# Machine Translation

## A New Frontier?

**Martin Kay**

**Stanford University and**

**The University of the Saarland**

# The European Union

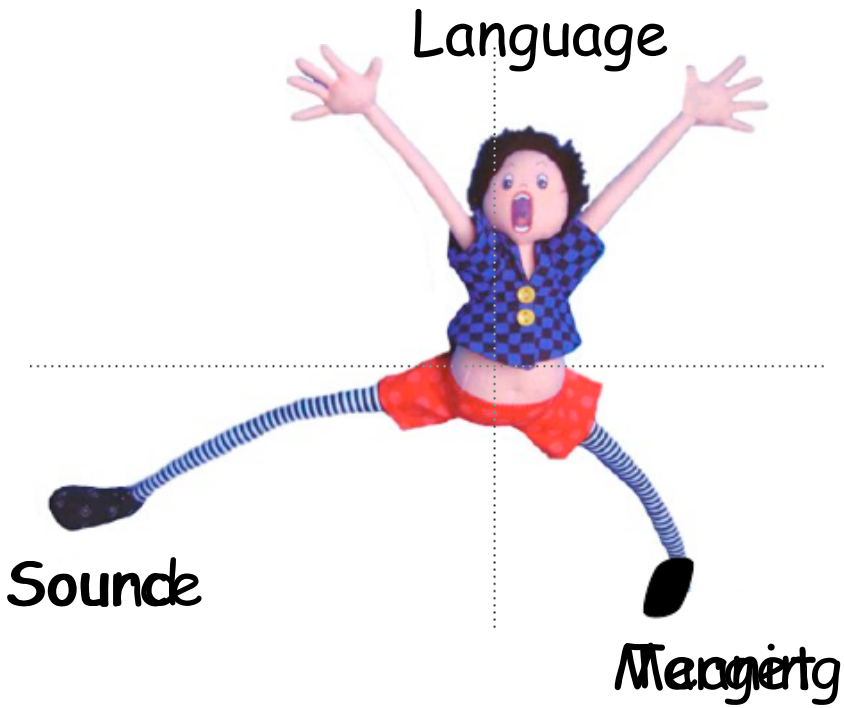
Danish
Dutch
English
Finnish
French
German
Greek
Italian
Portuguese
Spanish
Swedish

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

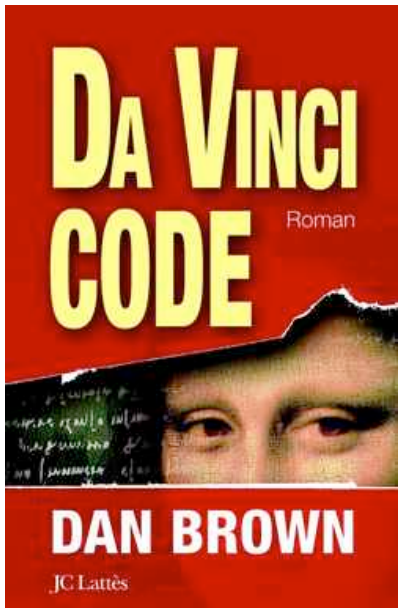
$$\binom{23}{2} = 253$$

Bulgarian
Czech
Estonian
Hungarian
Irish
Latvian
Lithuanian
Maltese
Polish
Romanian
Slovene
Slovak

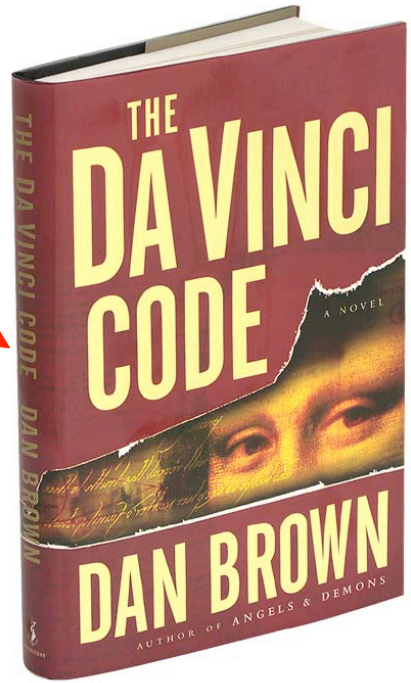
| CATERPILLAR | Maintenance Manuals<br>Operation and Troubleshooting Guides<br>Disassembly and Specifications Manuals<br>Assembly Manuals<br>Testing and Special Instructions<br>Adjustment Guides<br>Systems Operation Bulletins |
| --- | --- |
| 300 authors and illustrators<br>800 English pages per day<br>Translation into 14 languages | |

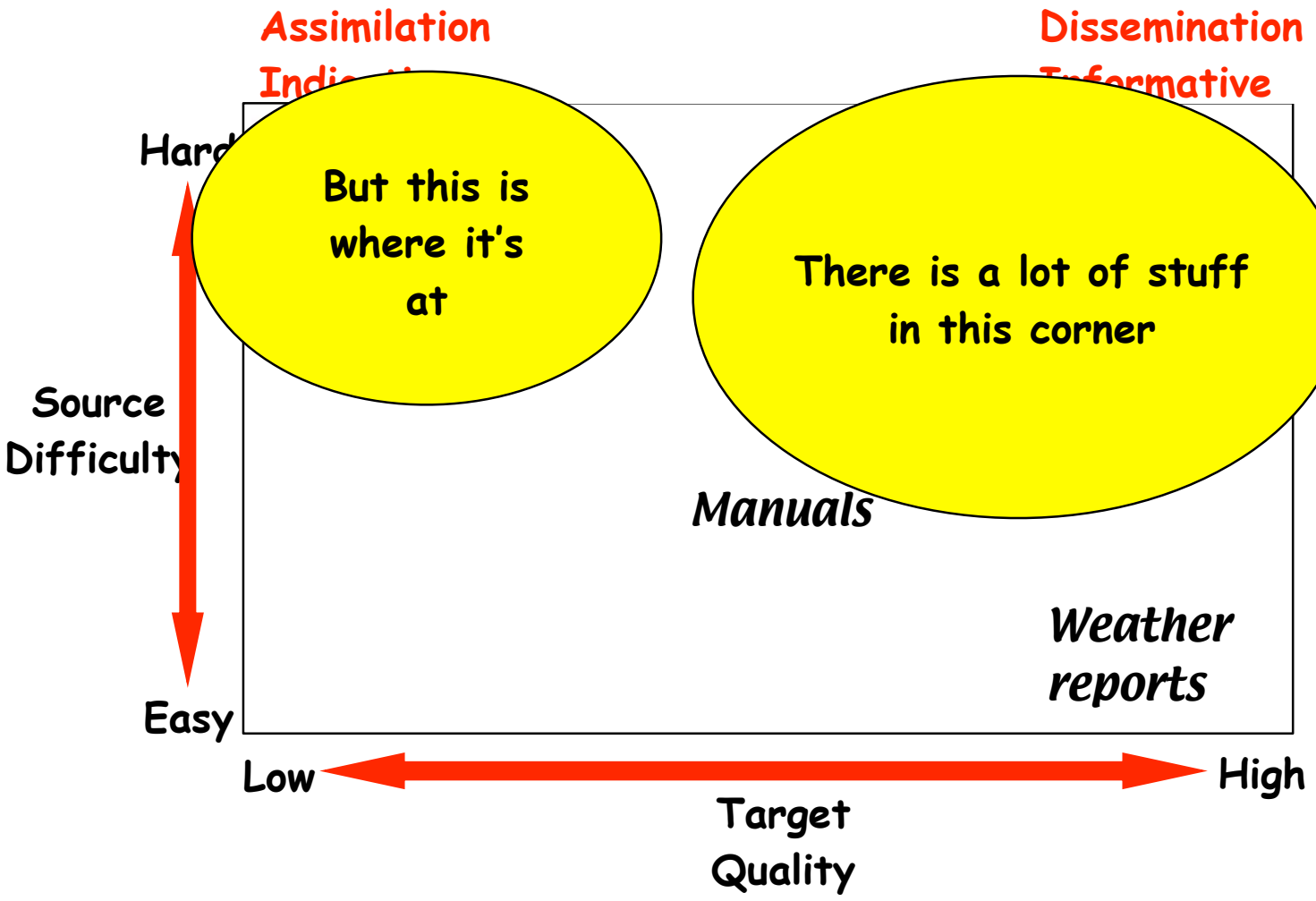

Language
Sounde
Meaning

When is this a translation of this?




When they have the same meaning ... ?




Assimilation
Dissemination
Hard
But this is where it's at
There is a lot of stuff in this corner
Source Difficulty
Manuals
Weather reports
Easy
Low
High
Target Quality

# What is Translation?

- A text that is based on a text in another language and which
  - has the same meaning
  - conveys the same information
  - has the same effect on its readers
  - gives the gist of the original
  - explains the original
- It depends what you want
- Generally a mixture of several of these

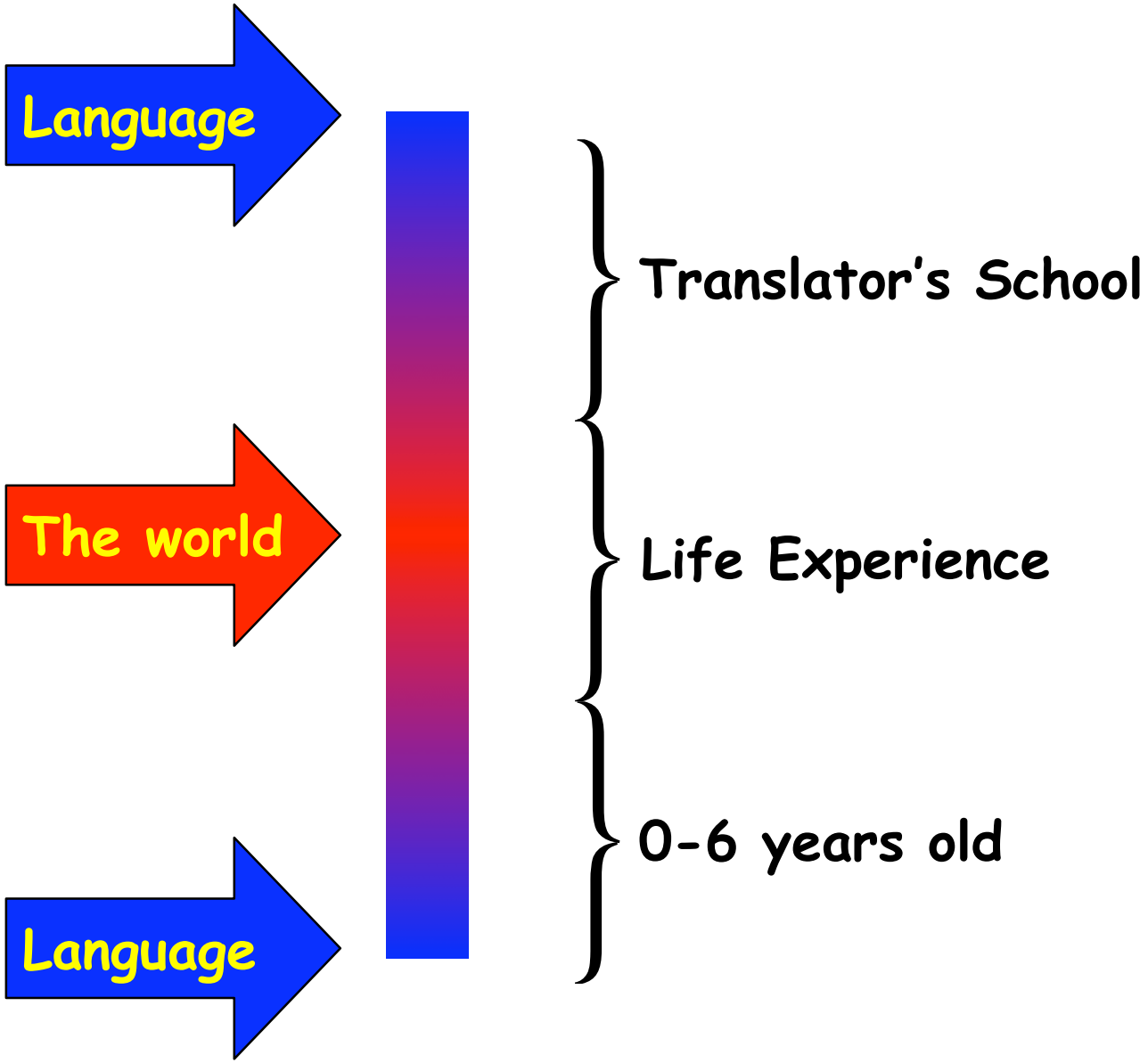
**A man and his two sons are on one side of a river and want to cross to the other side. There is a boat that can carry no more than 80 kilos. The father weighs 80 kilos and the sons 40 kilos each. How do they all get to the other side?**

# Broadly Speaking

**Noun phrases are used either to introduce new objects or to refer to previously introduced objects.**

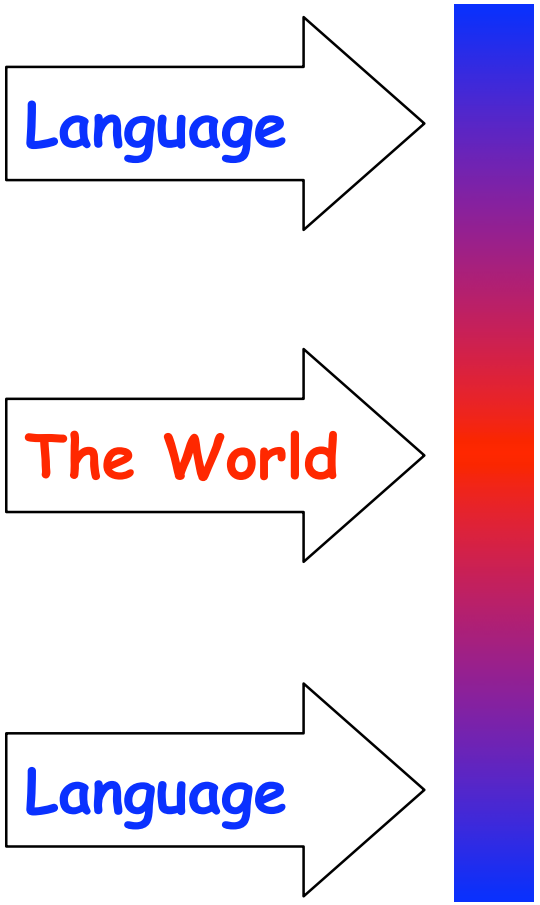
> Adam, Brian and Charles want to cross a river. Adam is the father of Brian and Charles and he weighs about the same as the two boys do together.

**Referring phrases need only be specific enough to distinguish among objects that have already been introduced.**


Language
The world
Language
Translator's School
Life Experience
0-6 years old

# For example

Où voulez-vous que je me mette?



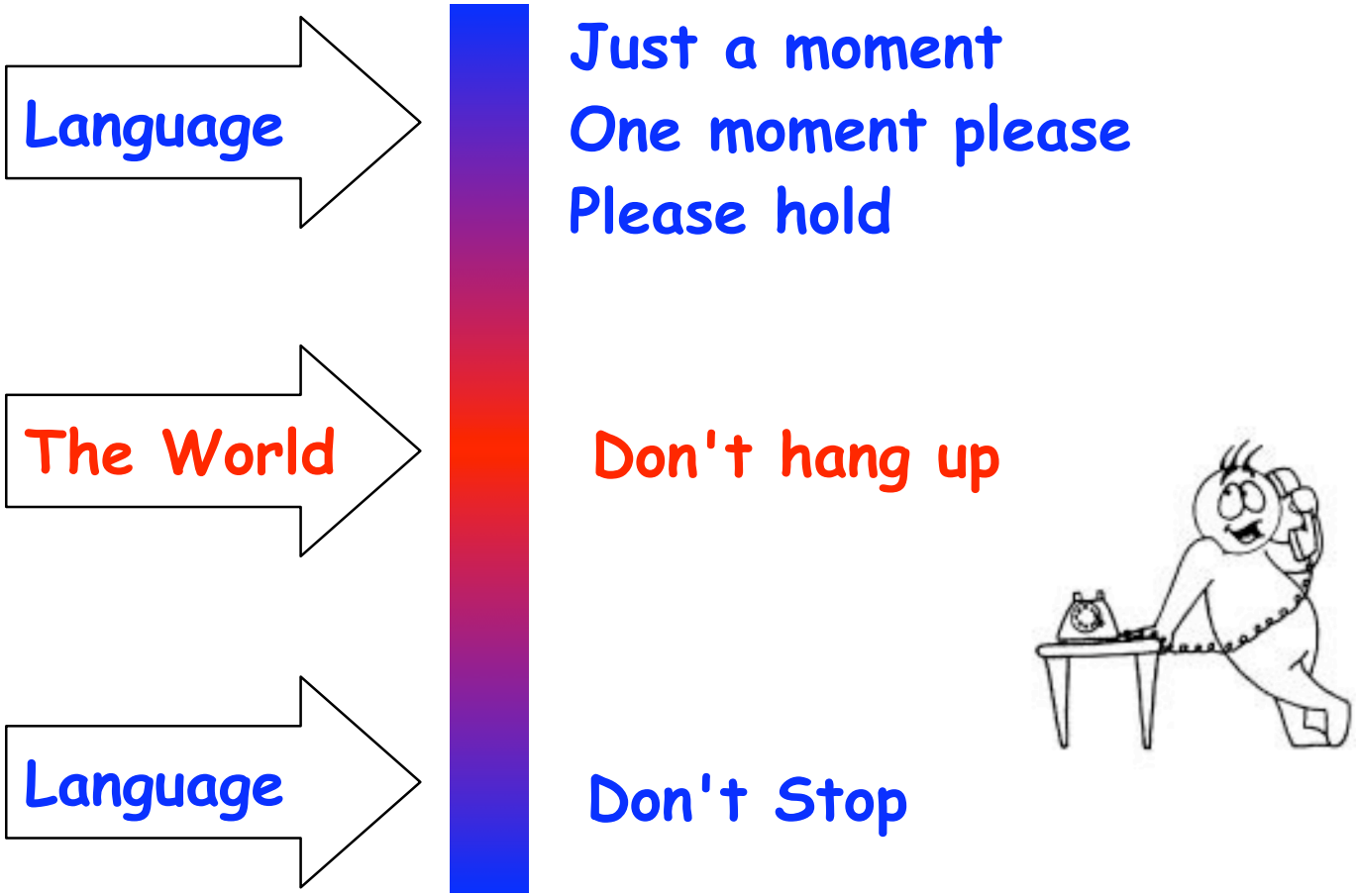

Where do you want me?

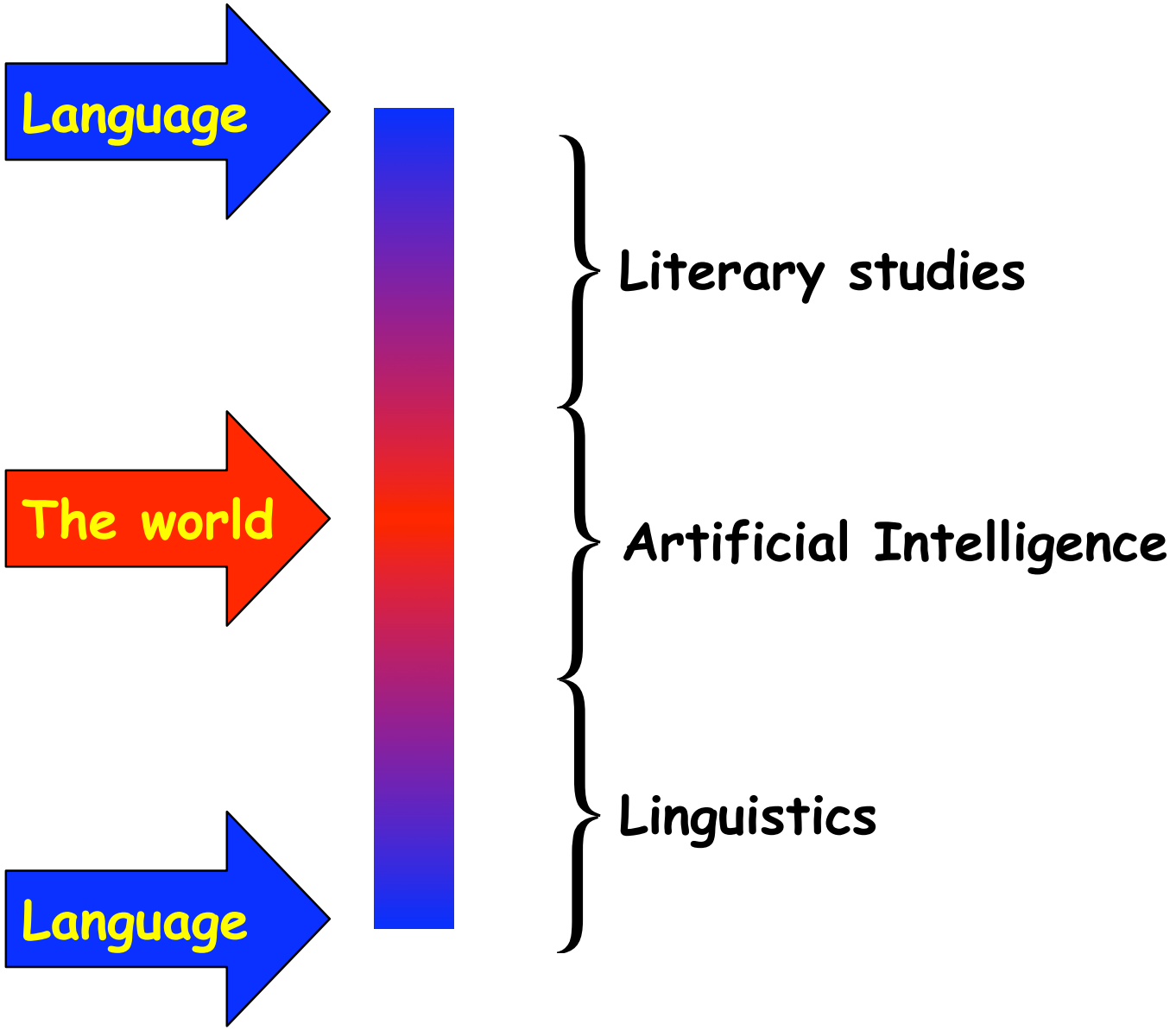
Where do you want me to ... sit?
stand?
sign?
tie up my boat?

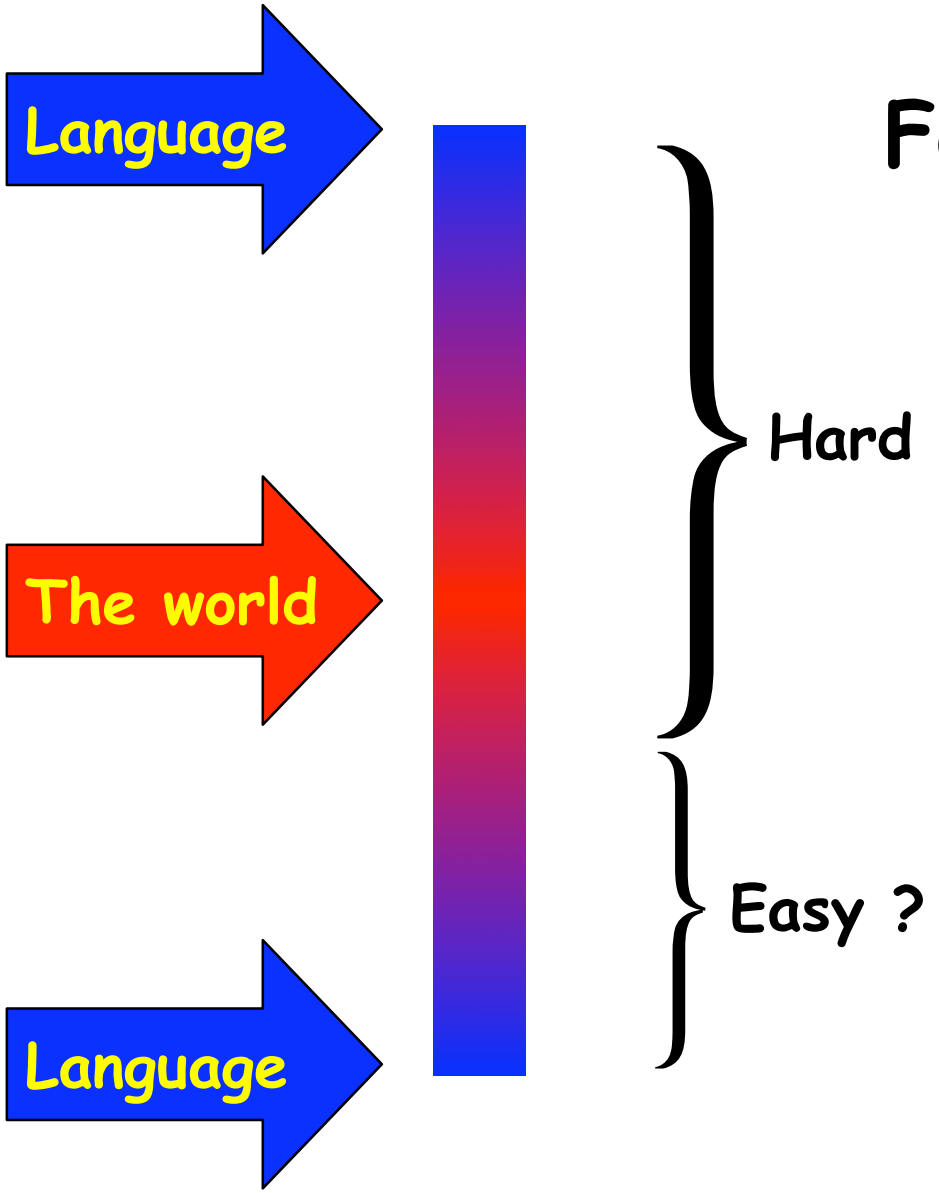
Where do you want me to put myself?

# For example

Ne quittez pas!



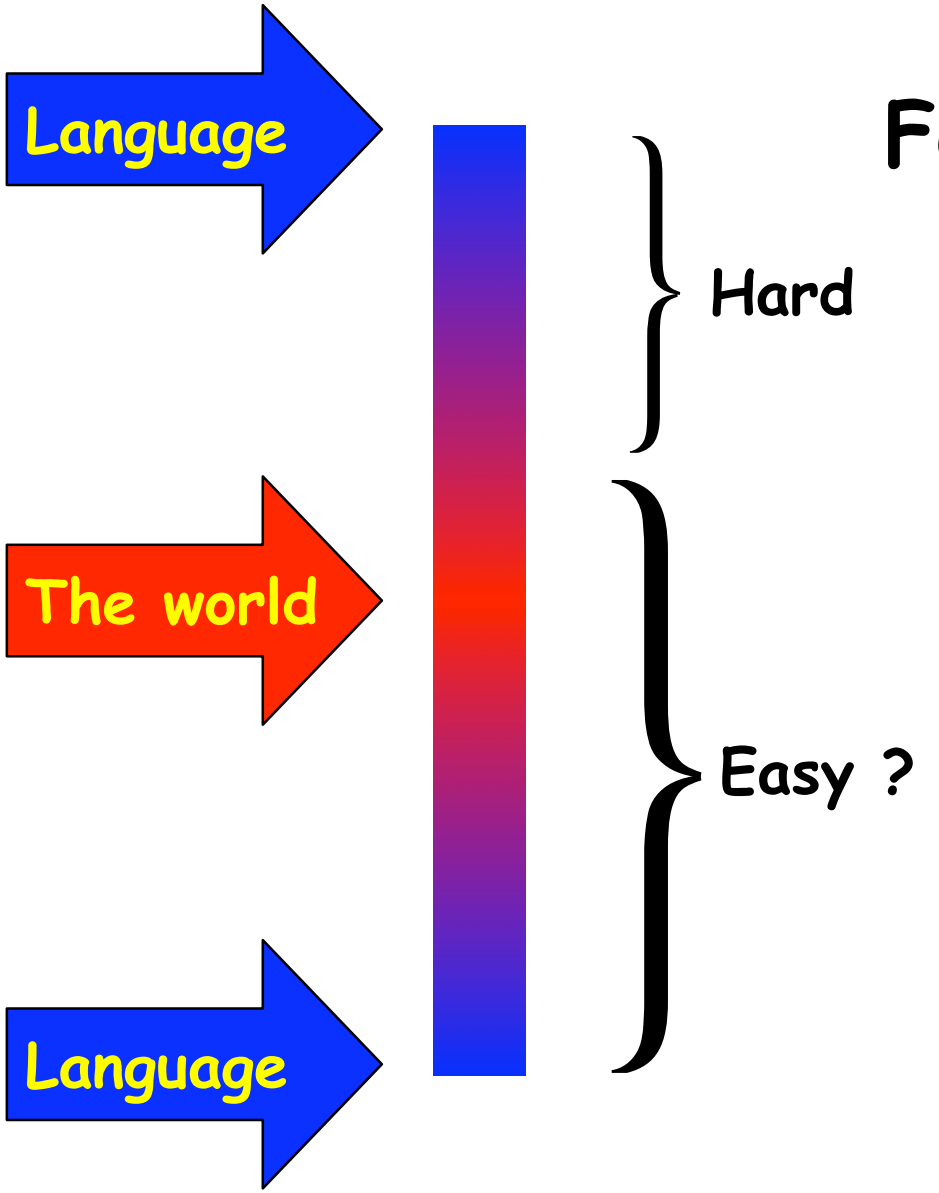

Language
The world
Language
Literary studies
Artificial Intelligence
Linguistics




# For Machines

Hard

Easy ?




# For People

Hard

Easy ?

# What is

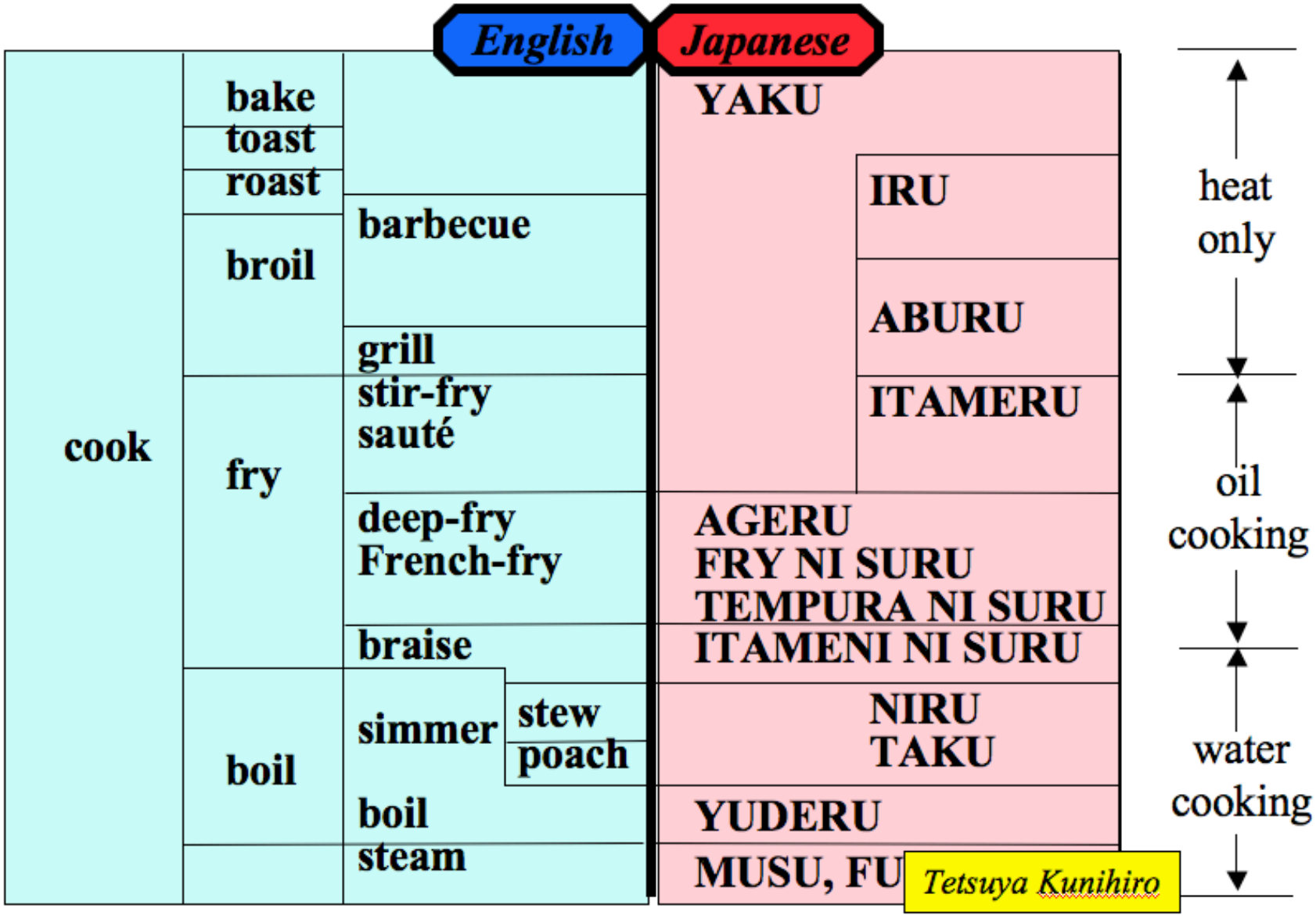
# Meaning?

***It depends what the meaning of "is" is.***

**William Jefferson Clinton**

# Meaning

- **Know**
  - **Connaître / savoir**
  - **Kennen / Wißen**
  - **Weißt du eine Kneipe ...?**
- **Go**
  - **Gehen / fahren / ...**
  - идти / ехать / ходить


English
Japanese
cook
bake
toast
roast
broil
barbecue
grill
fry
stir-fry
sauté
deep-fry
French-fry
braise
boil
simmer
stew
poach
boil
steam
YAKU
IRU
ABURU
ITAMERU
AGERU
FRY NI SURU
TEMPURA NI SURU
ITAMENI NI SURU
NIRU
TAKU
YUDERU
MUSU, FU
Tetsuya Kunihiro
heat only
oil cooking
water cooking

# Required additions/deletions

Progressive
tag questions
Japanese: determiners, zero pronouns, Yo/ne politeness

| | |
|---|---|
| **Gender** | **French** |
| **Tense** | **Chinese** |
| **Articles** | **Japanese** |
| **Aspect** | **Russian** |
| **Pronouns** | **Italian** |
| **...** | |

| | | |
|---|---|---|
| **chaise** | **fauteuil** | **siège** |
| **fleuve** | **rivière** | |
| **savoir** | **connaître** | |
| **livre** | **cahier** | **carnet** |
| **feu** | **phare** | **voyant** |
| **...** | | |

# Terminology

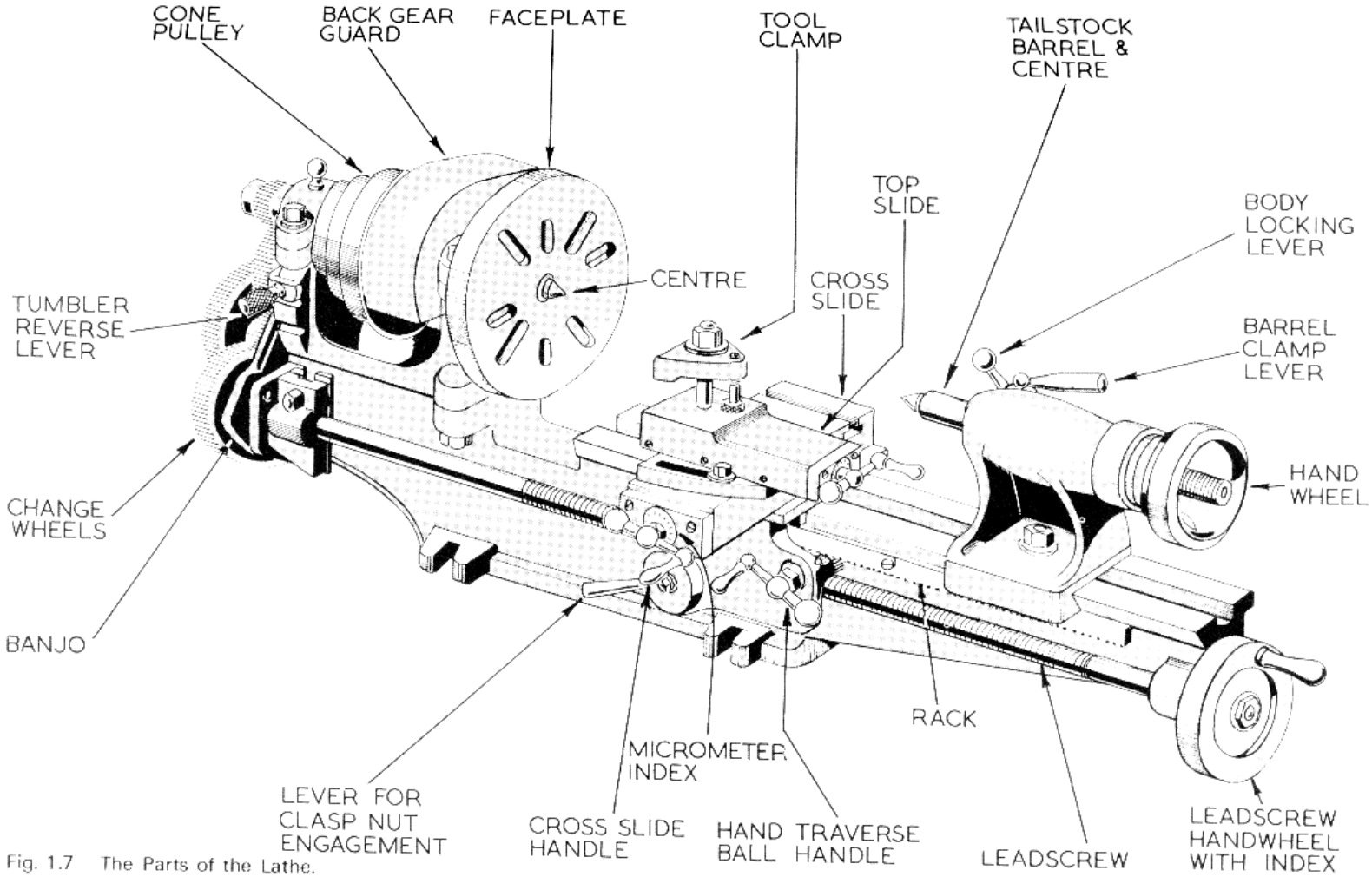
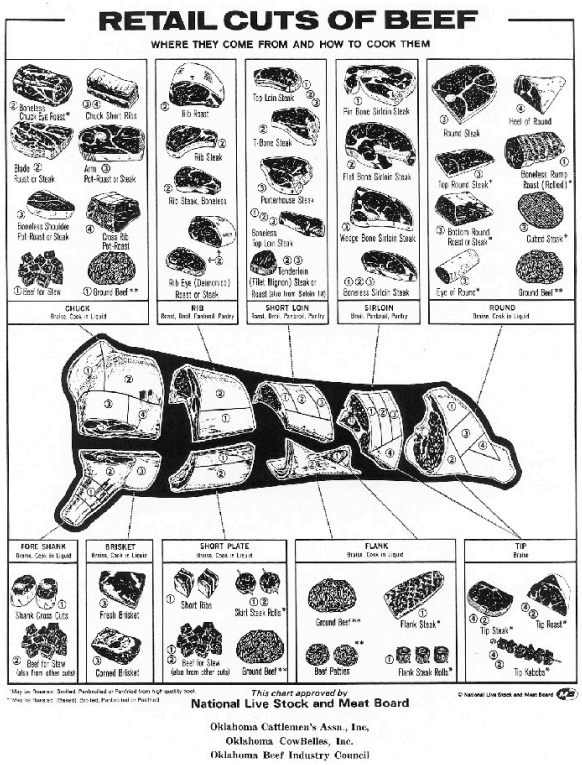



Fig. 1.7 The Parts of the Lathe.

# The Semantic Grid



**Ontological promiscuity**

**-- Hobbs**

**The bloated universe**

**-- Quine**

# Culture & the Semantic Grid

Two no trumps, short stop, goal keeper, end run

Happy hour, a hair of the dog

Alimony, juge d'instruction

value-added tax, home owner's policy

nut, hot tea, café/espresso

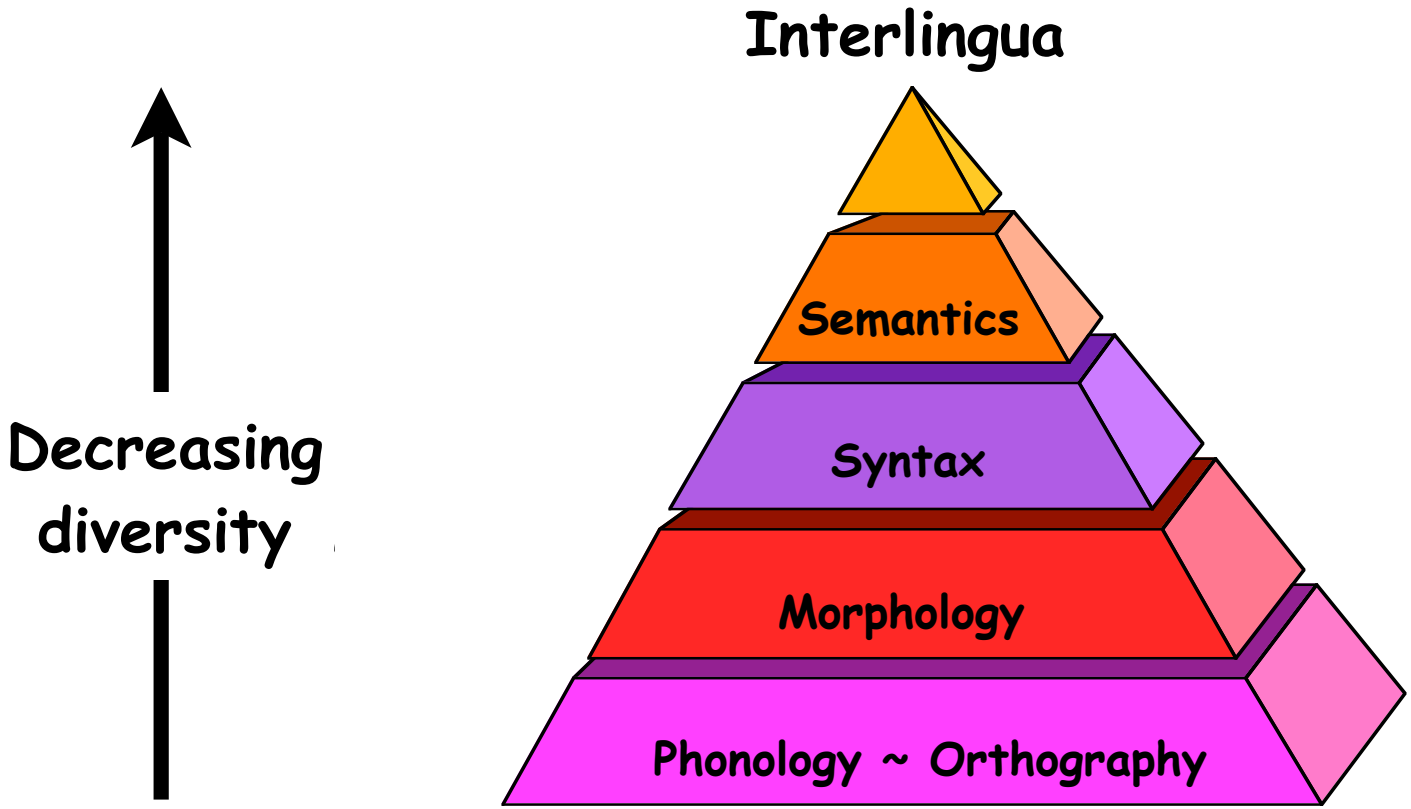
n-th floor, n pièces

2-piece, 2-seater, deux roues, 6-pack

Second reading.  Do I have a second?

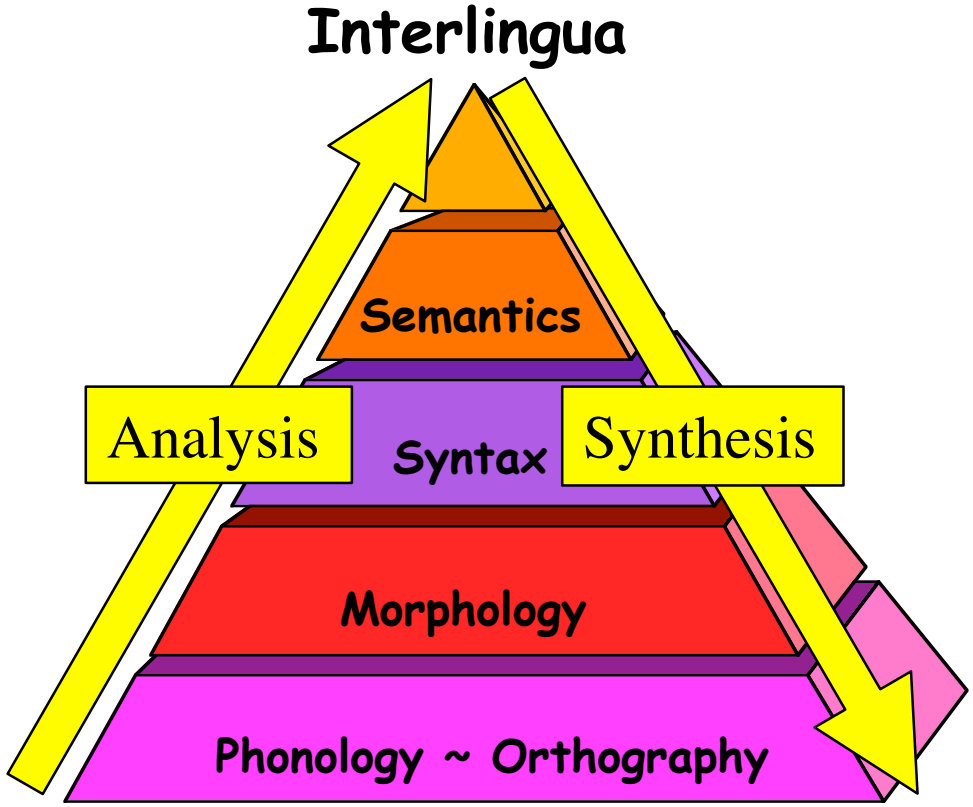
# From a Linguistic Point of View

# Vauquois' Triangle



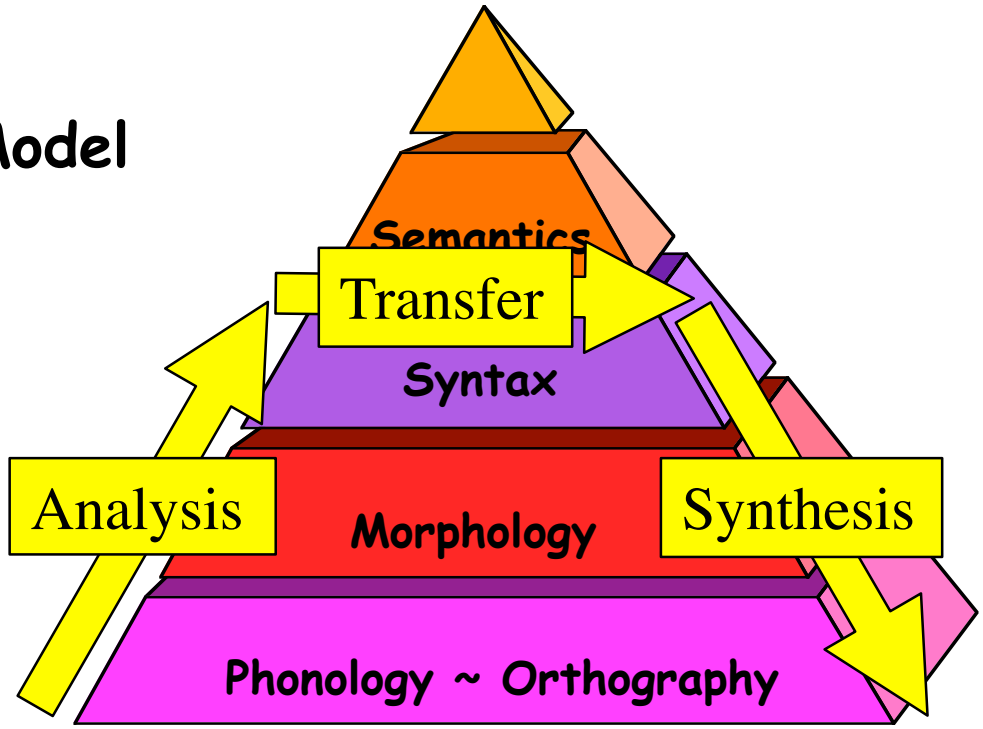
# Vauquois' Triangle

Interlingual Translation



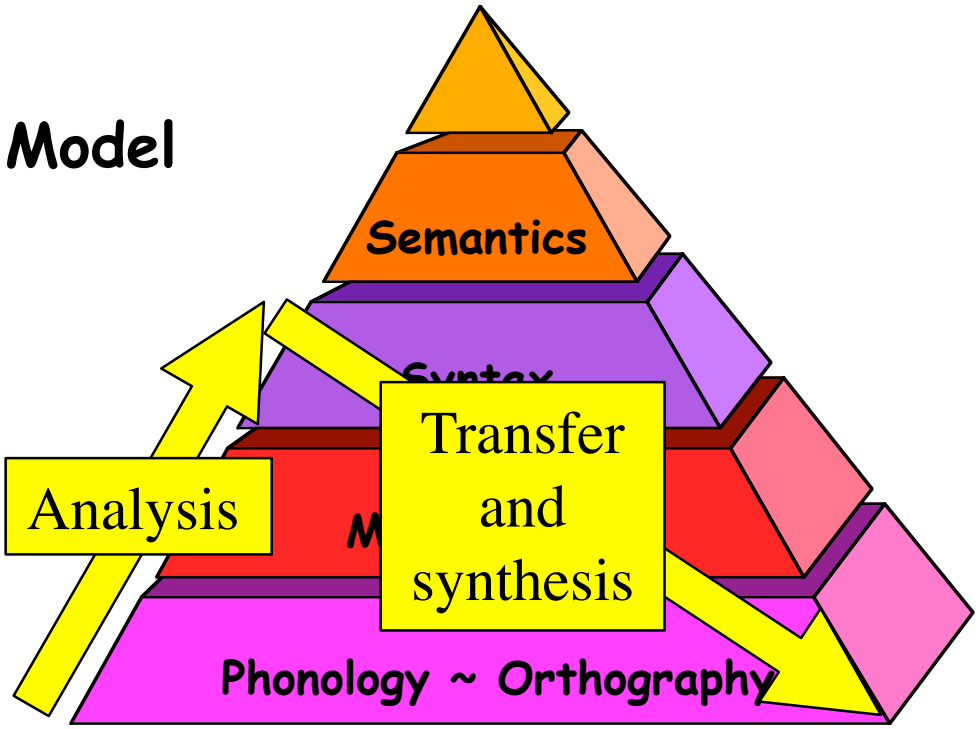
# Vauquois' Triangle

The Academic Model



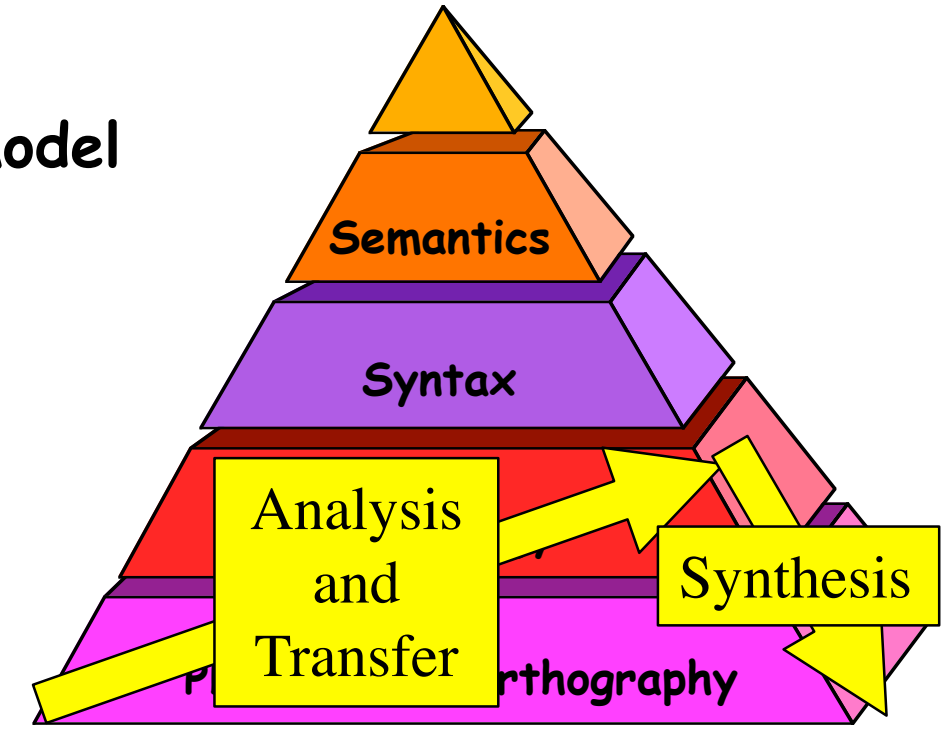
# Vauquois' Triangle

The Commercial Model



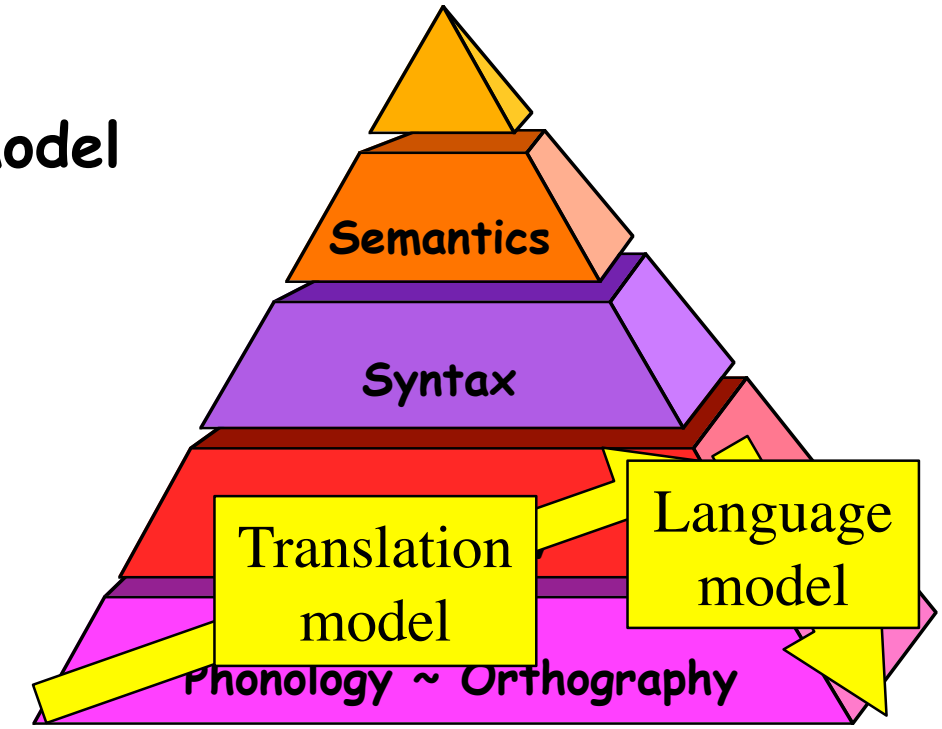
# Vauquois' Triangle

The Statistical Model



# Vauquois' Triangle

The Statistical Model



# The perception

## LINGUISTICS HAS FAILED TECHNOLOGY

- It has too narrow a focus
  - It concentrates on fringe phenomena
  - It luxuriates in ambiguities but is not interested in resolving them
  - It rarely gets beyond the sentence
- It is not robust
  - It is too laborious
  - Human judgements are not objective or consistent
- It is not about communication

# The response

## LANGUAGE PROCESSING IS ONLY PARTLY LINGUISTIC

**It has too narrow a focus**

- It focuses on crucial cases
- It luxuriates in ambiguities and is not responsible for resolving them
- It rarely gets beyond the sentence because that's where the action is

**It is not robust**

- It is too laborious without appropriate (horizontal) abstractions
- Human judgements are not objective or consistent But it's human language!

**It is not about communication**

- It's about part of it

# Crucial Cases

This is the violin that the sonatas are easy to play ◆ ◆ on

*These are the sonatas that the violin is easy to play ◆ ◆ on

Every farmer that owns a donkey beats it

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

# Ambiguities

## Lexical

They met at the bank of the river

He works at the bank by the river

## Morphology

The fish seemed very expensive

This is an untiable knot

They are unionized

## Syntactic

I sent the letter to Adams

The university graduate student admissions policy manual

## Semantic

I didn't take it back because I needed it here.

# Sentences

**Dialog and discourse seem to be structured**

- **weekly**
- **pragmatically**

**<u>Nobody</u> is working on larger units?**

# Horizontal Abstraction

Features ~ Properties ~ Attributes

Vowels are ±front, ±rounded, low/mid/high ...

German nouns and NPs are Nom/Acc/Gen/Dat × Masc/Fem/Neut × Sing/Plur × Count/Mass (48 combinations). Nouns pluralize with ±umlaut × suffixes -0/-e/-en/-er (48 × 2 × 4 = 384).

French nonperifrastic finite verbs are 1st/2nd/3rd person × sing/plur × (pres/imperf × indic/subj + fut/cond) (36 combinations)

# Horizontal Abstraction

NP.nom.masc.sg ➜ Det.nom.masc.sg N.nom.masc.sg

NP.nom.masc.pl ➜ Det.nom.masc.pl N.nom.masc.pl

NP.nom.fem.sg ➜ Det.nom.fem.sg N.nom.fem.sg

⋮

NP.dat.neut.pl ➜ Det. dat.neut.pl N.dat.neut.pl

**Zimmer (room) is 7 ways ambiguous**

**[dat plur is Zimmern]**

# Horizontal Abstraction

**This book is hard to believe a student could read ◆ quickly**

**This is a book I believe a student could read ◆ quickly**

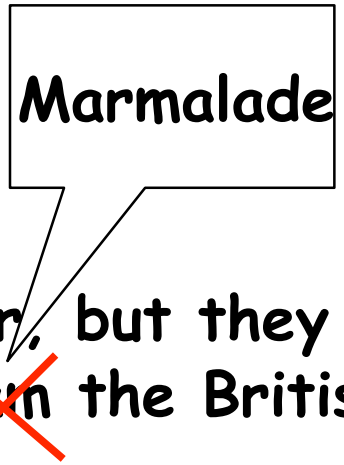
**Which of these books do you believe a student could read ◆ quickly?**

***A* sentence but for the lack of one noun phrase**

# Linguistic facts

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

# Linguistic facts

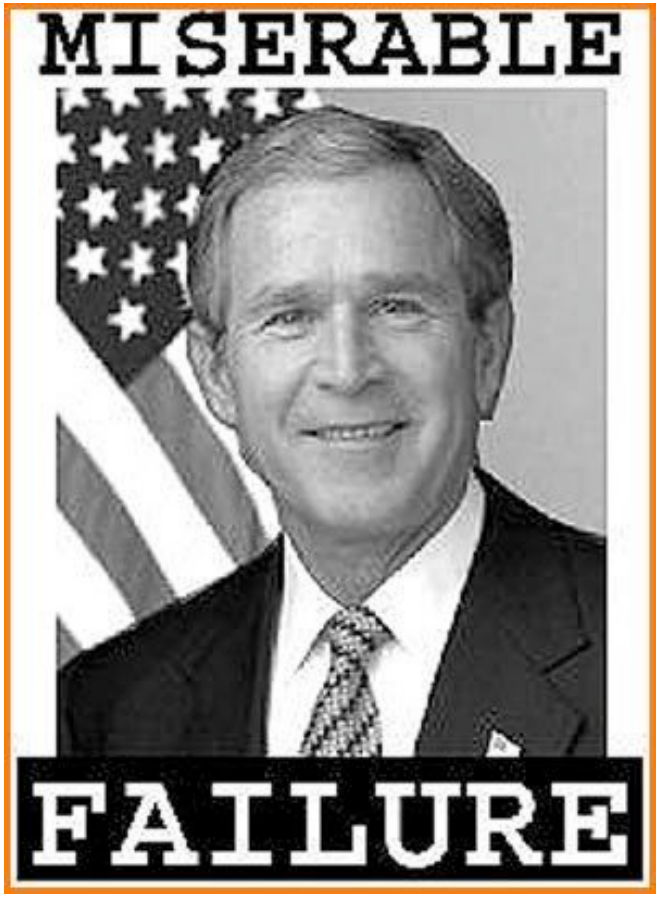



Seville oranges are quite bitter, but they are good for making the kind of ~~jam~~ the British like with their breakfast.

# Linguistic Facts

I usually go to work on the bus

# But it was all thought to be a




So ...

# 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

# Linguistic rules require addition of nonlinguistic Information

He sat {in / on} the chair

Il {s'est assis / était assis} {sur la chaise / dans la fauteil}

Elle écrivait des lettres

She {wrote / was writing} {letters / some letters}

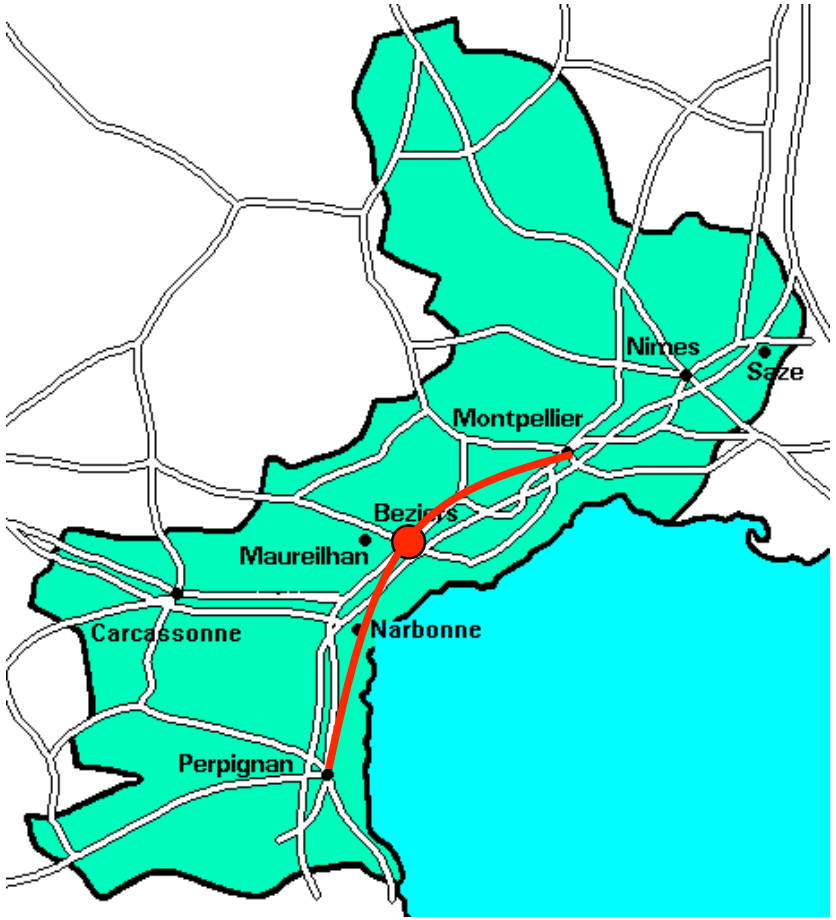

Est_ce que ce train va a Perpignan?

Does this train go to Perpignan?

No, it stops in Beziers.

Fährt dieser Zug nach Perpignan?

Nein, er {hält / endet} in Béziers



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?

# 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

# Doing it by numbers

- What words are most likely to occur in a translation of this sentence, given the source words that it contains and the translations we have seen?
- What order should they be in, given what we know about other sentences in the target language?

# The Statistical Approach: Training

## The translation model

Find pairs of words ("phrases") that have a high probability of occurring opposite one another in sentences that are translations of one another.

## The Language Model

Find short sequences of words (N-grams) that have a high probability of occurring together.

## Other stuff

Fertility

Distortion ...

# Model Evaluation

**Compare translations to human gold standard(s) using a similarity measure.**

**"Bleu" score—number of trigrams shared by candidate and gold standard(s)**

**N.B. The better the system gets, the less reliable the measure becomes.**

# Unfortunately we have …

Zipf's law

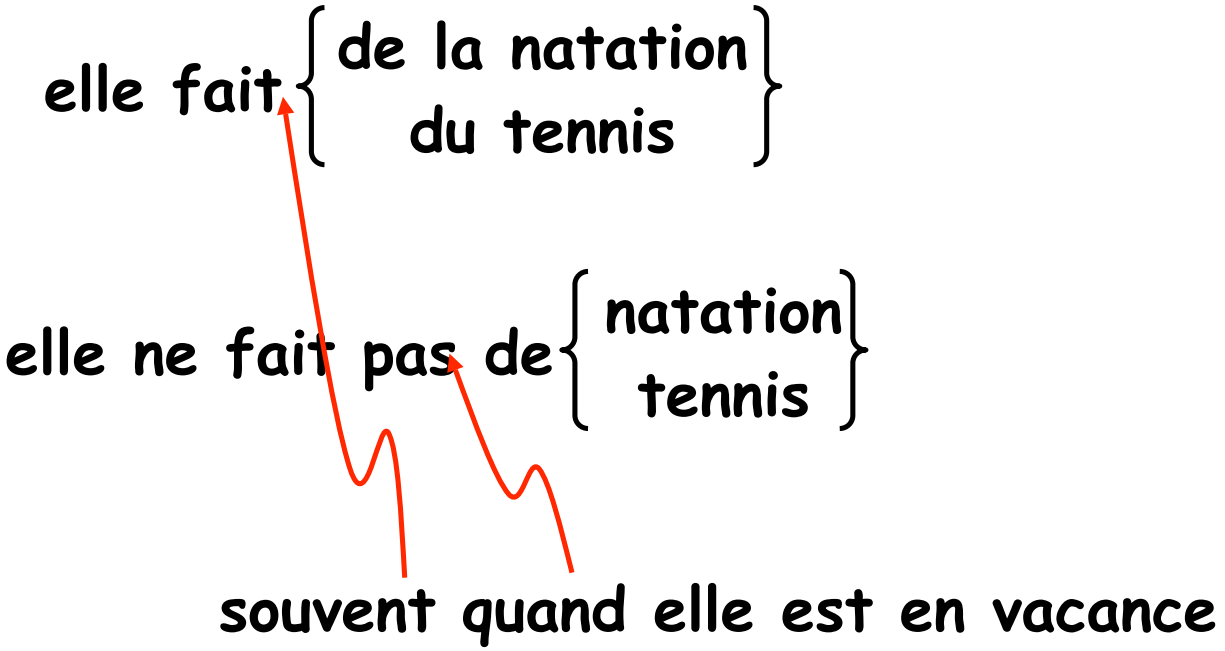
Locality

Emergent Properties

AI

Bleu score

# Linguistic Facts—Locality



# 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

**Pick up the red token off the table**

**Puts it in the box**

# Proposals

- Hybrids
- Monolingual human consultants
  - Reflective Editing
- Triangulation

# Reflective Editing

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.

# Triangulation

There are three windows in the room

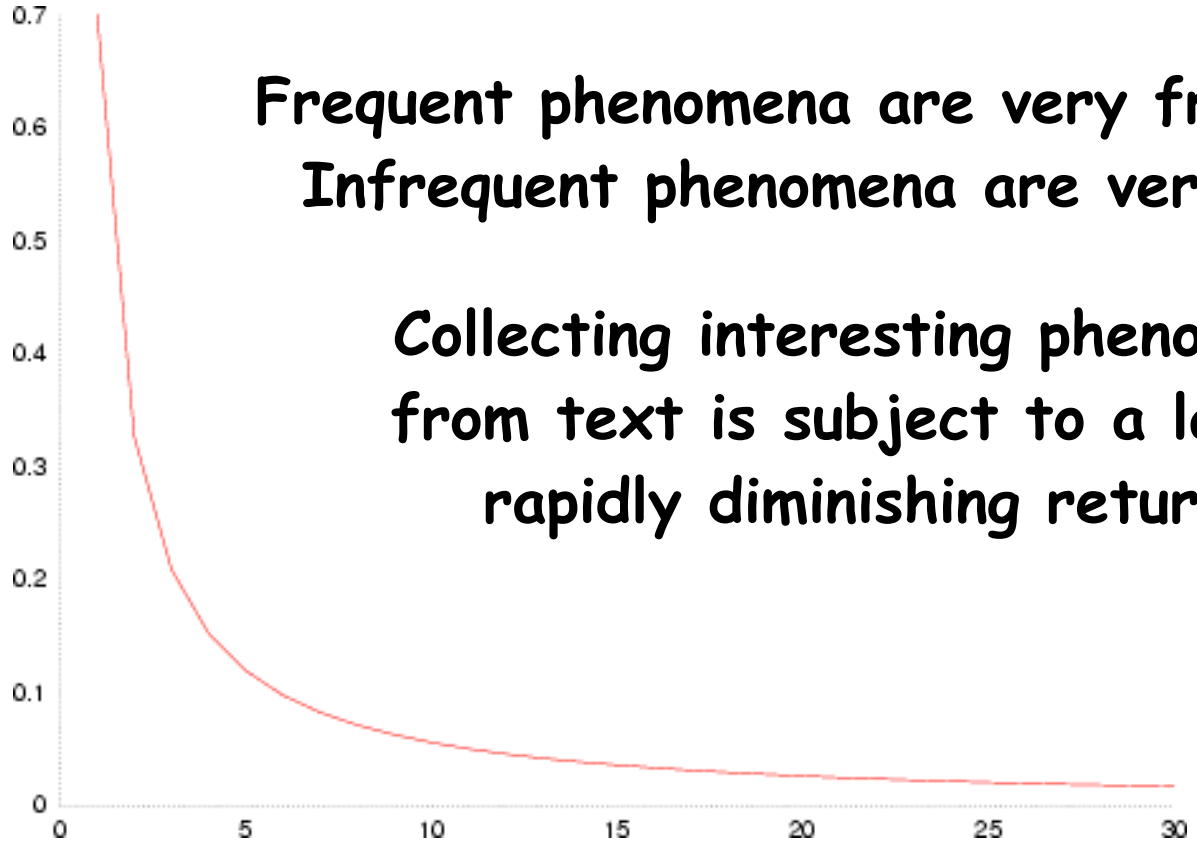
**Il y a trois fenêtres dans la salle.**

**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

# 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

# Emergent Properties

**The important facts about language may not be 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**