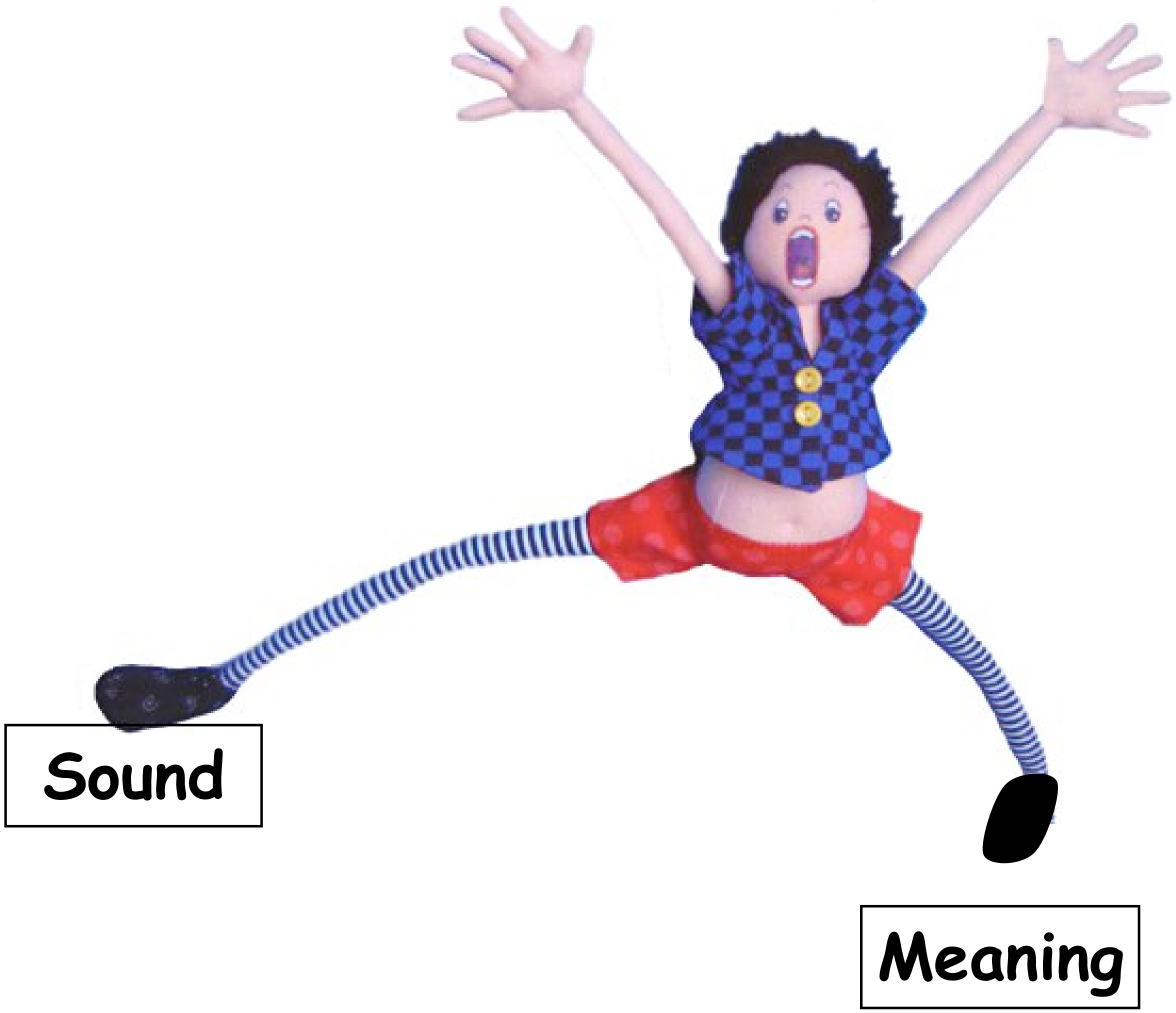
Machine Translation

Martin Kay

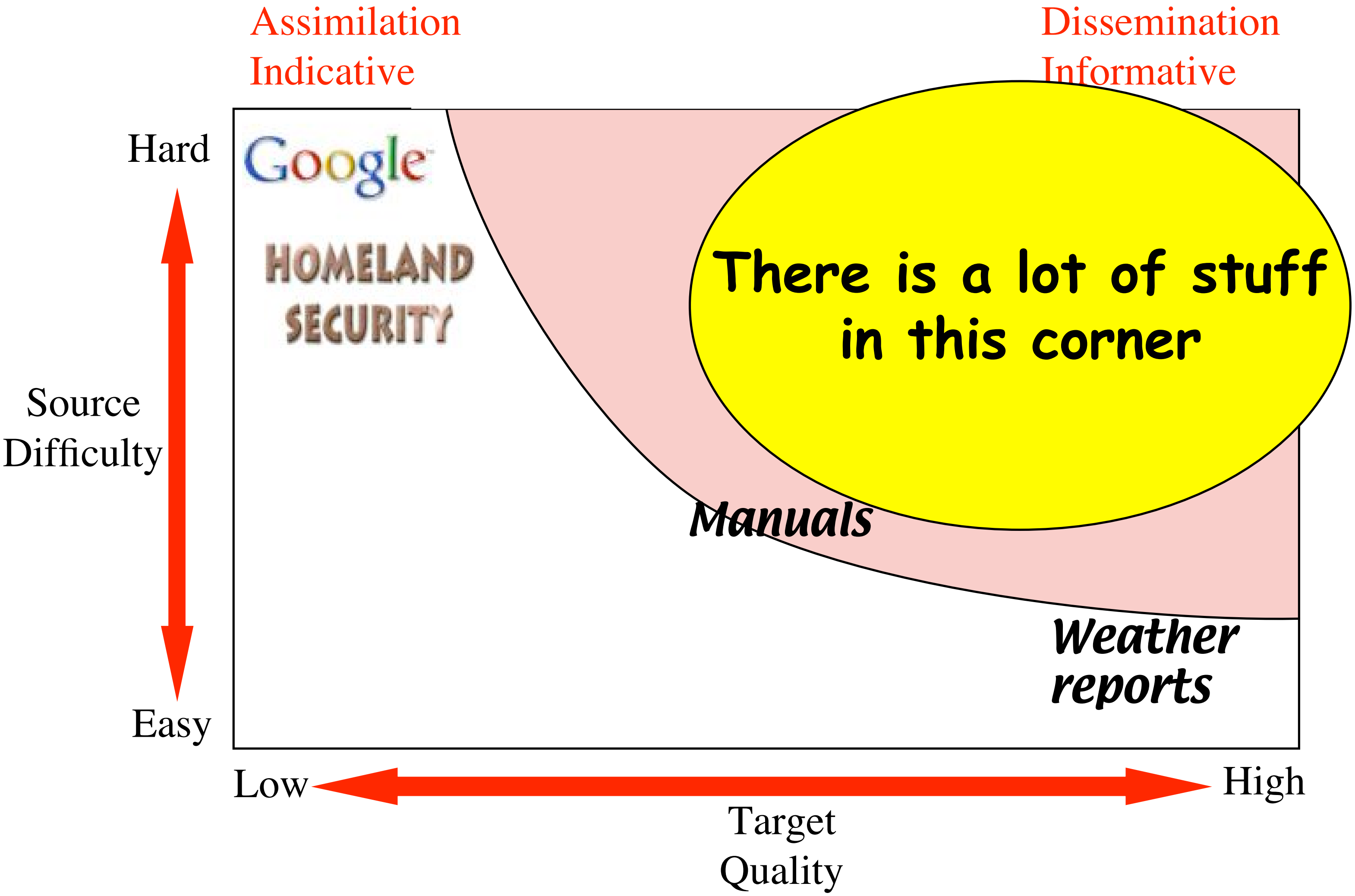
Stanford University and
The University of the Saarland

Language



Sound

Meaning



Assimilation
Indicative

Dissemination
Informative

Hard

Source
Difficulty

Easy

Low

High

Target
Quality

There is a lot of stuff in this corner

Manuals

Weather reports

The European Union

Danish
Dutch
English
Finnish
French
German
Greek
Italian
Portuguese
Spanish
Swedish

11 languages
2,500 (12.5%) of 20,000 staff
1% of the annual budget
40% of administration costs.

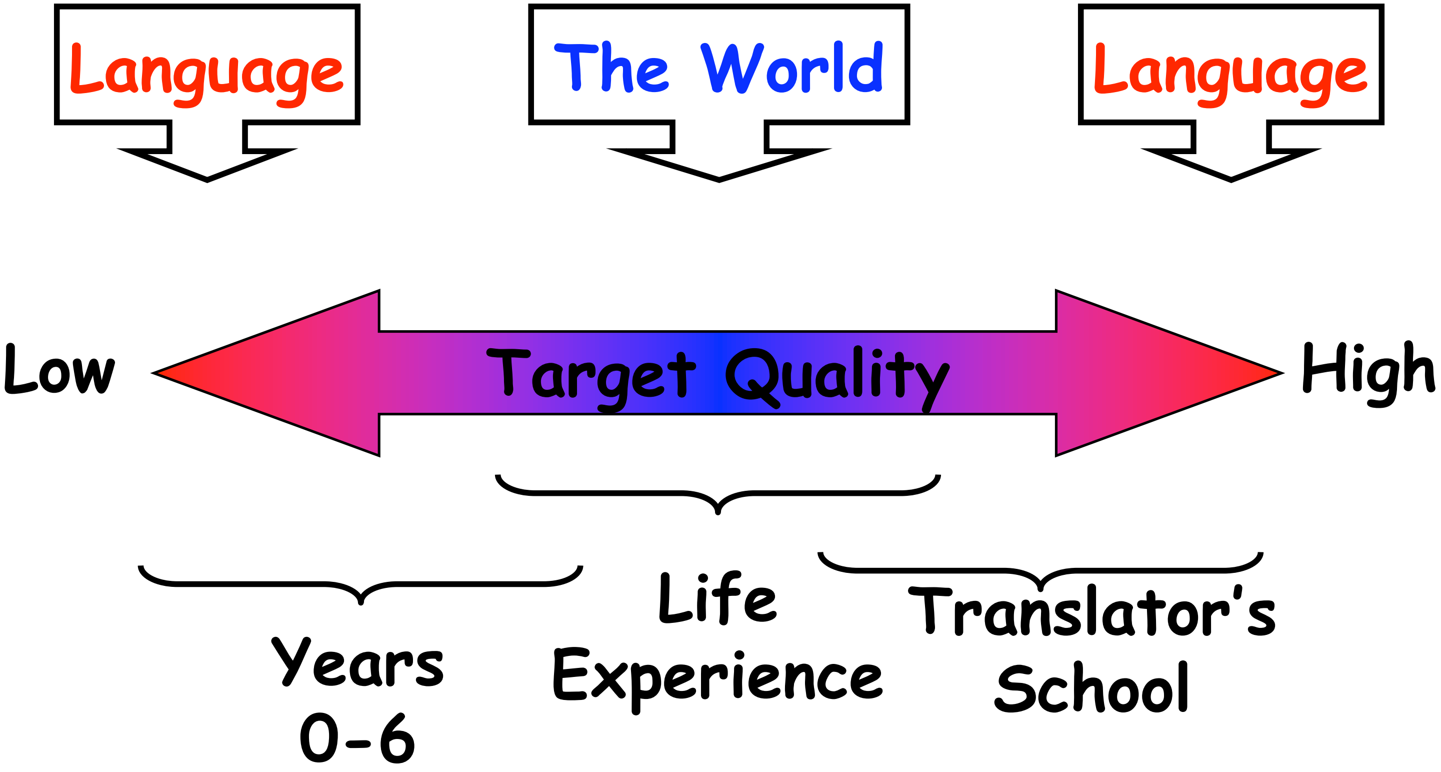
Czech
Estonian
Hungarian
Lithuanian
Latvian
Maltese
Polish
Slovene
Slovak

CATERPILLAR

300 authors and illustrators
800 English pages per day
Translation into 14 languages

Maintenance Manuals
Operation and Troubleshooting Guides
Disassembly and Specifications Manuals
Assembly Manuals
Testing and Special Instructions
Adjustment Guides
Systems Operation Bulletins

Why are the hard cases hard?



Language

The World

Language

Low

Target Quality

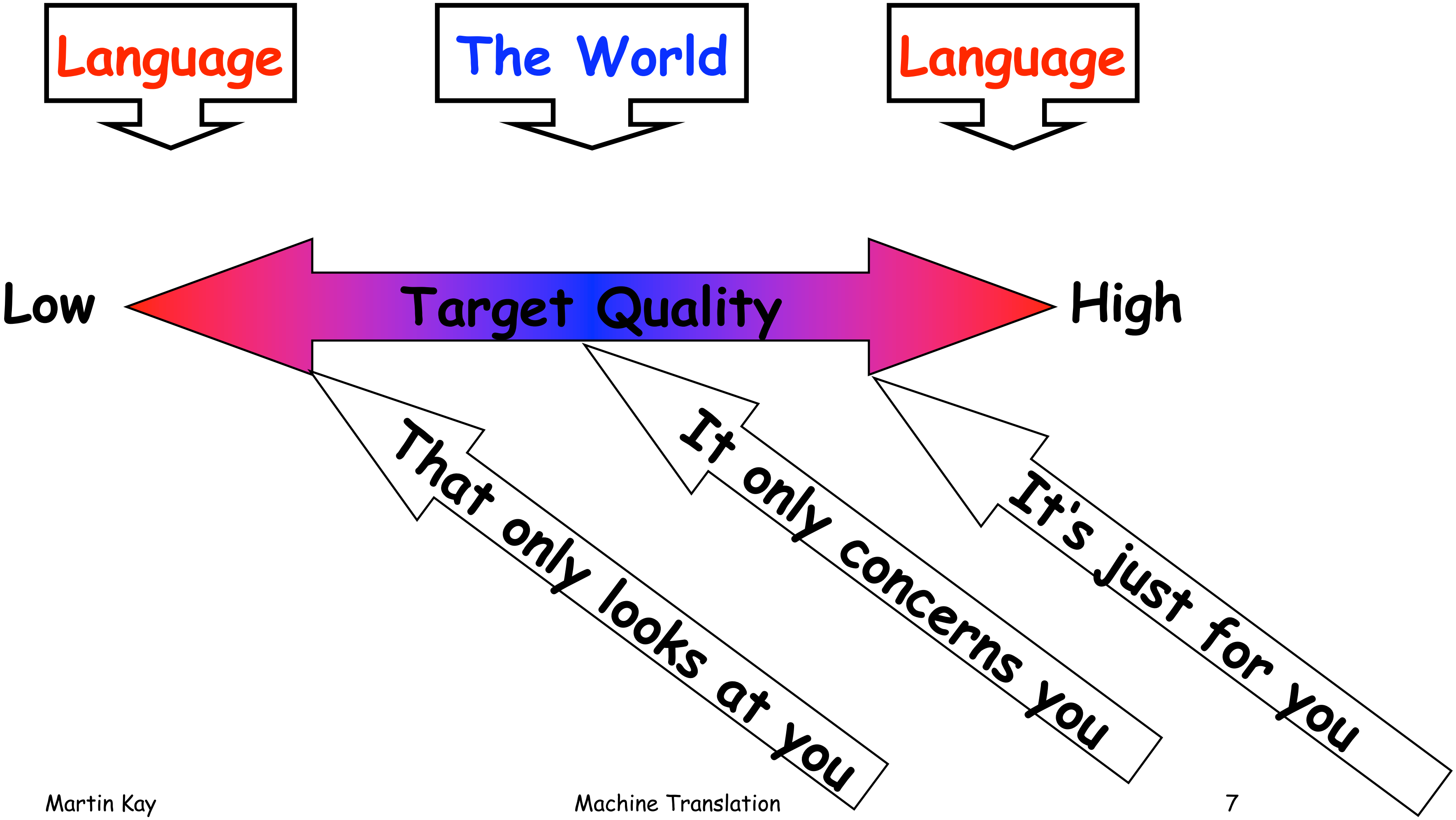
High

Years
0-6

Life
Experience

Translator's
School

Le téléphone sonne? Doucement. Ça ne regarde que vous



Language

The World

Language

Low

Target Quality

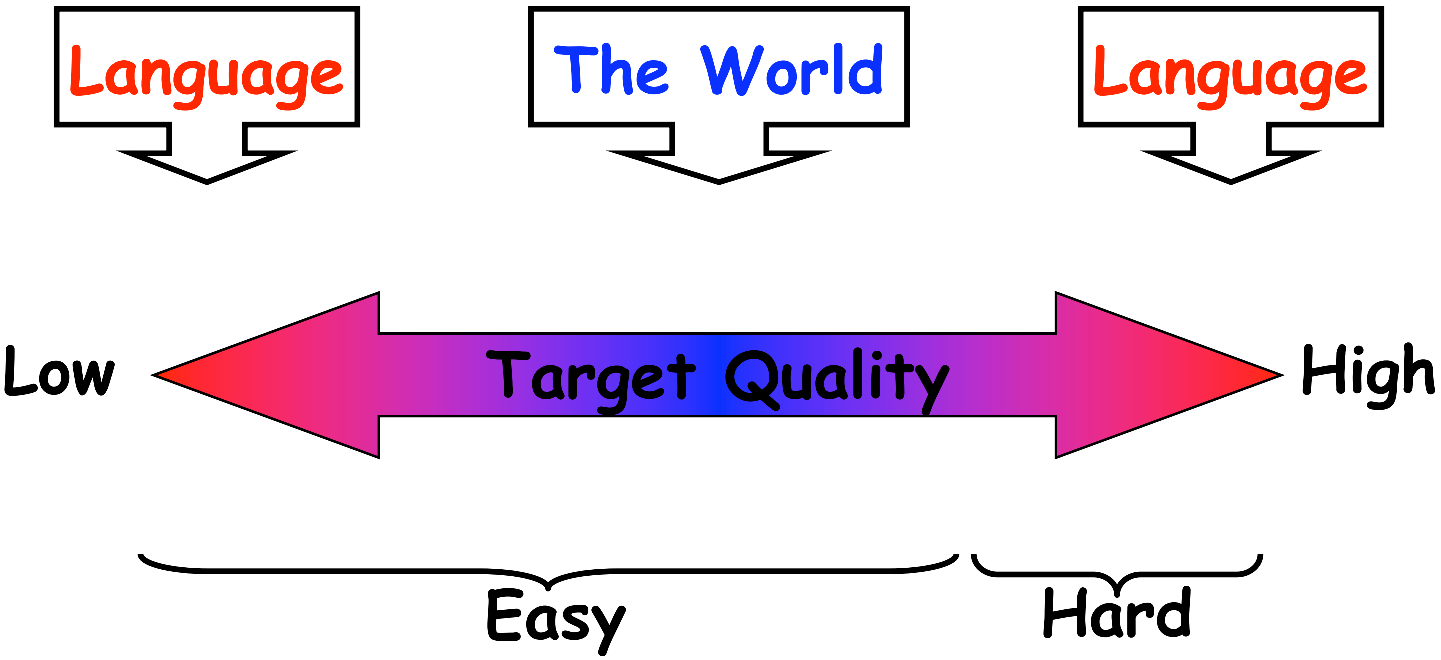
High

That only looks at you

It only concerns you

It's just for you

For People


Language

The World

Language

Low

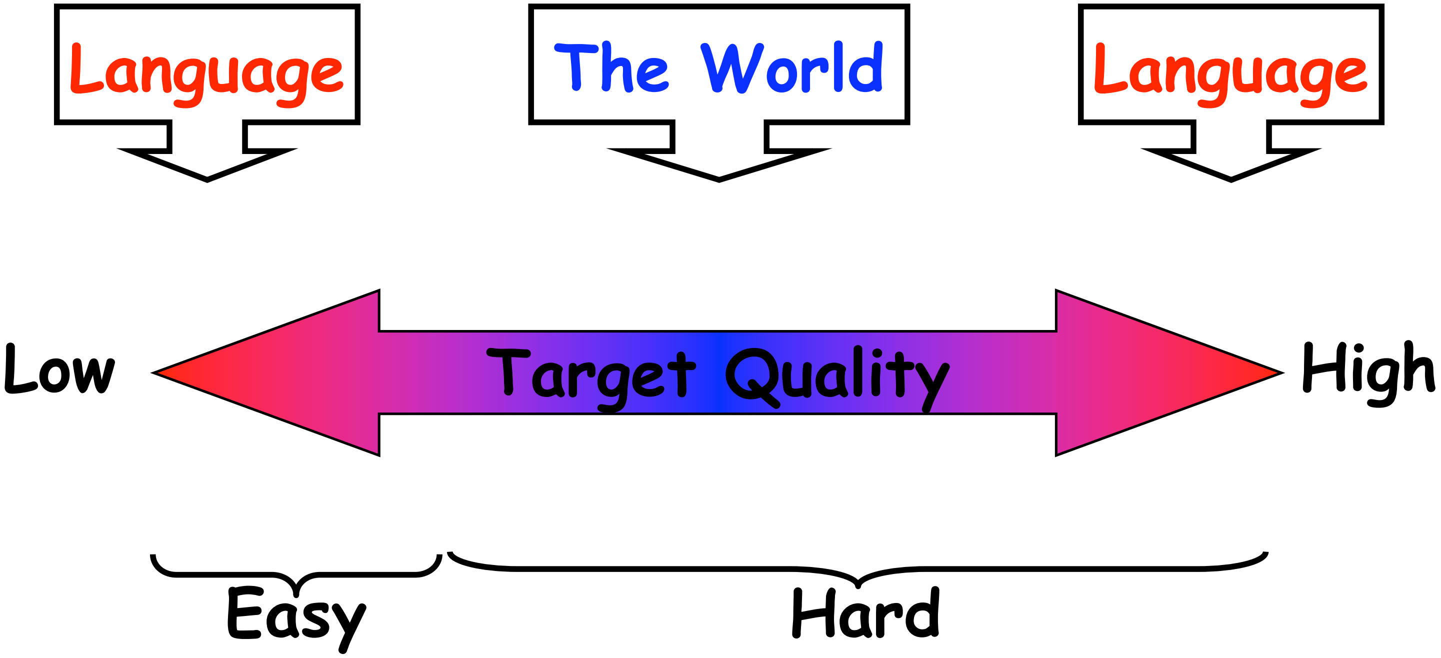
Target Quality

High

Easy

Hard

For Machines



Linguistic Rules are *Rules*

Il a cessé immédiatement de parler

Il espère vivement parler

Because the door was locked, he could not go into the office

Da dir Tür abgeschloßen war, konnte er
nicht ins Büro gehen

Because the train drivers were on strike, he could not go to the office

Da die Lokomotivführer streikten, konnte er
nicht ins Büro fahren

Linguistic facts

The sheep that was attacked by the mountain lion apparently does not belong to the current owner of the property

Linguistic facts

The sheep that were attacked by the mountain lion apparently do not belong to the current owner of the property

Linguistic facts

This is an important matter and it is a fact that the paper claims the president concealed from the public.

Linguistic facts

Seville oranges are quite bitter, but they are good for making the kind of jam the British like with their breakfast.

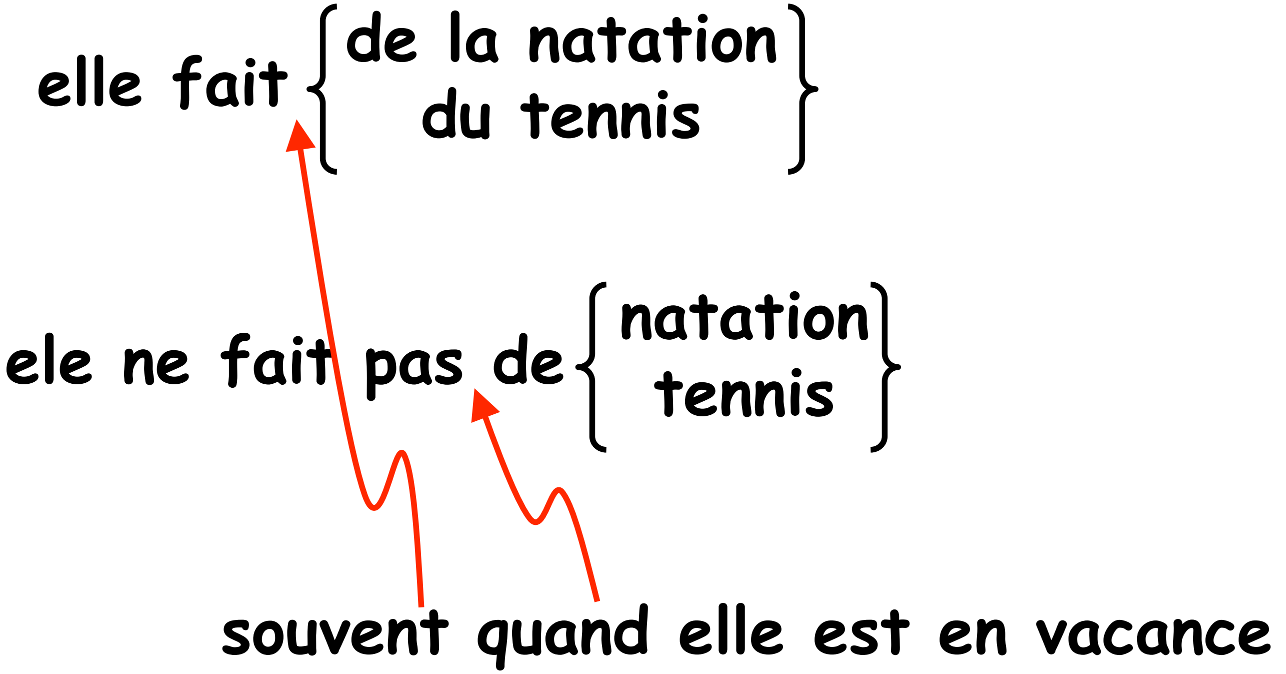
Linguistic facts

Seville oranges are quite bitter, but they are good for making the kind of jam the British like with their breakfast.

Linguistic Facts

The representative requested that the hearing be continued until February 6, 2003. Paul Willard made a motion to continue the hearing until February 6th at 8:00 pm. Don Ritchie seconded the motion. The vote was unanimous.

Linguistic Facts



elle fait {de la natation / du tennis}

ele ne fait pas de {natation / tennis}

souvent quand elle est en vacance

Linguistic Facts

I usually go to work on the bus

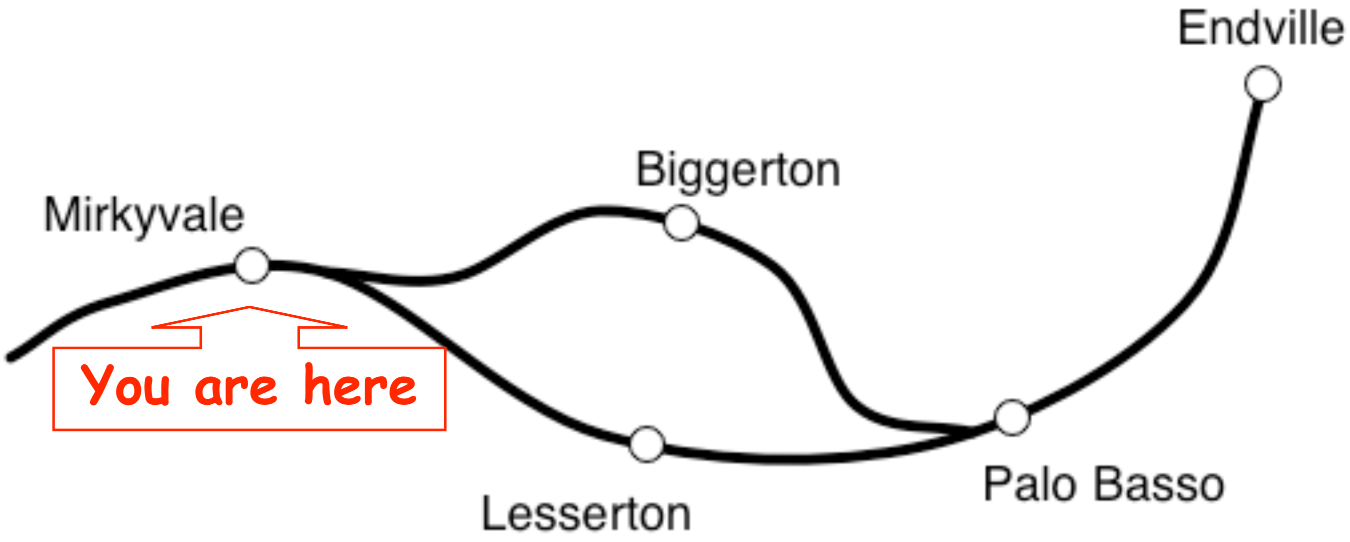
Nonlinguistic Facts

The representative requested that the hearing be continued until February 6, 2003. Paul Willard made a motion to continue the hearing until February 6th at 8:00 pm. Don Ritchie seconded the motion. The vote was unanimous.

Nonlinguistic Facts

Does this train go to Biggerton? No, it stops in Lesserton.

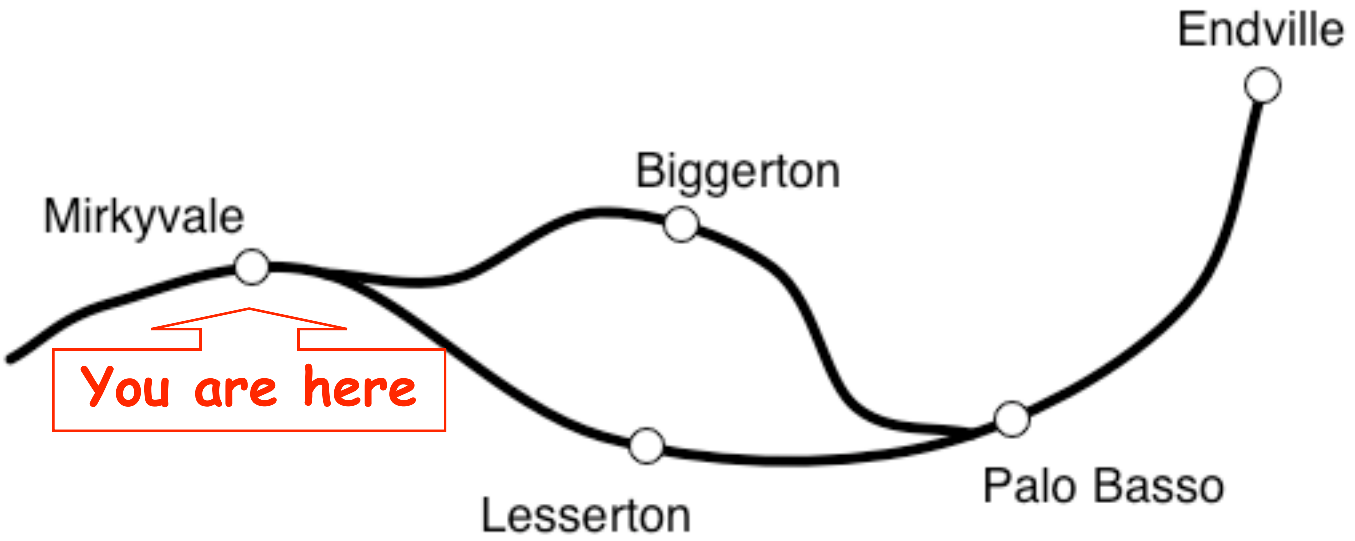
Fahrt dieser Zug nach Biggerton? Nein, er hält in Lesserton.



Nonlinguistic Facts

Does this train *go* to Endville? No, it stops in Palo Basso.

Fahrt dieser Zug nach Endville? Nein, er endet in Palo Basso.



Est-ce que c'est ta cousine?

Non, je n'ai pas de cousine.

female

Is that your cousin?

^

female

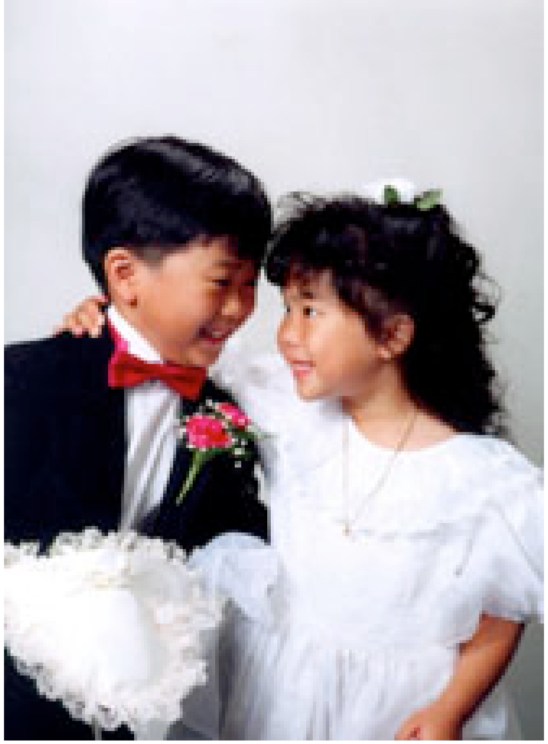
No. I don't have a cousin.

^

Is that woman your cousin?

Est-ce que c'est ta cousine?

Non, je n'ai pas de cousine.



female
Is that your cousin?
^

female
No. I don't have a cousin.
^

girl
Is that ~~woman~~ your cousin?

Linguistic rules require addition of nonlinguistic Information

$$\text{He sat}\left\{\begin{array}{c}\text{in}\\ \text{on}\end{array}\right\}\text{the chair} \qquad \text{Il}\left\{\begin{array}{c}\text{s'est assis}\\ \text{était assis}\end{array}\right\}\left\{\begin{array}{c}\text{sur la chaise}\\ \text{dans la fauteuil}\end{array}\right\}$$

$$\text{Elle écrivait des lettres} \qquad \text{She}\left\{\begin{array}{c}\text{wrote}\\ \text{was writing}\end{array}\right\}\left\{\begin{array}{c}\text{letters}\\ \text{some letters}\end{array}\right\}$$

The perception

Linguistics has failed technology

Linguistics is not about communication

It focuses on fringe phenomena

It is not robust

It luxuriates in ambiguities but is not interested in resolving them

It never gets beyond the sentence

Generative Linguistics

The *generative* vein in linguistics has run out because most problems have

- —been solved
- —turned out to belong to a wider domain

A new paradigm is required based in acknowledging that language is about communication

Translation is about communication

In many ways Linguistics has been a phenomenal success

Where problems have competing solutions, fringe phenomena can decide the issue

The strength and flexibility of language comes in large measure from its openness to ambiguity. Resolving ambiguity is not a linguistic enterprise

Much of what there is to study about language is within the sentence

Lexical semantics

Statistics to the Rescue!

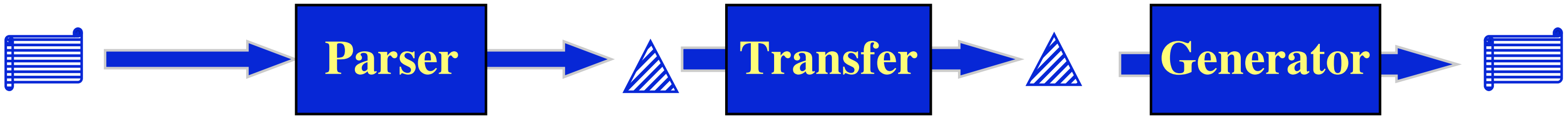
- Rests on primary data
- No linguistic/nonlinguistic distinction
- Treats all phenomena impartially
- Deterministic
- Local
- Rapid development cycle
- People annotate rather than analyze
- Good enough results for government work

Unfortunately we have …

Early binding
Zipf’s law
Locality
Emergent Properties
AI
Bleu score

Machine Translation: The Standard Approach

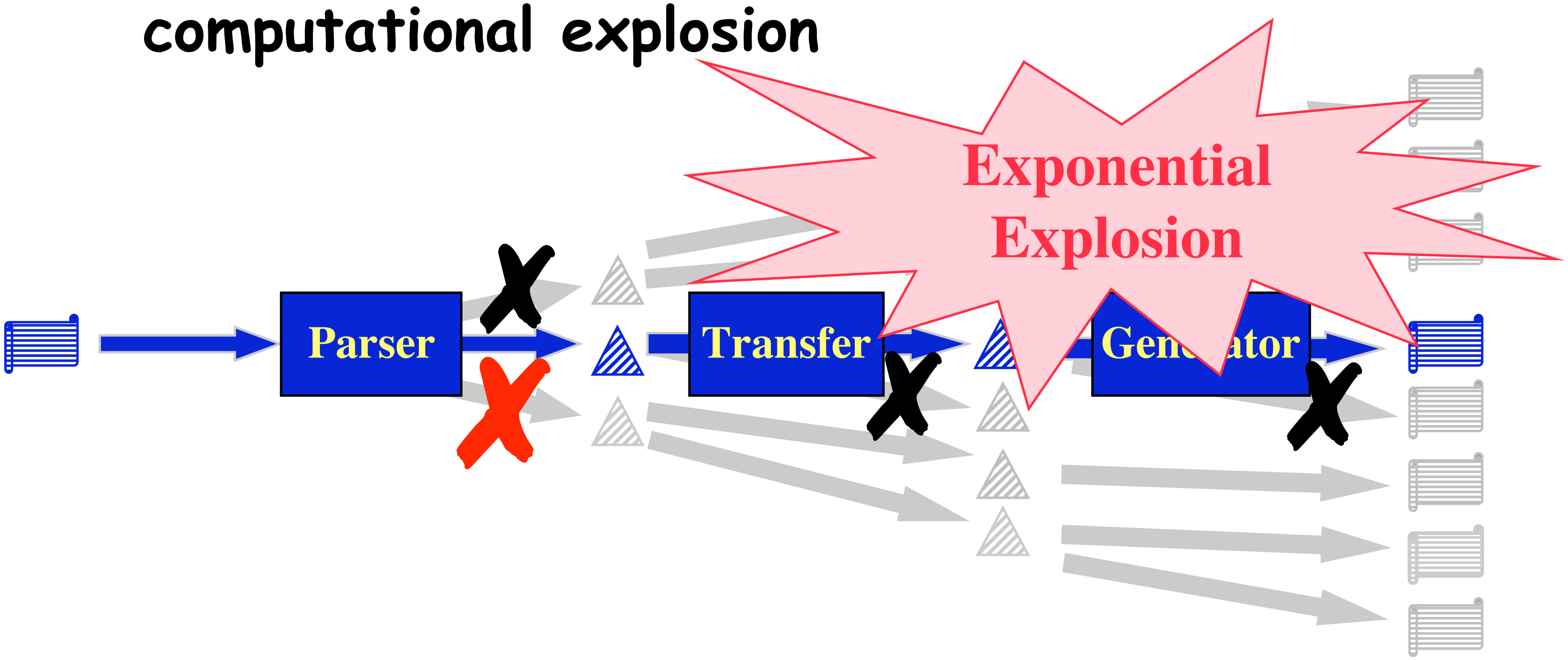
Separate modules for simplicity, maintainability, reuse



Machine Translation: A Common Approach

Separate modules for simplicity, maintainability, reuse

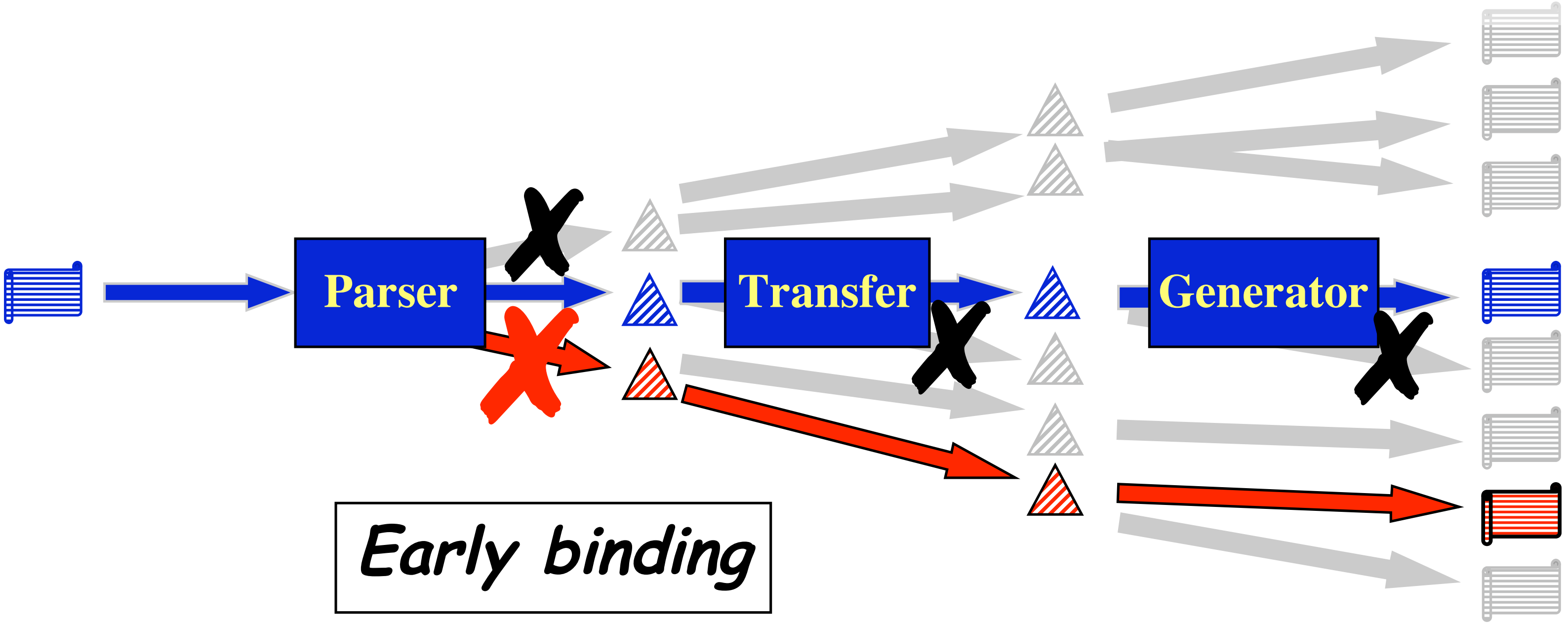
Heuristic filters are applied early to avoid computational explosion



Machine Translation: A Common Approach

Separate modules for simplicity, maintainability, reuse

Heuristic filters are applied early to avoid computational explosion



Proposal

Use human input but do no more human work than would be necessary in any case

Reflective Editing

Translation is nondeterministic—it implements a choice tree

We can identify an outcome and, in particular, the preferred one, by giving the answers to the questions on the path that leads to it.

If any of those questions arise in a subsequent translation into another language, they should presumably be answered the same way.

Strategy

Produce many translations

Display one of them—the *best* one.

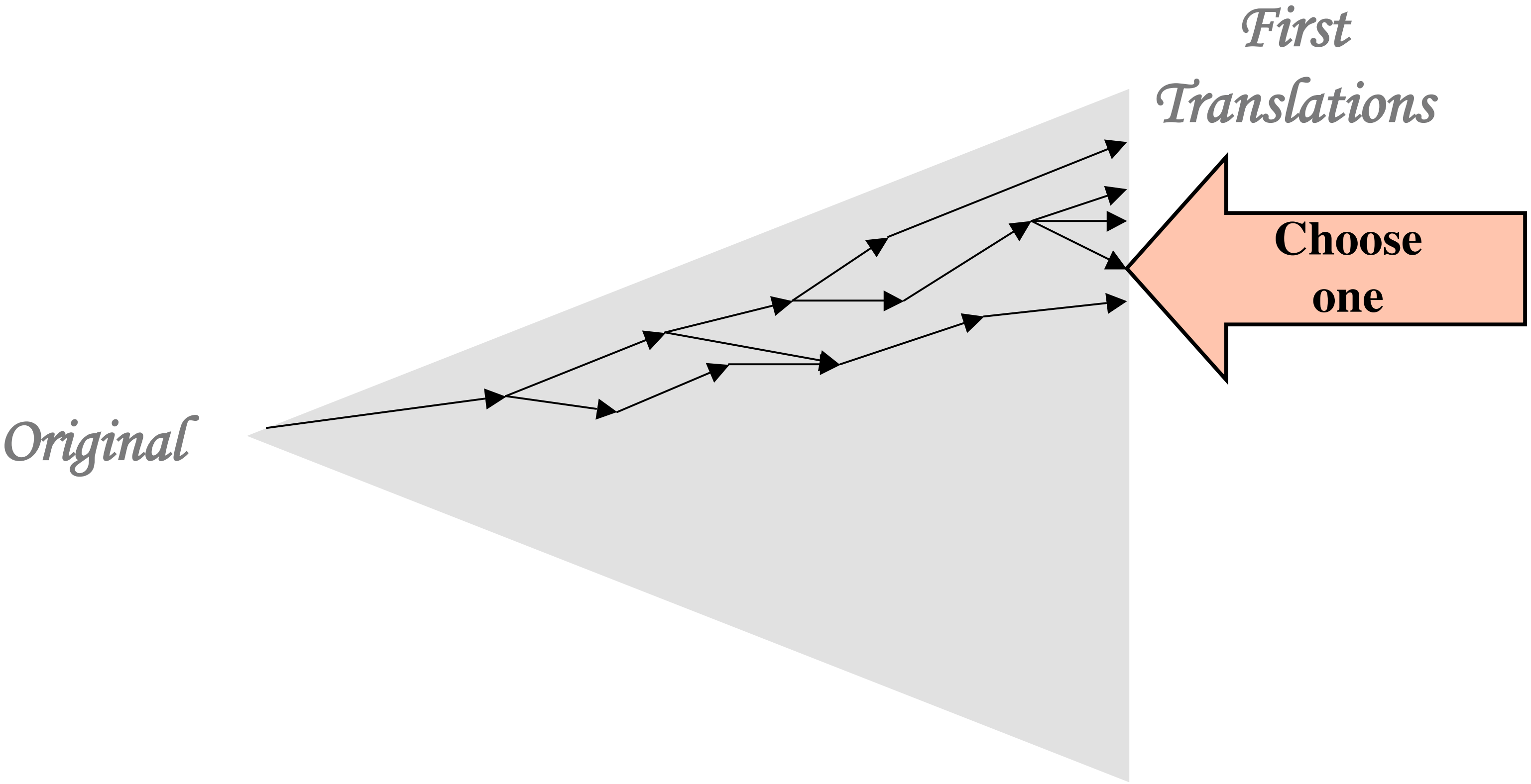
The editor changes it into …

A version that the system had already foreseen, but not chosen as the preferred version.

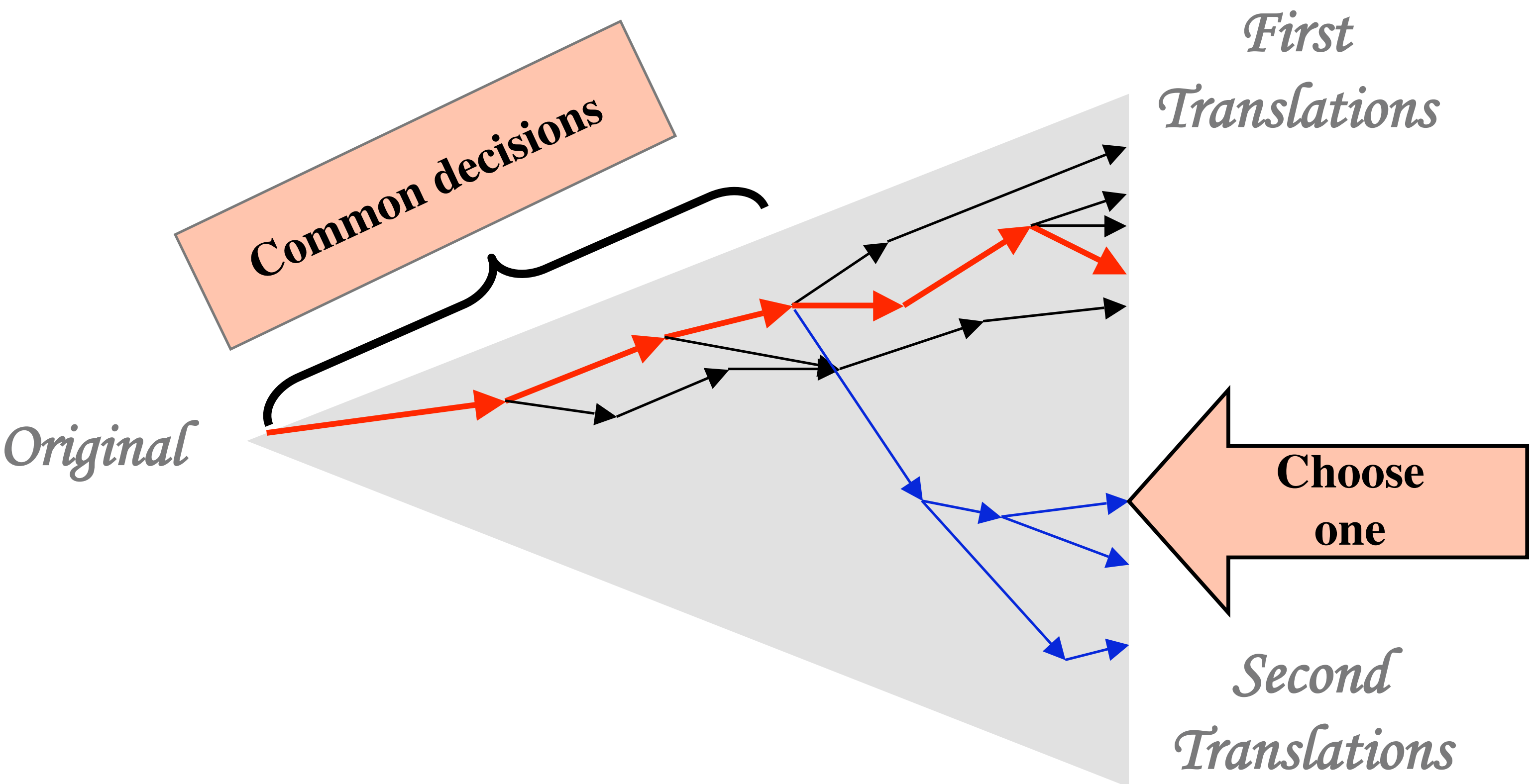
∴ We know what choices the system would have had to make to reach that version.

∴ We will make those choices when translating into the next language.

Reflective Editing



Reflective Editing



There are three windows in the room

Il y a trois fenêtres dans la salle.
Il y a trois guichets dans la salle.

Es gibt drei Fenster in dem Zimmer.
Es gibt drei Schalter in dem Zimmer.

fenêtre ~ Fenster
guichet ~ Schalter

It is cold

Il/elle est froid(e)
Il/elle a froid
Il fait froid

Er/sie/es ist kalt
Ihm/ihr ist kalt
Es ist kalt

[faire] froid/chaud ... ~ Es [sein] kalt/warm ...

X [avoir] froid/chaud ... ~ [Dat] [sein] kalt/warm ...

Wir haben noch zwei

We still have two.
We have two more.

Il nous en reste deux.
Nous en avons encore deux

still ~ [rester]
more ~ encore

The Automatic Linguist

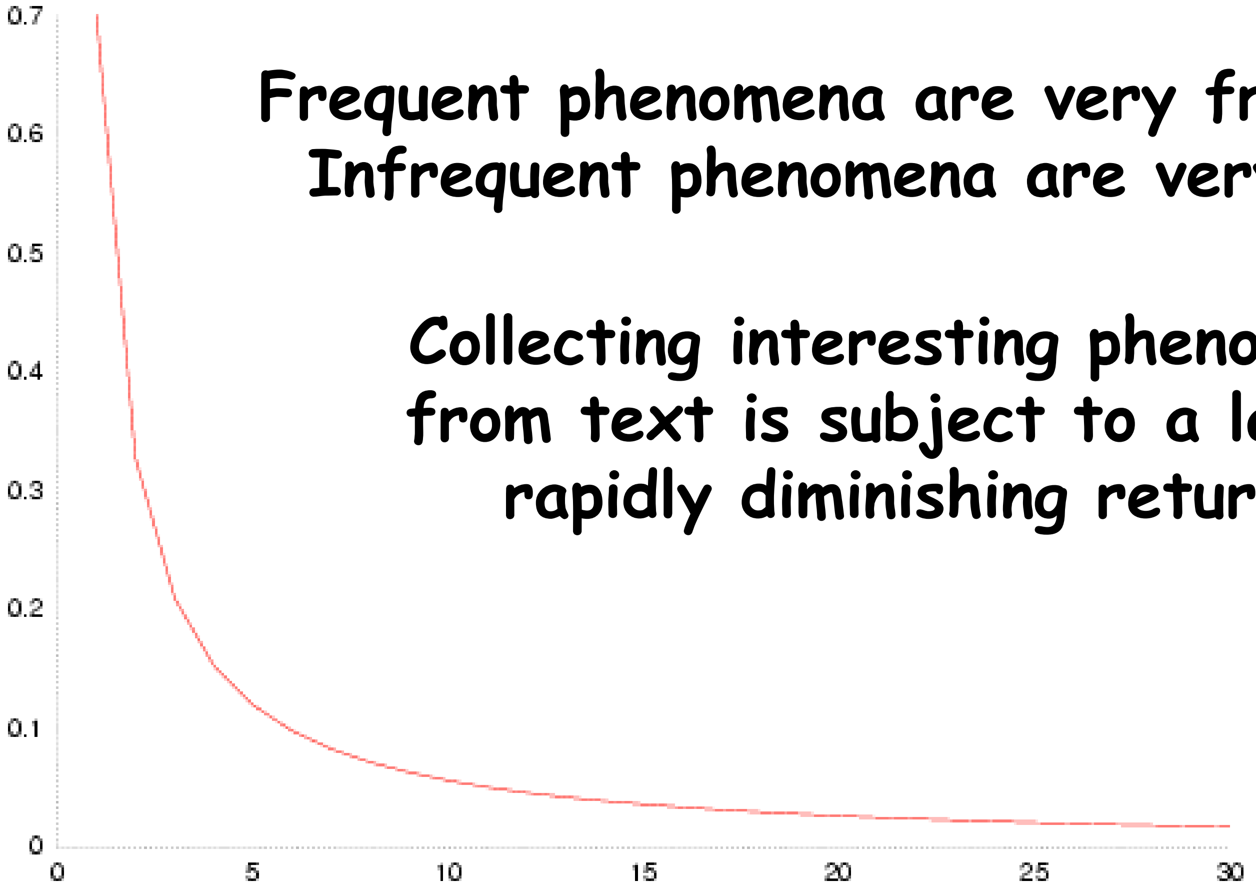
Don't analyze; annotate!

Analyze annotations

… a little

Don't worry about translators;
just look at translations

Zipf's Law



Frequent phenomena are very frequent;
Infrequent phenomena are very rare

Collecting interesting phenomena from text is subject to a law of rapidly diminishing returns

Facts about translation

… are not all reflected in emergent properties of translations

Does this train go to Endville?

Est-ce que c'est ta cousine?

I just got back from Texas/Utah. I had forgotten how good beer tastes.

Ich hatte vergeßen, wie gut[es] Bier schmekt.

It may be necessary to reduce condenser steam side pressure

pression latérale de la vapeur

pression côté vapeur

Robustness and Optimality

Are you a ditransitive verb?

Could you translate French *cousine*?

Could you refer to the same thing as that phrase?

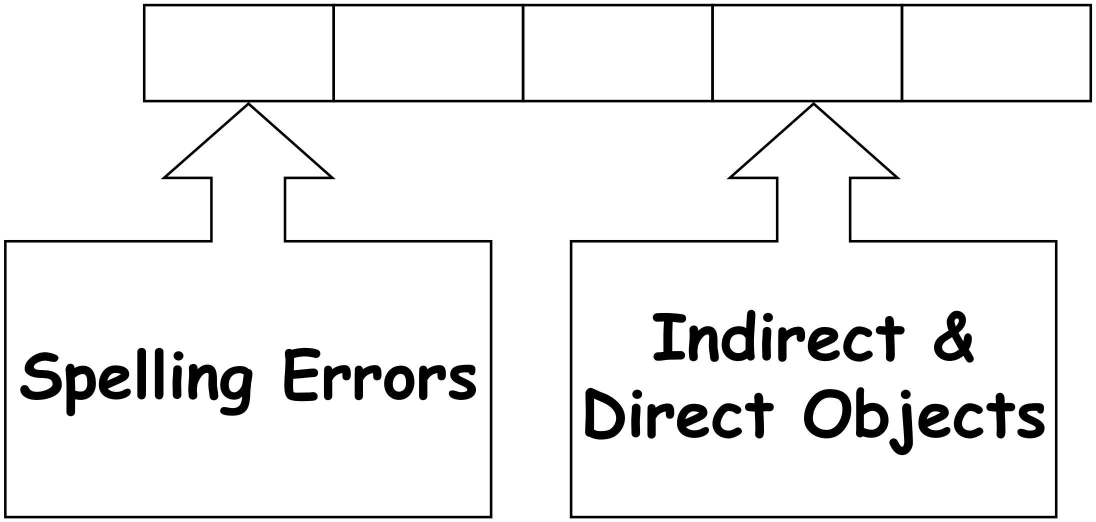
for $29.95

Hans hat dem Kind das Wasser gegeben
?Hans hat das Wasser dem Kind gegeben

Hans hat Maria Wein gegeben
***Hans hat Wein Maria gegeben**

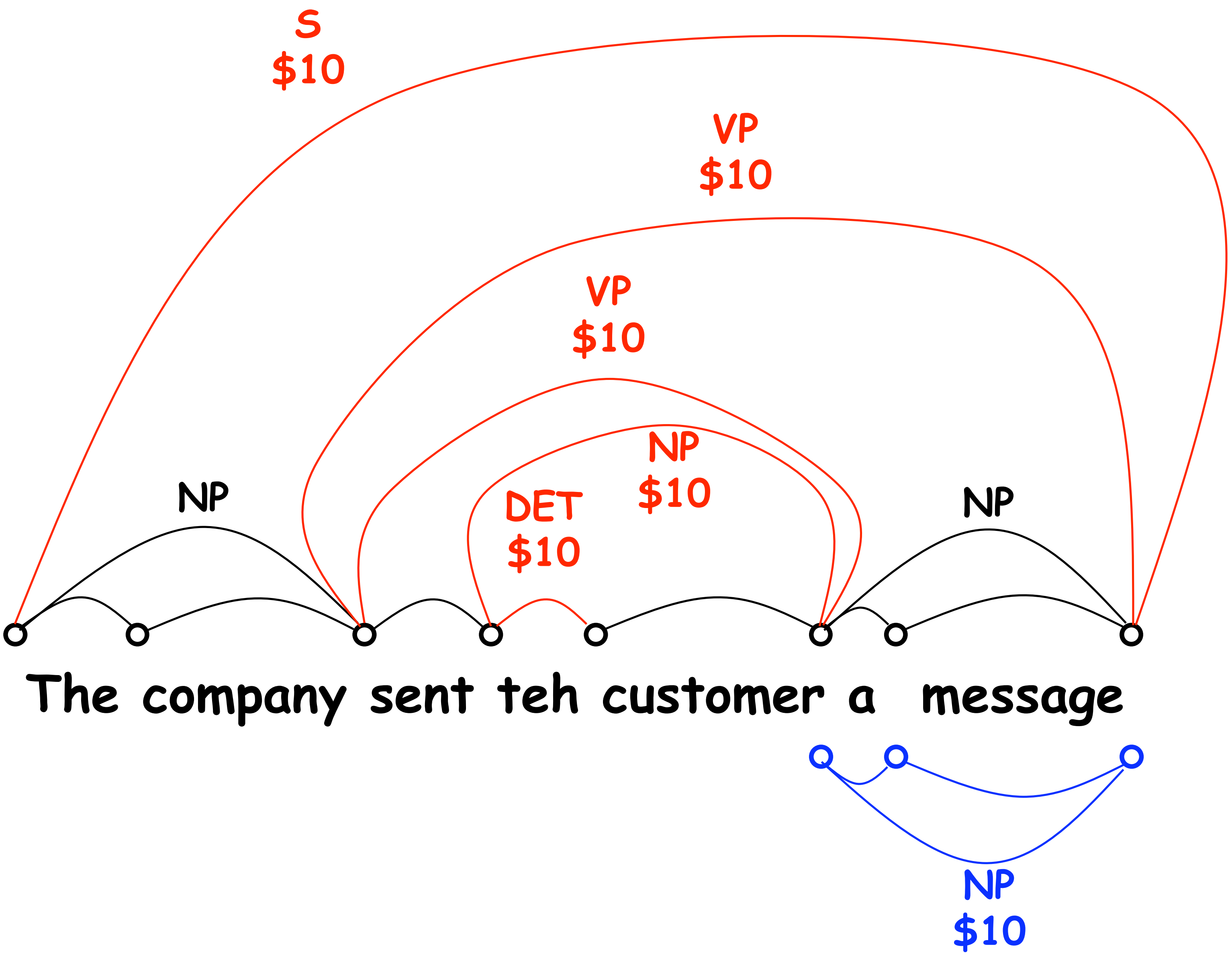
Optimality Charges

Place-value notation
No carries



Spelling Errors

Indirect & Direct Objects



The company sent teh customer a message

$10 ACC < DAT

$0 Hans hat dem Kind das Wasser gegeben

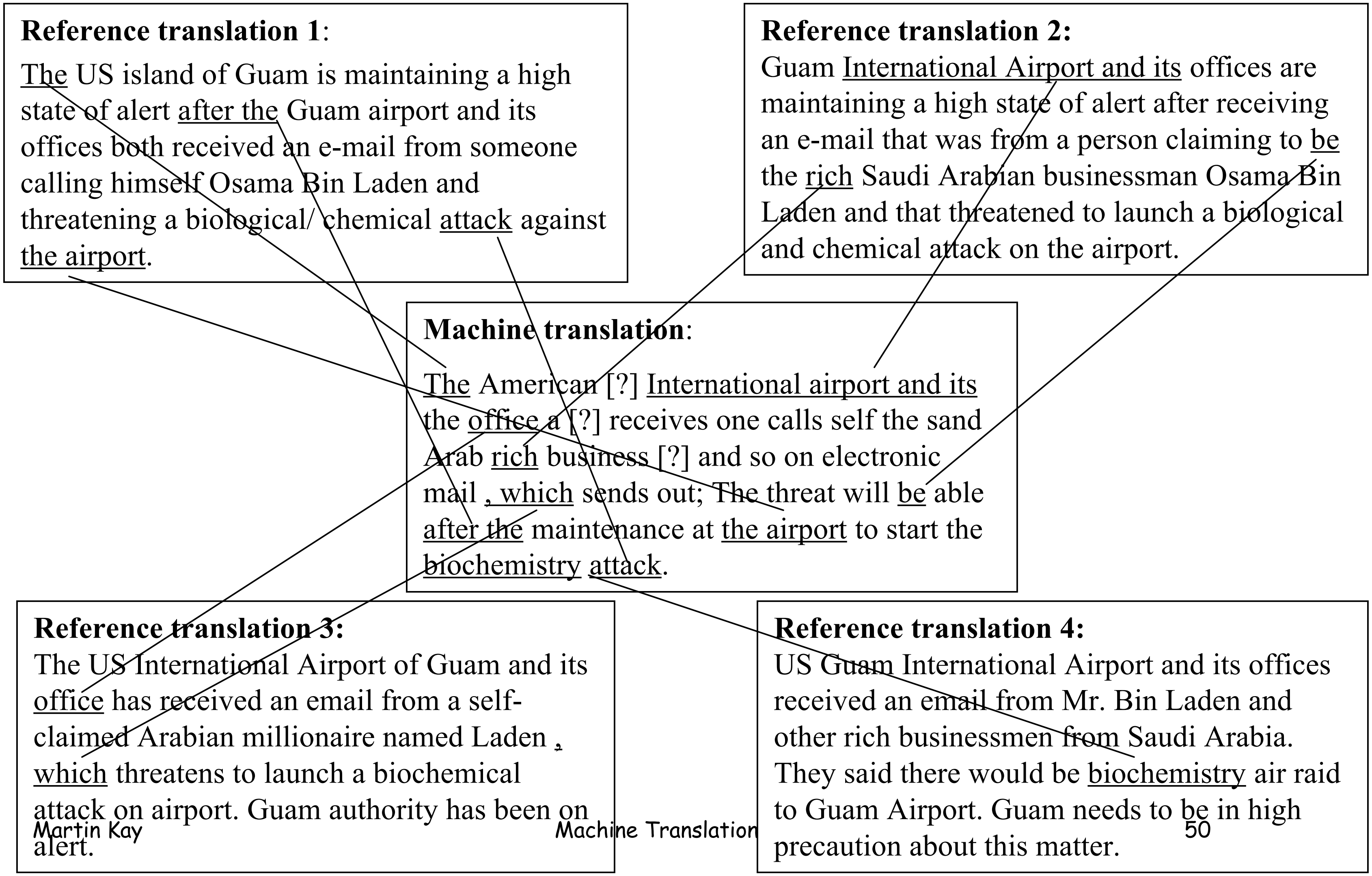
$10 ?Hans hat das Wasser dem Kind gegeben

$0 Hans hat Maria Wein gegeben

$10 Hans hat Maria Wein gegeben

Evaluation

More Reference Translations is Better



Reference translation 1:

The US 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 Osama Bin Laden and threatening a biological/ chemical attack against the airport.

Reference translation 2:

Guam International Airport and its offices are maintaining a high state of alert after receiving an e-mail that was from a person claiming to be the rich Saudi Arabian businessman Osama Bin Laden and that threatened to launch a biological and chemical attack on the airport.

Machine translation:

The American [?] International airport and its the office a [?] receives one calls self the sand Arab rich business [?] and so on electronic mail , which sends out; The threat will be able after the maintenance at the airport to start the biochemistry attack.

Reference translation 3:

The US International Airport of Guam and its office has received an email from a self-claimed Arabian millionaire named Laden , which threatens to launch a biochemical attack on airport. Guam authority has been on alert.

Reference translation 4:

US Guam International Airport and its offices received an email from Mr. Bin Laden and other rich businessmen from Saudi Arabia. They said there would be biochemistry air raid to Guam Airport. Guam needs to be in high precaution about this matter.

Will get worse as translation gets better!

All words that are not in the reference translation are equally bad

Only considers local phenomena—no syntax, anaphora, reference …

Emergent Properties

The important facts about language are not emergent properties of text.

L'arbitraire du signe

The important facts about translation *may* not *all* be emergent properties of translations.

The End

Fin

Ende

The Bleu Score

- Inspired by the Word Error Rate metric used by ASR research
- Measuring the “closeness” between the MT hypothesis and human reference translations
- Precision: n-gram precision
- Recall:
  - Against the best matched reference
  - Approximated by brevity penalty
- Cheap, fast
- Highly correlated with subjective evaluations
- MT research has greatly benefited from automatic evaluations
- Typical metrics: IBM BLEU, NIST MTeval, NYU GTM

- **Adding information**
- **Elle ne fait pas de voile**
- **Quel**

Why we need some kind of linguistics

- Syntax ~ Locality
- When different rules apply, different messages are conveyed
- We want grammatical results

Approaches

- Deep linguistics
- Shallow Linguistics
- Statistics
- Hybrid Schemes

Computational Linguistics

- Everything must contribute to a complete analysis
- Nondeterministic
- Only linguistic factors contribute

Undestanding Language

It is important to understand how language works

- —for its technological value
- —because it is quintessentially human

Understanding is presently thwarted by

- —Linguistics that denies meaning
- —Statistics that denies understanding

Understanding translation

- —encompasses understanding language
- —cannot be in denial

Statistics

Statistics is no substitute for thought, and science, and observation, and understanding because

- Linguistics works much better
- Language is not fundamentally statistical
- Zipf's law implies diminishing returns
- Important facts about language are not emergent properties of texts

Paradigms

- **Linguistics**
  - Translation is quintessentially linguistic. Language is based on special mental capacities that may be innate. Language cannot be understood in terms of emergent properties of texts.
- **Artificial Intelligence**
  - Translation consists in creating a text that reexpresses what has been understood from another text. A good translator must be a renaissance man.
- **Statistics**
  - Translation is best understood by examining examples.

So what went wrong?

- There are no practical tasks that are entirely, or even primarily linguistic
  - Summarization
  - Information extraction
  - Translation
- Real tasks that seem to be linguistic almost always require a complete artificial intelligence

What to do?

- Simplify the task
  - —Limit the domain
  - —Reduce expectations
  - —Use ignorance modeling

Probabilistic linguistic phenomena

- Word choice
- Semantic vs. syntactic agreement (Mädchen)
- Commencer à/de

Who needs translation?

- Producers
  - Control the domain the domain
  - Require high quality
- Consumers
  - Little control of domain
  - Varying quality requirements

Triangulation

Optimality

Linguistics

- This is one of the manuscripts that he is supposed to have donated the collection.
- I think my car is *going* to need a new motor.
- The British like to have orange jam for breakfast
- When people read the fine print, they decid(ed) not to *go* through with it.

Statistics as a Stand-in for AI

Syntactic Ambiguity

- airport long term car park courtesy vehicle pickup point
- I bought a car with four doors/dollars
- Attach the end of the wire from the power supply of the unit to the red terminal on the panel at the back of the amplifier (1430 structures)
- Connect pressure and return lines to pump
- I just got back from Texas/Utah//Germany/Saudi Arabia. I had forgotten how good beer tastes.
  - —Ich hatte vergeßen, wie gut[es] Bier schmekt.
- His paper shows that smoking can cause cancer

Understanding Language

It is important to understand how language works

- for its technological value
- because it is quintessentially human

Understanding is presently thwarted by

- Linguistics that denies meaning
- Statistics that denies understanding

Understanding translation

- encompasses understanding language
- cannot be in denial

Generative Linguistics

The generative vein in linguistics has run out because most problems have

- been solved
- turned out to belong to as wider domain

A new paradigm is required based in acknowledging that language is about communication

Translation <u>is</u> about communication

Statistics

Statistics is no substitute for thought, and science, and observation, and understanding because

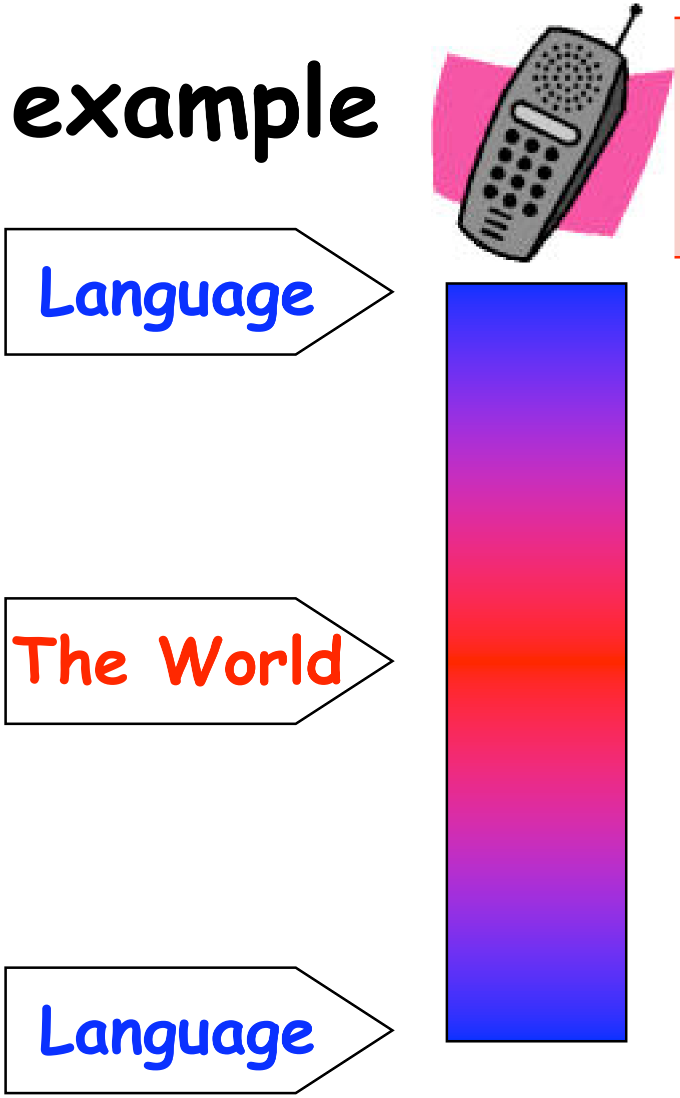
- Linguistics works much better
- Language is not fundamentally statistical
- Zipf's law implies diminishing returns
- Important facts about language are not emergent properties of texts

Translation Problems

Hard

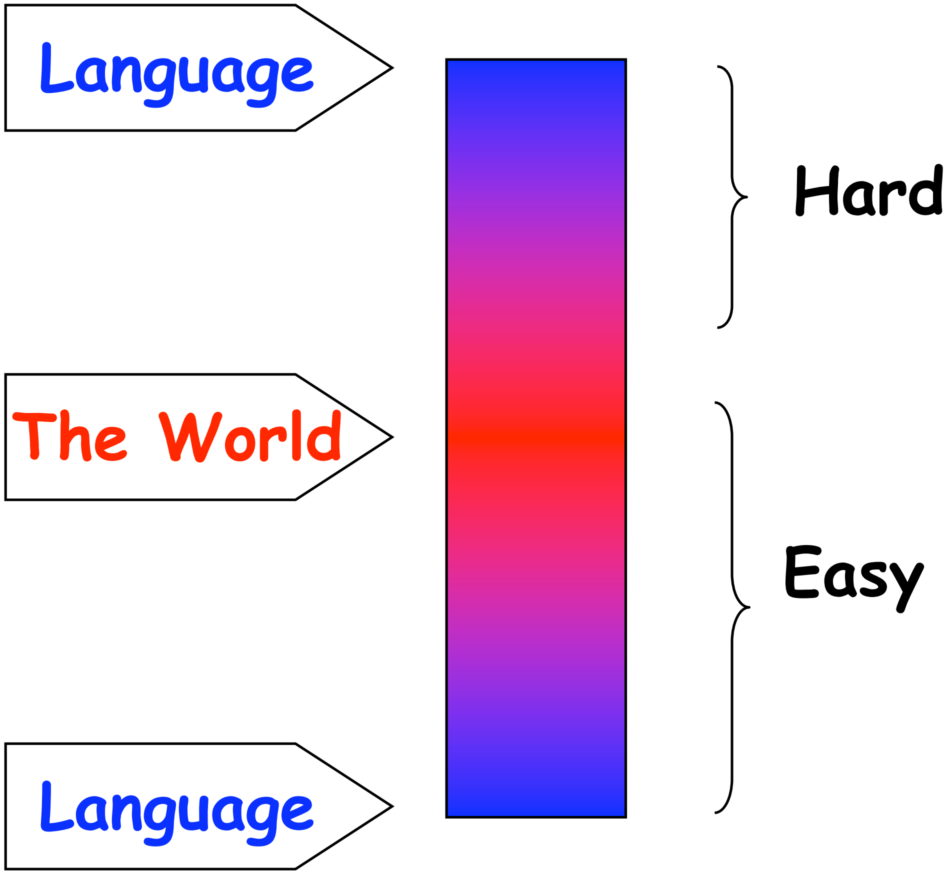
Easy

For example

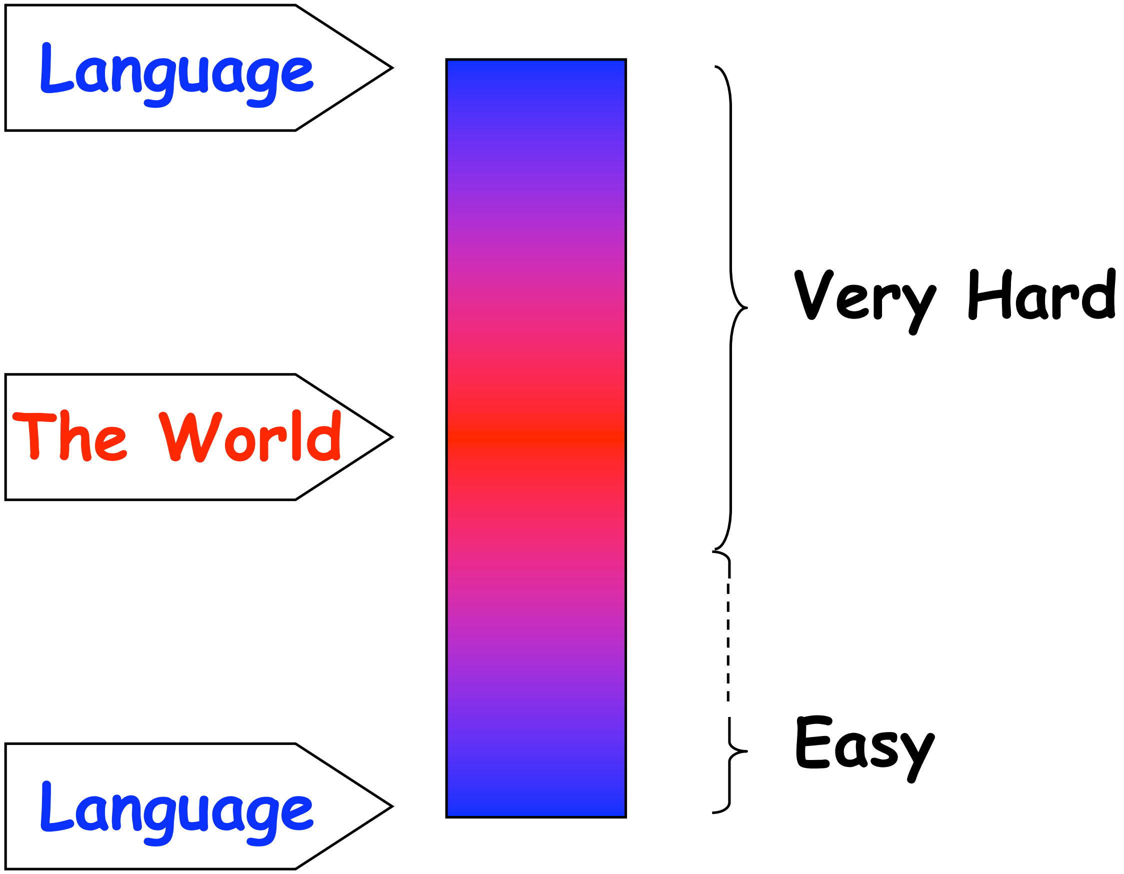


Le téléphone sonne? Doucement. Ça ne regarde que vous

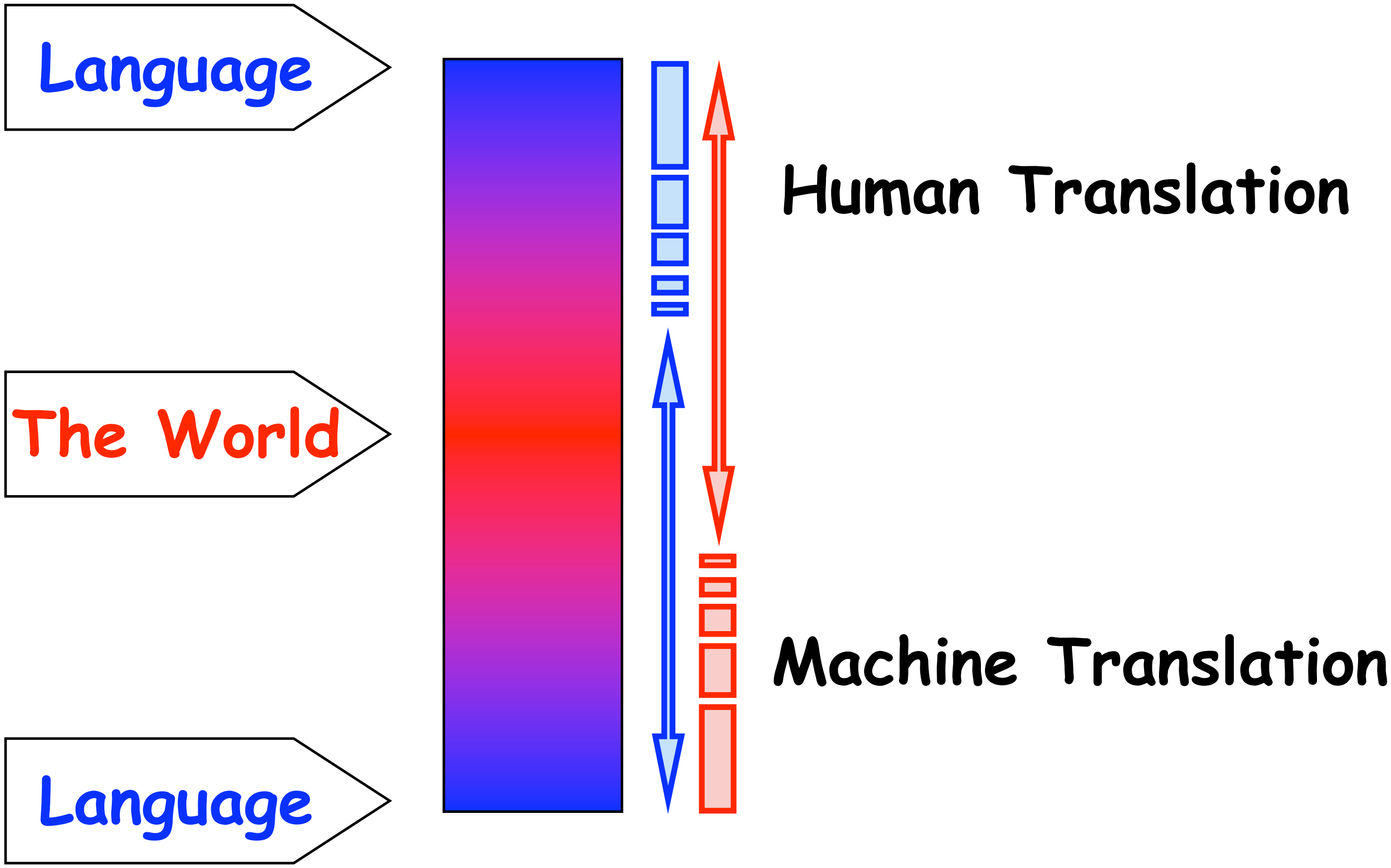
For people ...



For theories and machines ...



Humans and Machines



Language

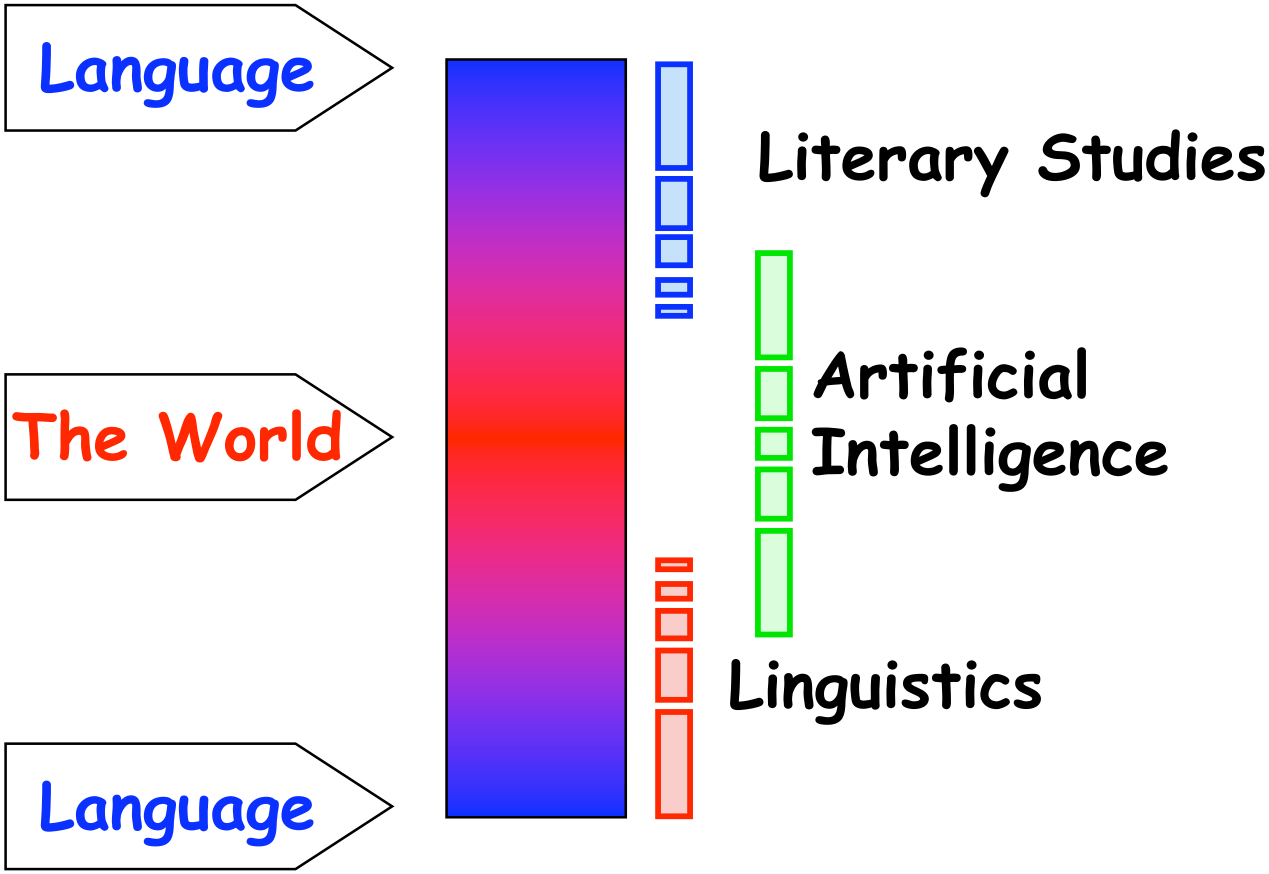
The World

Language

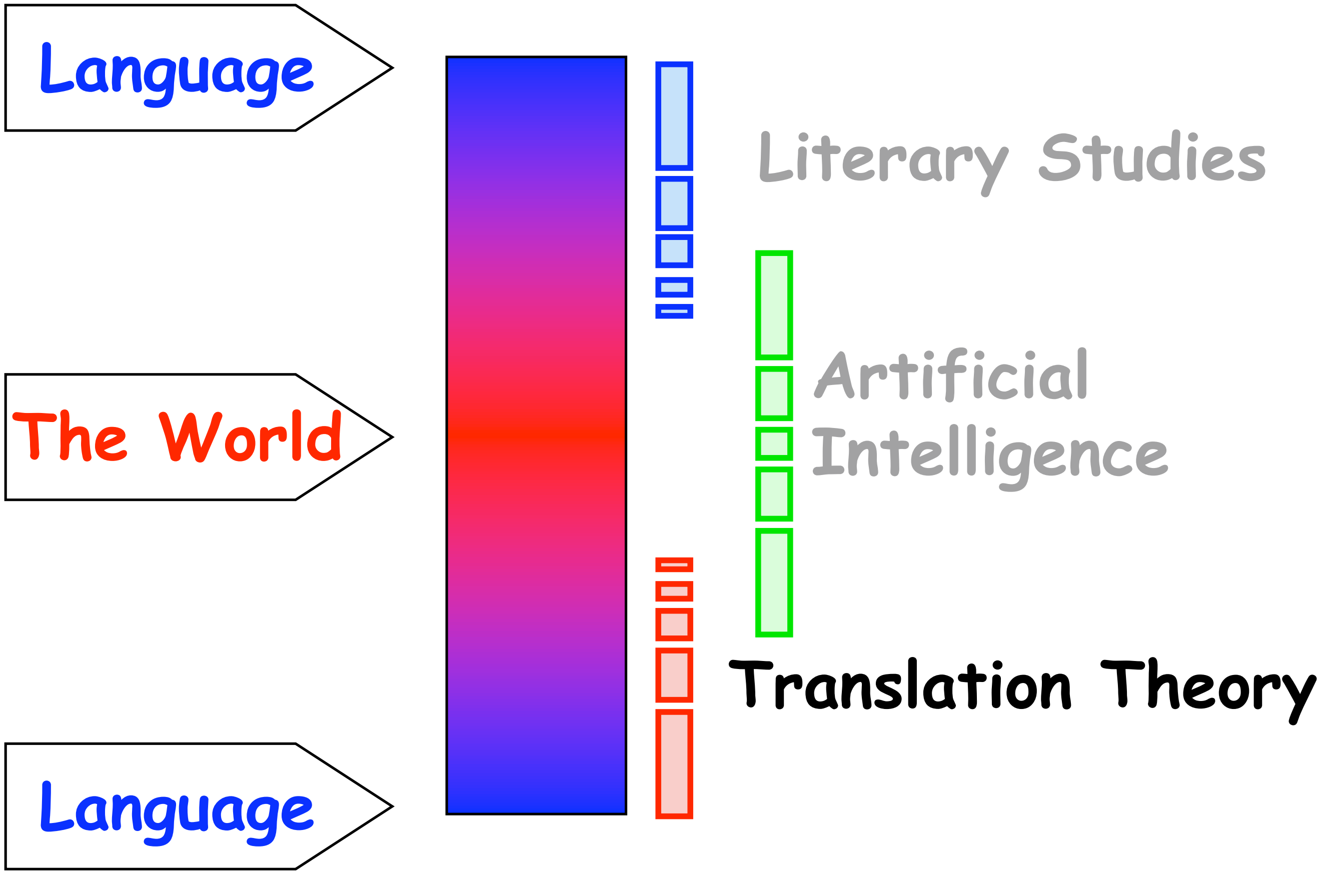
Human Translation

Machine Translation

The Disciplines



The Disciplines



Language

The World

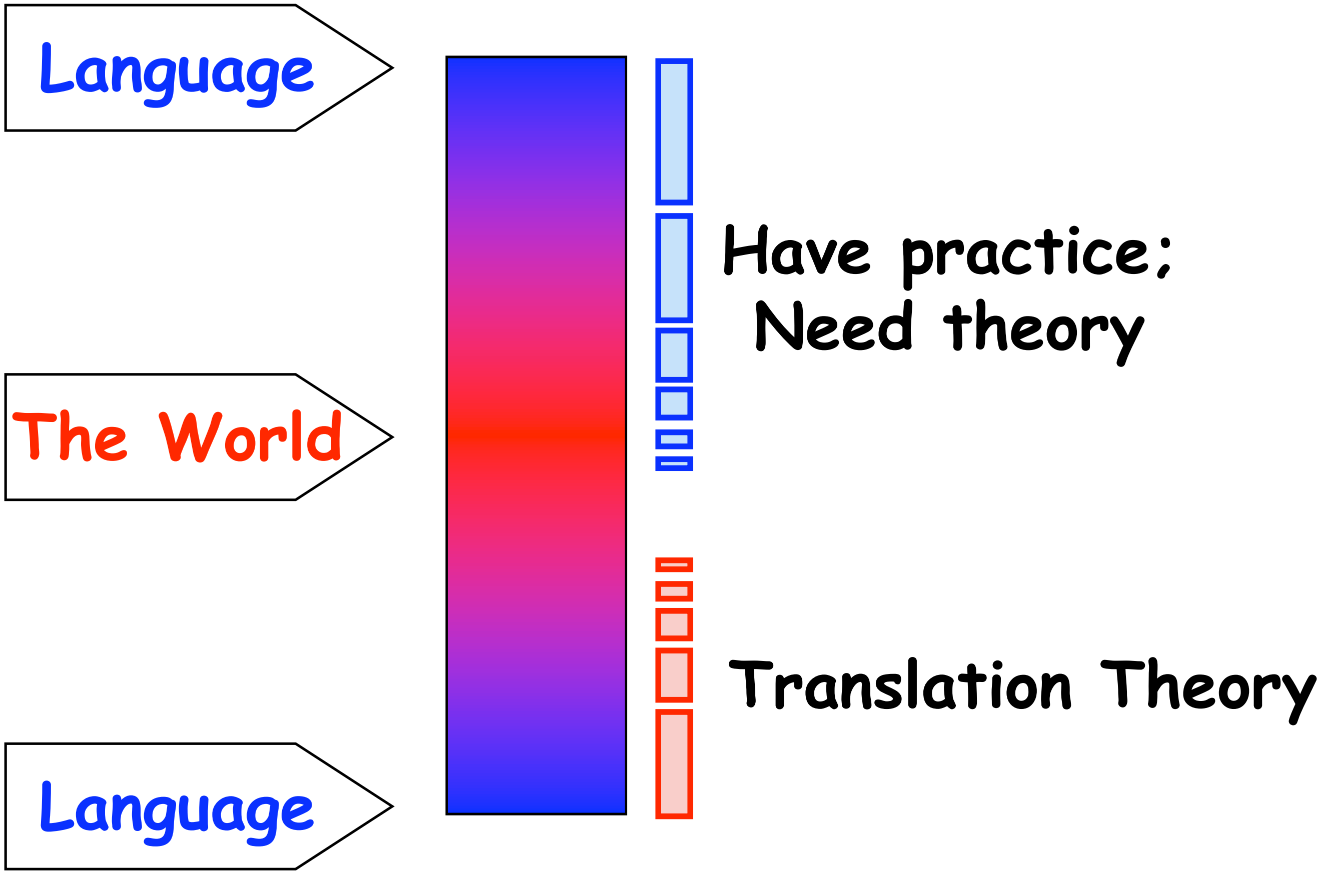
Language

Literary Studies

Artificial Intelligence

Translation Theory

The Disciplines


Language
The World
Language
Have practice;
Need theory
Translation Theory