# Languages of the World
## (or why is MT so hard!)

Cristina España-Bonet
DFKI GmbH

Summer Semester 2023

11th April 2023

# Why is MT hard?

## Inspired by Josef van Genabith's slides

# Outline

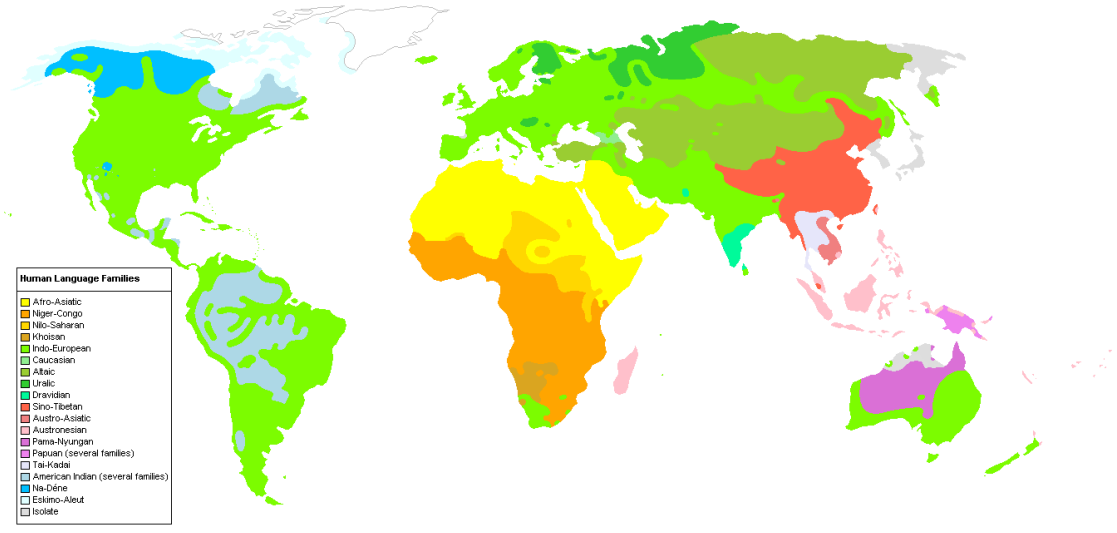
# Languages of the World

## Some Numbers

- There are more than 7000 languages
  (even if the definition of language is not straightforward!)
- 141 language families
  (6 of them account for 2/3 of all languages and 5/6 of the world's population)

# Languages of the World

## Language Families



# Languages of the World

## Some Numbers

- There are more than 7000 languages
  (even if the definition of language is not straightforward!)
- 141 language families
  (6 of them account for 2/3 of all languages and 5/6 of the world's population)

Explore:

| | |
|---|---|
| Ethnologue | `https://www.ethnologue.com/` |
| Glottolog | `http://glottolog.org/` |
| Linguistic Maps | `http://linguisticmaps.tumblr.com/` |

# Languages of the World

## Distribution by number of first-language speakers

| Population range | Living languages | | | Number of speakers | | |
|---|---|---|---|---|---|---|
| | *Count* | *Percent* | *Cumulative* | *Total* | *Percent* | *Cumulative* |
| 100,000,000 to 999,999,999 | 8 | 0.1 | 0.1% | 2,736,892,880 | 40.38033 | 40.38033% |
| 10,000,000 to 99,999,999 | 85 | 1.2 | 1.3% | 2,699,413,660 | 39.82735 | 80.20768% |
| 1,000,000 to 9,999,999 | 308 | 4.3 | 5.7% | 962,414,866 | 14.19954 | 94.40722% |
| 100,000 to 999,999 | 969 | 13.7 | 19.3% | 309,471,921 | 4.56597 | 98.97319% |
| 10,000 to 99,999 | 1,803 | 25.4 | 44.7% | 61,598,950 | 0.90884 | 99.88203% |
| 1,000 to 9,999 | 1,969 | 27.7 | 72.5% | 7,516,041 | 0.11089 | 99.99292% |
| 100 to 999 | 1,047 | 14.8 | 87.2% | 466,977 | 0.00689 | 99.99981% |
| 10 to 99 | 316 | 4.5 | 91.7% | 12,150 | 0.00018 | 99.99999% |
| 1 to 9 | 151 | 2.1 | 93.8% | 608 | 0.00001 | 100.00000% |
| 0 | 248 | 3.5 | 97.3% | 0 | 0.00000 | 100.00000% |
| Unknown | 193 | 2.7 | 100.0% | | | |
| *Totals* | 7,097 | 100.0 | | 6,777,788,053 | 100.00000 | |

http://www.ethnologue.com/ethno_docs/distribution.asp?by=size#3

# Languages of the World

## Ranking by number of speakers

| Rank | Language | Primary Country | Total Countries | Speakers (millions) |
|---|---|---|---|---|
| *1* | *Chinese [zho]* | *China* | *38* | *1,299* |
| 2 | Spanish [spa] | Spain | 31 | 442 |
| 3 | English [eng] | United Kingdom | 118 | 378 |
| *4* | *Arabic [ara]* | *Saudi Arabia* | *58* | *315* |
| 5 | Hindi [hin] | India | 4 | 260 |
| 6 | Bengali [ben] | Bangladesh | 4 | 243 |
| 7 | Portuguese [por] | Portugal | 15 | 223 |
| 8 | Russian [rus] | Russian Federation | 18 | 154 |
| 9 | Japanese [jpn] | Japan | 2 | 128 |
| *10* | *Lahnda [lah]* | *Pakistan* | *6* | *119* |
| 11 | Javanese [jav] | Indonesia | 3 | 84.4 |
| 12 | Turkish [tur] | Turkey | 8 | 78.5 |
| 13 | Korean [kor] | South Korea | 6 | 77.2 |
| 14 | French [fra] | France | 53 | 76.8 |
| 15 | German, Standard [deu] | Germany | 28 | 76.0 |
| 16 | Telugu [tel] | India | 2 | 74.8 |
| 17 | Marathi [mar] | India | 1 | 71.8 |
| 18 | Urdu [urd] | Pakistan | 7 | 69.2 |
| 19 | Vietnamese [vie] | Viet Nam | 3 | 68.0 |
| 20 | Tamil [tam] | India | 7 | 66.7 |

# Characteristics of Human Languages

## Description

Human languages are a tool to
communicate thoughts and they are

elegant, efficient, flexible, complex

# Characteristics of Human Languages

## Description

Human languages are a tool to
communicate thoughts and they are

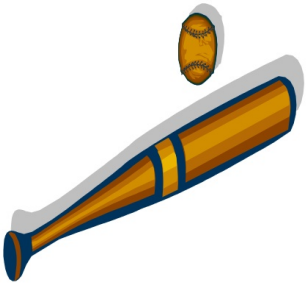
elegant, efficient, flexible, complex

Cool for a human, but for a machine...

# Characteristics of Human Languages

## One word/sentence may mean many things

### Homonymy and Polysemy

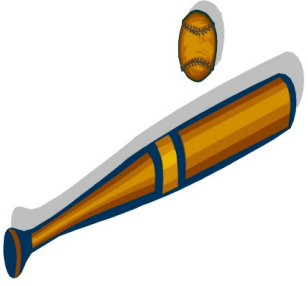


# Characteristics of Human Languages

## One word/sentence may mean many things

### Homonymy and Polysemy

bat



# Characteristics of Human Languages

## Meaning depends on context

Bats!

# Characteristics of Human Languages

## Meaning depends on context

Bats! (I'm inside a cave)

# Characteristics of Human Languages

## Meaning depends on context

Bats! (I'm inside a cave)

This bat is brown

# Characteristics of Human Languages

## Meaning depends on context

Bats! (I'm inside a cave)

This bat is brown (I'm watching a baseball match)

# Characteristics of Human Languages

## Meaning depends on context

Bats! (I'm inside a cave)

This bat is brown (I'm watching a baseball match)

Streaks and slumps are as common to baseball as bats

He bats the flies away

We bat around a wide variety of issues

When he told me what he'd done, I didn't bat an eye

# Characteristics of Human Languages

## Many ways of saying the same thing

### Synonymy

gift present

# Characteristics of Human Languages

## Literal and figurative language (metaphors)

This coffee shop is an ice box

vs.

It is very cold in this coffee shop

# Characteristics of Human Languages

## Language is culture

The early bird catches the worm

vs.

Morgenstund hat Gold im Mund

vs.

Qui matina fa farina

vs.

A quien madruga, Dios le ayuda

# Characteristics of Human Languages

## Language is culture

Jurafsky & Martin, 2007 (Ch.25)

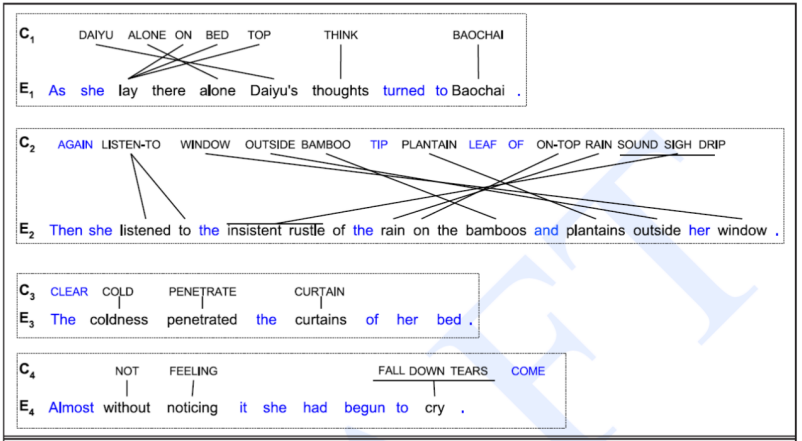



**Figure 25.1** A Chinese passage from *Dream of the Red Chamber*, with the Chinese words represented by English glosses IN SMALL CAPS. Alignment lines are drawn between 'Chinese' words and their English translations. Words in blue are Chinese words not translated into English, or English words not in the original Chinese.

# Universal Aspects

## Cultures & Language

Different cultures share basic concepts and actions
and communicate them with words


- Vocabulary: We all have a mother, a father, and see trees, water...
- Grammar: nouns and verbs

# Non-Universal Aspects

## Cultures & Language

But different cultures communicate them with words in different ways and cover different necessities

46 words for snow in Icelandic


https://reportsfromtherock.wordpress.com/page/2/

# (Non-)Universal Aspects

## WEAT1 and WEAT2 Original Lists

25 items of selected concepts:

| | | |
|---|---|---|
| WEAT1 target items | | |
|  | Flowers | orchid... |
|  | Insects | ant... |
| WEAT2 target items | | |
|  | Instruments | guitar... |
|  | Weapons | arrow... |
| WEAT1 and WEAT2 attributes | | |
|  | Pleasant | caress... |
|  | Unpleasant | abuse... |

# (Non-)Universal Aspects

## Original and X-WEAT Lists

### Original version (WEAT1, WEAT2)

[Battig and Montague, 1969; Bellezza et al., 1986; Greenwald et al., 1998]

- Collected from college students in Eastern US
- Frequent terms
- Non-ambiguous terms

# (Non-)Universal Aspects

## Original and X-WEAT Lists

### Original version (WEAT1, WEAT2)

[Battig and Montague, 1969; Bellezza et al., 1986; Greenwald et al., 1998]

- Collected from college students in Eastern US
- Frequent terms
- Non-ambiguous terms

### Multilingual version (X-WEAT)

[Lauscher and Glavaš, 2019; Lauscher et al., 2020]

- Literal translation
- Arabic, Croatian, German, Italian, Russian, Spanish and Turkish

# (Non-)Universal Aspects

## Features and Issues with WEAT and X-WEAT

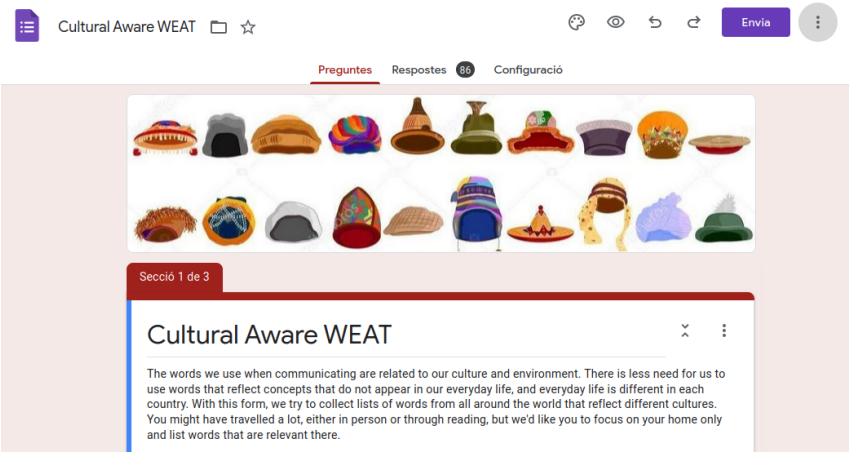
- WEAT: American English, represents the culture of the (Eastern) US
- X-WEAT: Multilingual, but represents the culture of the (Eastern) US!
  —and this applies to all NLP using translation—
  - duplicates? (violin, fiddle $\rightarrow$ violín)
  - frequent terms? (gnat $\rightarrow$ jején)
  - non-ambiguous terms? (blade $\rightarrow$ hoja)

Do we always want/need MT?

- CA-WEAT: Multilingual and culturaly aware

# (Non-)Universal Aspects

## Collecting CA-WEATs. Let's discuss!




https://github.com/cristinae/CA-WEAT

# Non-Universal Aspects

## Differences among languages

- Lexicon (and idioms!) are the most evident differences at semantic level
  - Do we always want literal translation?
  - How do we reflect cultural differences in MT? Can we do automatic localisation?
  - Can multimodality help?
- But other differences among languages make MT difficult!

# Non-Universal Aspects

## Differences among languages

Besides different necessities for concepts and lexicon, languages differ in morphology

Aliikkusersuillammassuaanerartassagaluarpaalli

# Non-Universal Aspects

## Differences among languages

Besides different necessities for concepts and lexicon, languages differ in morphology

Western Greenlandic:
Aliikkusersuillammassuaanerartassagaluarpaalli

English:
However, they will say that he is a great entertainer but

# Non-Universal Aspects

## Differences among languages

Besides different necessities for concepts and lexicon, languages differ in morphology

Aliikkusersuillammassuaanerartassagaluarpaalli

```
aliikku-sersu-i-llammas-sua-a-nerar-ta-ssa-galuar-paal-li
```

```
entertainment-provide-SEMITRANS-one.good.at-COP-say.that-
REP-FUT-sure.but-3.PL.SUBJ/3SG.OBJ-but
```

However, they will say that he is a great entertainer but

# Non-Universal Aspects

## Morphology

Analytic/Isolating Languages

- low morpheme-per-word ratio and low inflection

Synthetic Languages

- high morpheme-per-word ratio

  Polysynthetic one word, many morphemes
  Fusional each morpheme denotes several features
  Agglutinative one morpheme per feature

# Non-Universal Aspects

## Morphology Examples

Analytic

Mandarin

我 给 你 一本 书

I give you one book

# Non-Universal Aspects

## Morphology Examples

**Analytic** Mandarin

我 给 你 一本 书

I give you one book

**Polysynthetic** Western Greenlandic

Aliikkusersuillammassuaanerartassagaluar...

# Non-Universal Aspects

## Morphology Examples

**Analytic** Mandarin

我 给 你 一本 书

I give you one book

**Polysynthetic** Western Greenlandic

Aliikkusersuillammassuaanerartassagaluar...

**Fusional** Catalan

sortiré

-é: 1PS, FUT., IND., ACTIVE

# Non-Universal Aspects

## Morphology Examples

**Analytic** Mandarin

我 给 你 一本 书

I give you one book

**Polysynthetic** Western Greenlandic

Aliikkusersuillammassuaanerartassagaluar...

**Fusional** Catalan

sortiré

-é: 1PS, FUT., IND., ACTIVE

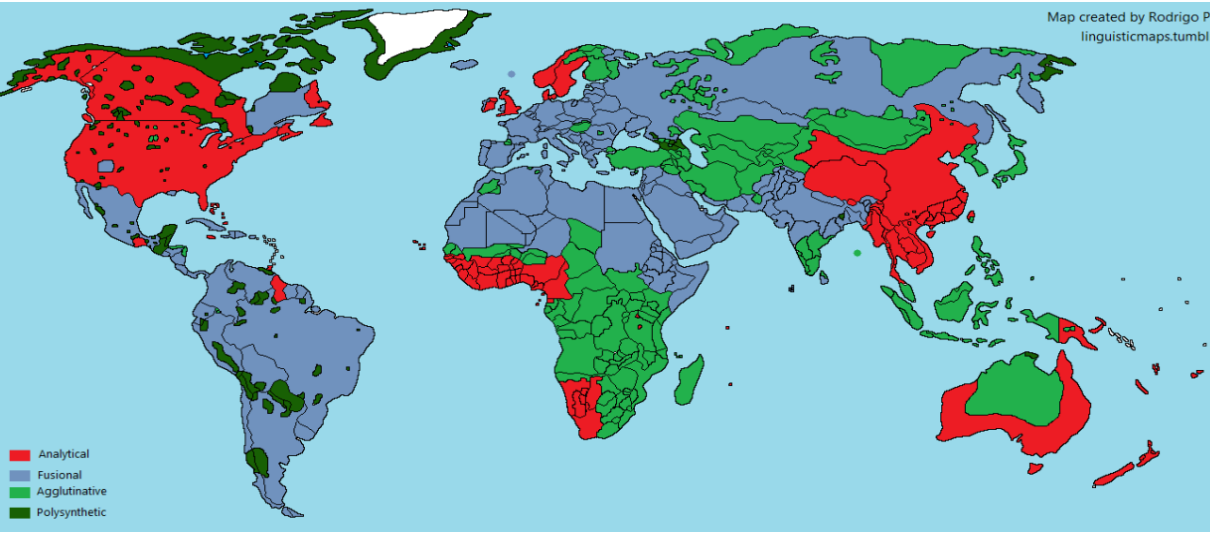
**Agglutinative** Turkish

evlerinizden

ev-ler-iniz-den: house, PL., your, from

# Non-Universal Aspects

## Distribution of Languages by Morphology Type



# Non-Universal Aspects

## Syntax

Declarative sentences with a Subject, a Verb and an Object can follow different orders

SOV: Japanese, Hindi...
`Inu ga` (subject) `neko` (object) `o oikaketa` (verb)

SVO: English, French, Mandarin...
`The dog` (subject) `chased` (verb) `the cat` (object)

VSO: Irish, Arabic...
`yutarid` (verb) `alkalb` (subject) `alqut` (object)

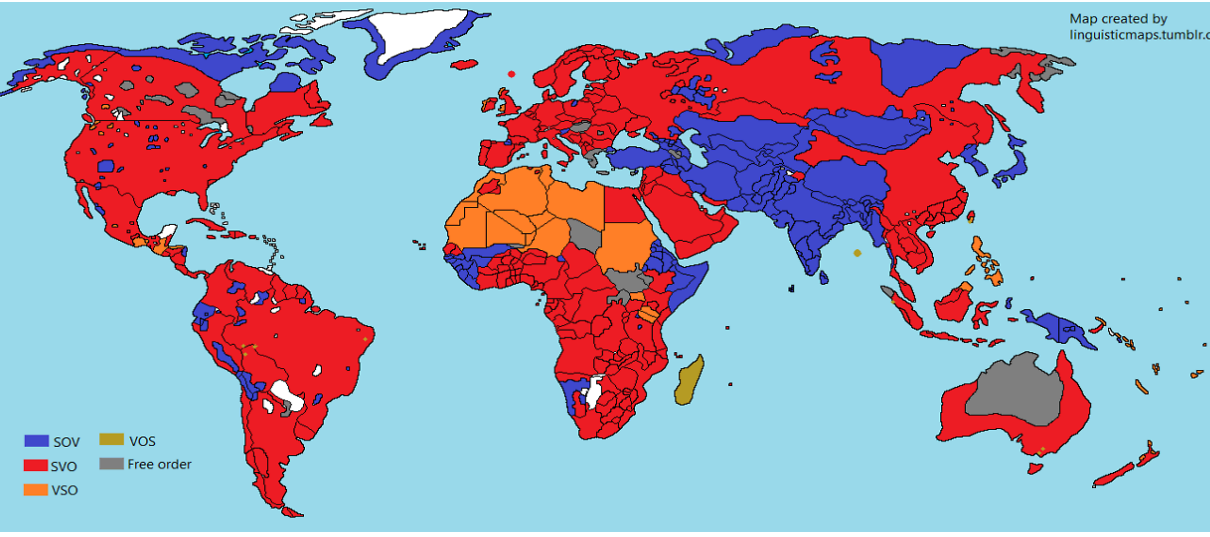
# Non-Universal Aspects

## Syntax – Observations

- Synthetic languages have more freedom in order than isolating languages
- SVO tend to have prepositions
- SOV tend to have postpositions

# Non-Universal Aspects

## Distribution of Languages by Syntax Type



# Non-Universal Aspects

## Other differences among languages

- Omision of elements (e.g. pronouns)
- Short range order
  - Adj Noun Blue house
  - Noun Adj Casa blava
- Structure, e.g. dates
  - DD/MM/YY British English
  - MM/DD/YY American English
  - YY/MM/DD Japanese

# Summary

## Keep in mind

Remember all these aspects of languages when we design and evaluate machine translation systems

- Next week you'll learn with Koel about MT evaluation. Think about how can we measure translation quality with ambiguous words, polysemic, cultural differences?
- The following week we'll recover your CA-WEATs to observe cultural differences with US English and start with MT systems.

# Summary

## Wait!

# Summary

## By the way, I do:

How is German?

And your language?

# Summary

# References


- Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition. Daniel Jurafsky & James H. Martin. 2007. Chapter 25.
- Introduction to Linguistics – Morphological Typology. Slides. Jonathan Manker
- Languages of the world. Slides of the course. Gerhard Jäger