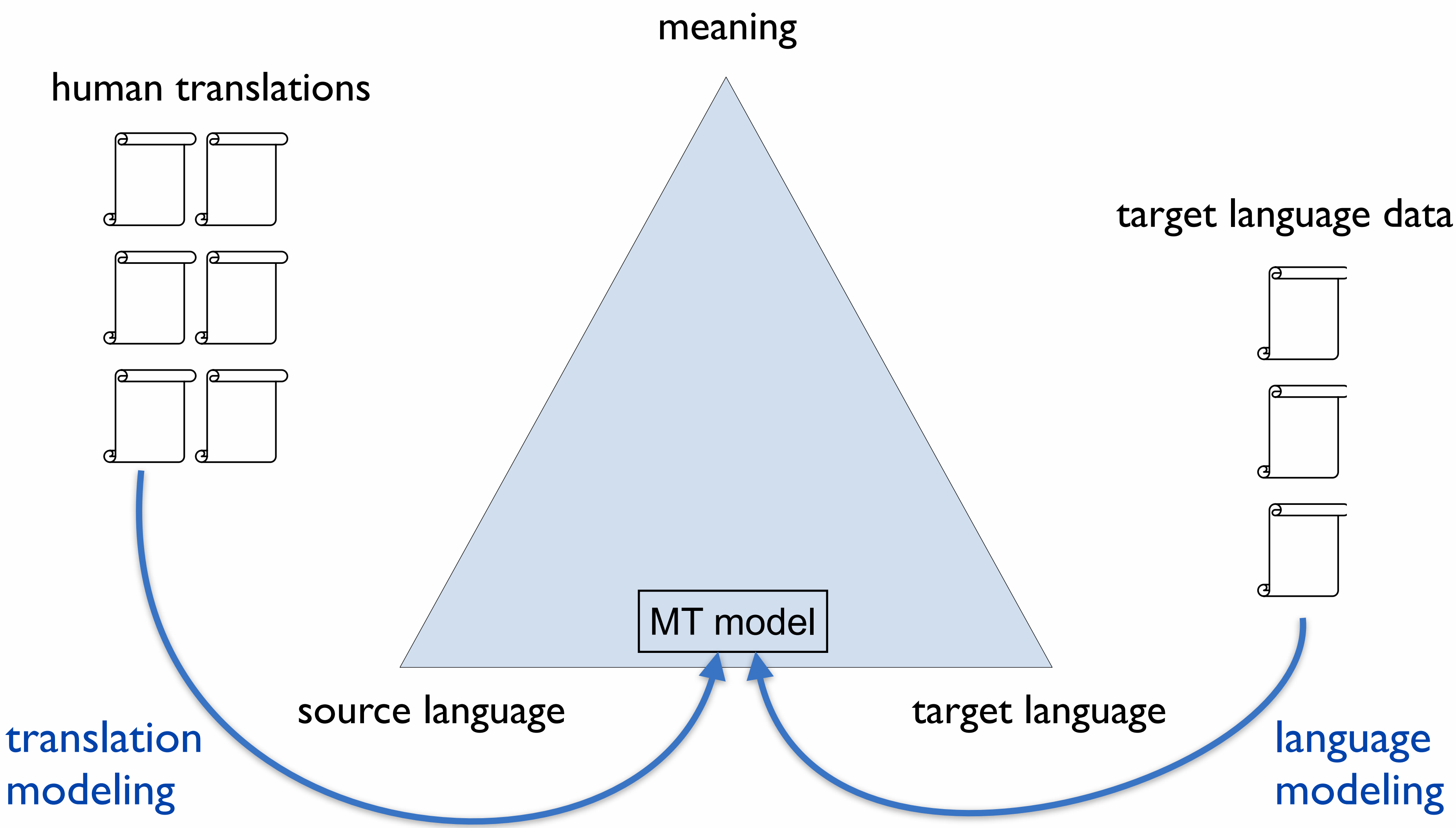


HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

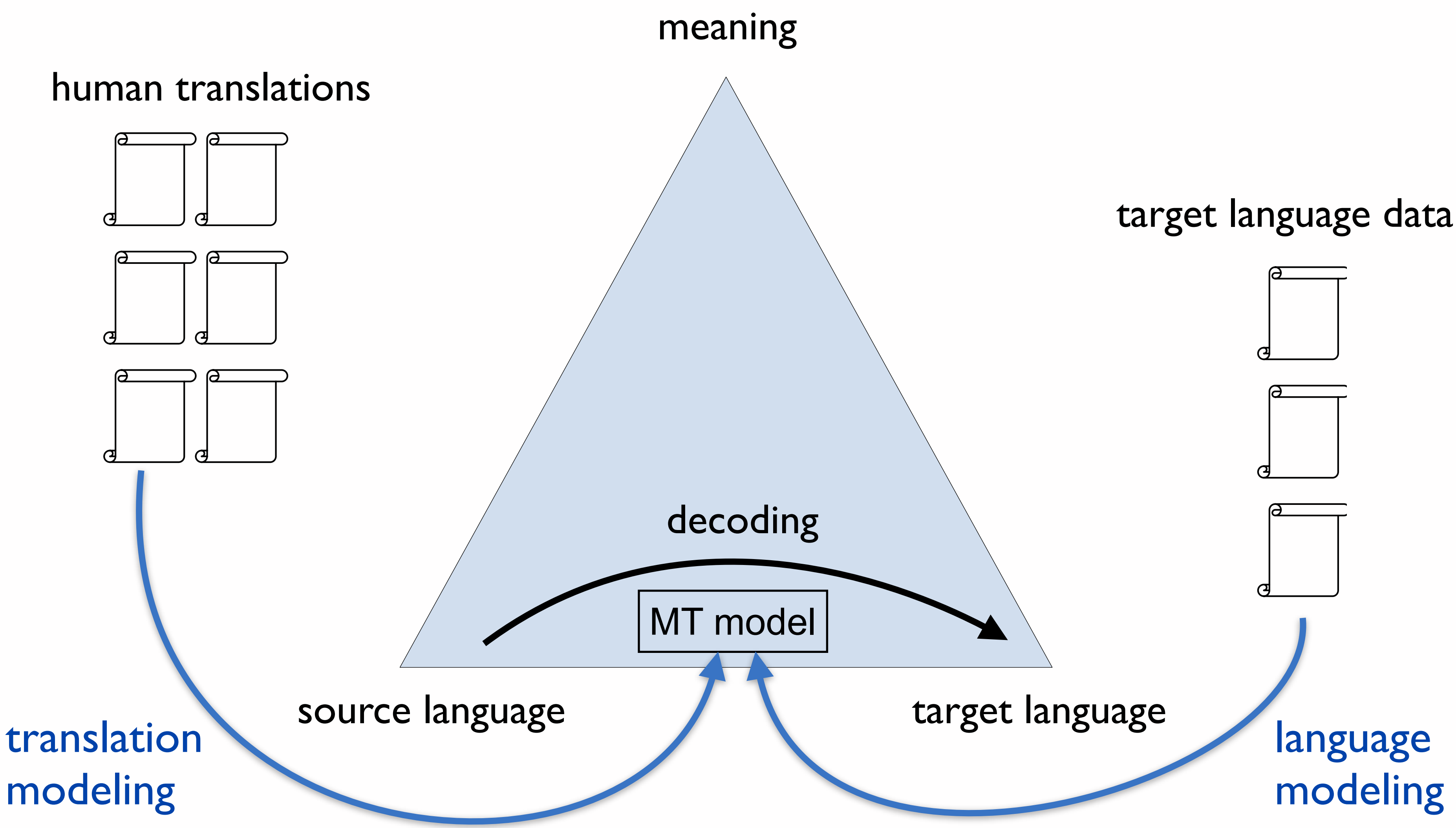
Data-Driven Machine Translation



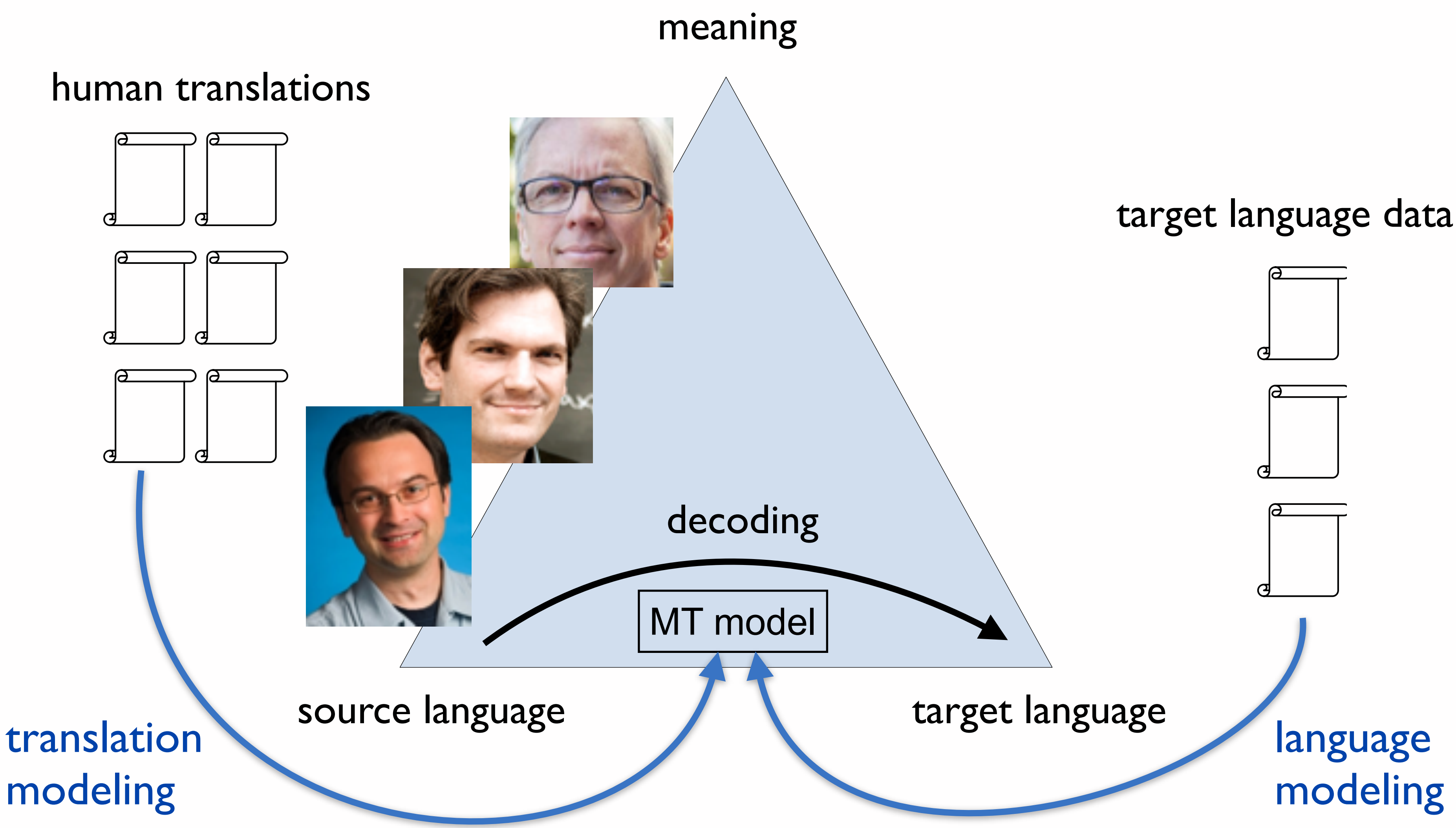
Data-Driven Machine Translation



Data-Driven Machine Translation

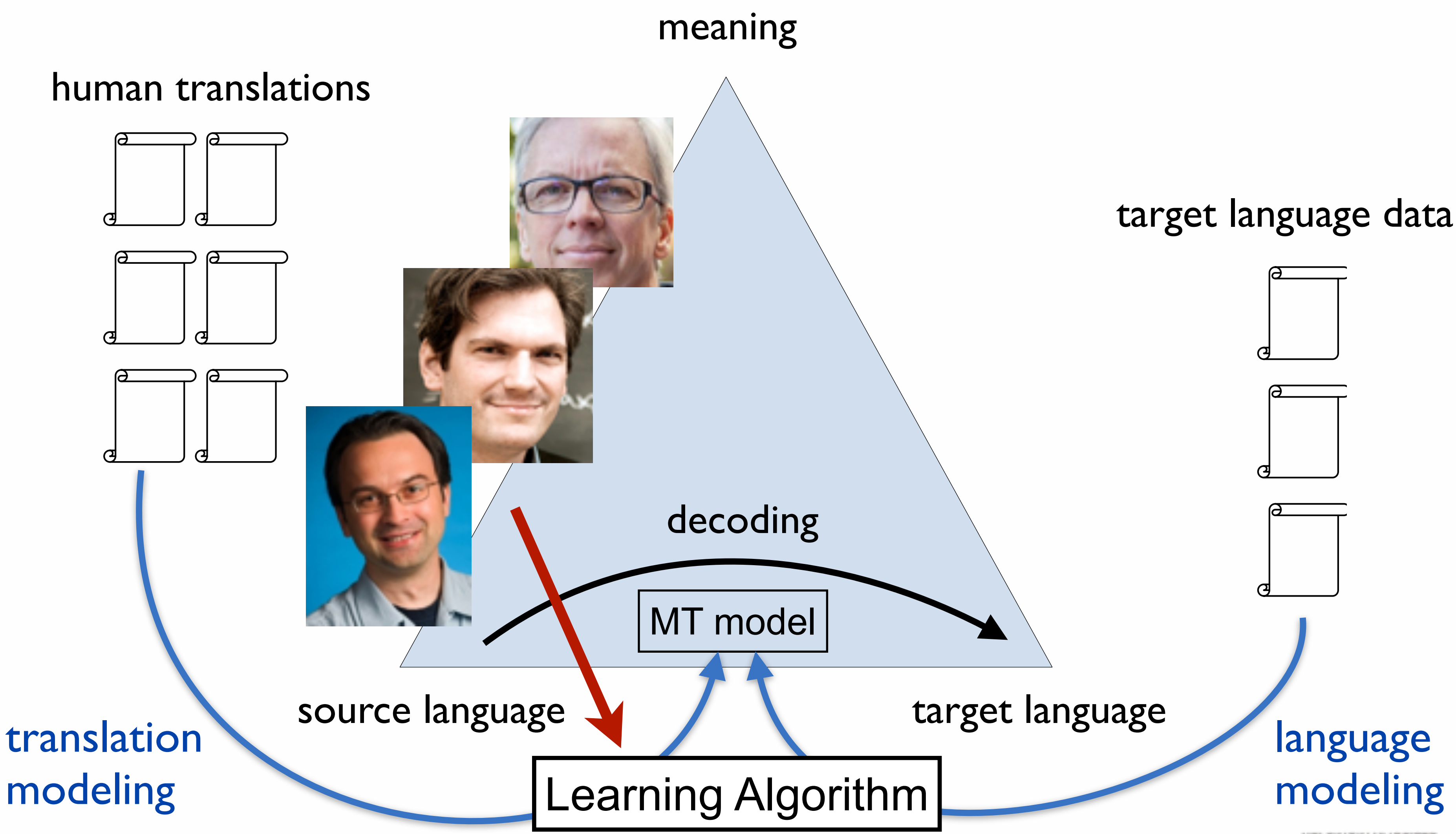


Data-Driven Machine Translation



HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Data-Driven Machine Translation



HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Estimating Parameters

Translation models

- collect statistics over mappings in training data
- estimate translation likelihoods

Language models

- collect statistics over words in context
- estimate probabilistic language models

MT models

- tune weights of various components

Workflow Integration

Sufficient quality

- (lots of) domain-specific training data
- customised systems
- task-specific optimisation

MT needs to be fast

- reasonable training times (especially for customisation)
- quick translation (a real problem)

Workbench integration

- accessibility from translation tools
- reliable service