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

Cristina España-Bonet
DFKI GmbH

Summer Semester 2023

24th April 2023

Starting off on the wrong Foot!

Starting off on the wrong Foot!

What we did:

1. As good IT people, we spent 40 minutes setting up the PC

Starting off on the wrong Foot!

What we did:

1. As good IT people, we spent 40 minutes setting up the PC
2. Present the course (1st set of slides)
   - `https://cristinae.github.io/teaching/mt2023/`
   - Questions?
   - Lab2 LCT people

Starting off on the wrong Foot!

What we did:

1. As good IT people, we spent 40 minutes setting up the PC
2. Present the course (1st set of slides)
   - `https://cristinae.github.io/teaching/mt2023/`
   - Questions?
   - Lab2 LCT people
3. Languages of the world (2nd set of slides)
   - Recap today
   - We did not finished anyway, just motivated the difficulty of MT evaluation

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

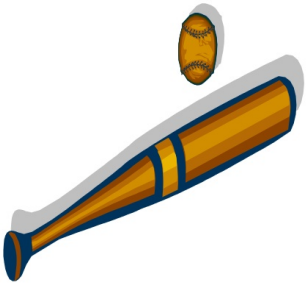
elegant, efficient, flexible, complex

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

Characteristics of Human Languages

One word/sentence may mean many things

Homonymy and Polysemy

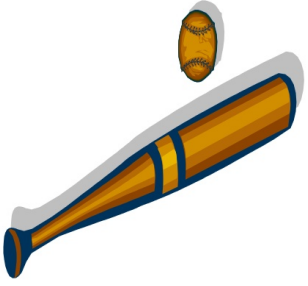


Characteristics of Human Languages

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

Starting off on the wrong foot

Does it have a literal translation in your language?

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

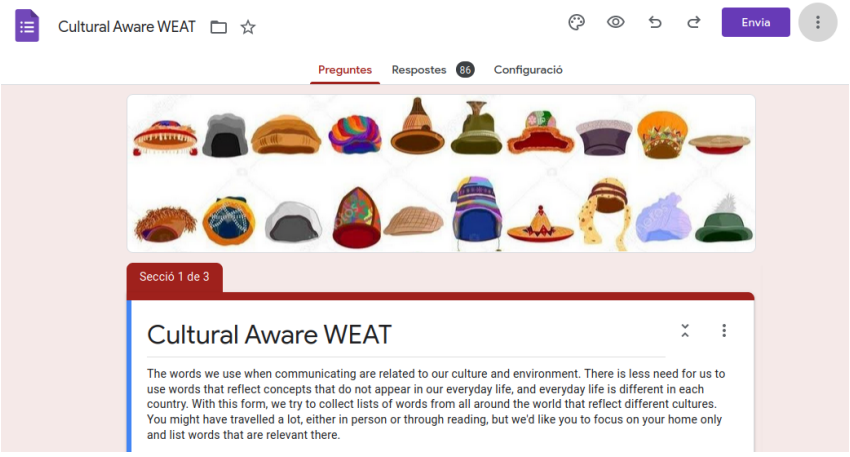
- WEAT: American English, represents the culture of the (Eastern) US
- X-WEAT: Multilingual, but represents the culture of the (Eastern) US!
  —and this applies to all NLP using translation—
  - duplicates? (violin, fiddle $\rightarrow$ violín)
  - frequent terms? (gnat $\rightarrow$ jején)
  - non-ambiguous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

Do we always want/need MT?

- CA-WEAT: Multilingual and culturaly aware

(Non-)Universal Aspects

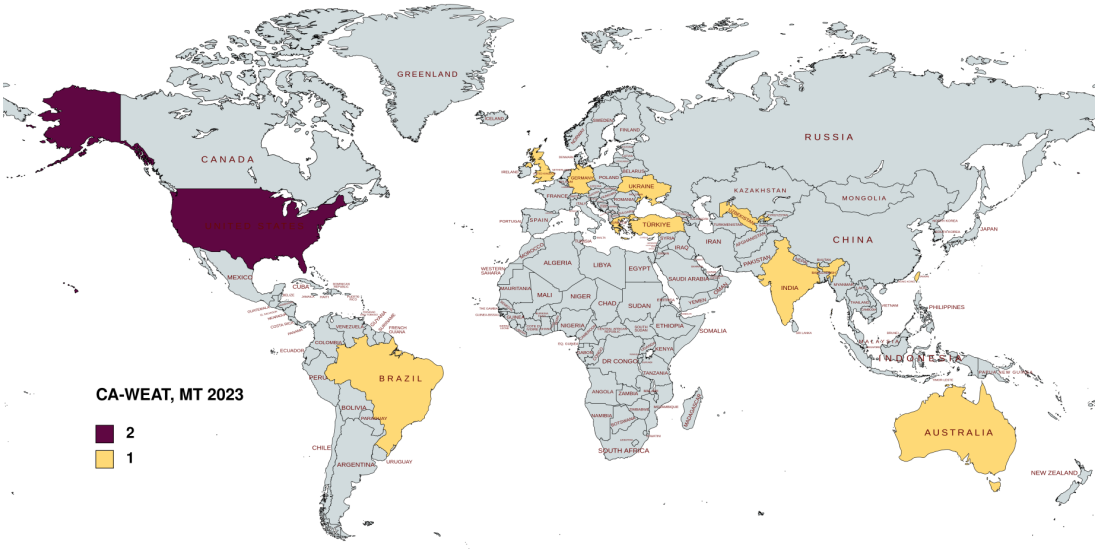
Collecting CA-WEATs. Let's discuss!



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

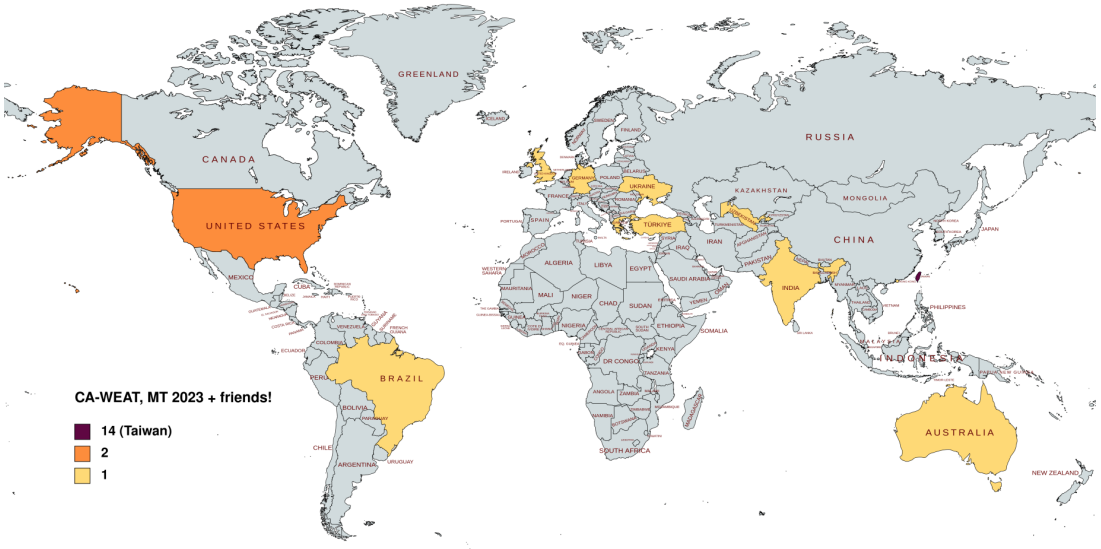
(Non-)Universal Aspects

11(+15!!) Contributions up to now, very Multilingual!



(Non-)Universal Aspects

11(+15!!) Contributions up to now, very Multilingual!



(Non-)Universal Aspects

Your Lists (Monolingual)

apple, pear, grape, strawberry, blackberry, blueberry, raspberry, plum, apricot, orange, tangerine, clementine, lemon, lime, watermelon, pepper, squash, pumpkin, tomato, banana, pineapple, fig, date, mango, papaya

apple, banana, orange, grape, cherry, strawberry, raspberry, blueberry, tangerine, mango, peach, nectarine, pineapple, plum, mandarin, kiwi, papaya, blackberry, blackcurrant, redcurrant, apricot, raisin, gooseberry, pear, melon

apple, orange, banana, kiwi, grape, lemon, cherry, pear, strawberry, blueberry, raspberry, blackberry, avocado, lime, mango, peach, plum, apricot, nectarine, pineapple, papaya, watermelon, lychee, longan, durian

(Non-)Universal Aspects

Your Lists (Monolingual)

apple, pear, grape, strawberry, blackberry, blueberry, raspberry, plum, apricot, orange, tangerine, clementine, lemon, lime, watermelon, pepper, squash, pumpkin, tomato, banana, pineapple, fig, date, mango, papaya

apple, banana, orange, grape, cherry, strawberry, raspberry, blueberry, tangerine, mango, peach, nectarine, pineapple, plum, mandarin, kiwi, papaya, blackberry, blackcurrant, redcurrant, apricot, raisin, gooseberry, pear, melon

apple, orange, banana, kiwi, grape, lemon, cherry, pear, strawberry, blueberry, raspberry, blackberry, avocado, lime, mango, peach, plum, apricot, nectarine, pineapple, papaya, watermelon, lychee, longan, durian

(Non-)Universal Aspects

Your Lists (Monolingual)

 US English	apple, pear, grape, strawberry, blackberry, blueberry, raspberry, plum, apricot, orange, tangerine, clementine, lemon, lime, watermelon, pepper, squash, pumpkin, tomato, banana, pineapple, fig, date, mango, papaya
 UK English	apple, banana, orange, grape, cherry, strawberry, raspberry, blueberry, tangerine, mango, peach, nectarine, pineapple, plum, mandarin, kiwi, papaya, blackberry, blackcurrant, redcurrant, apricot, raisin, gooseberry, pear, melon
 AU English	apple, orange, banana, kiwi, grape, lemon, cherry, pear, strawberry, blueberry, raspberry, blackberry, avocado, lime, mango, peach, plum, apricot, nectarine, pineapple, papaya, watermelon, lychee, longan, durian

(Non-)Universal Aspects

Your Lists (Multilingual). Disclaimer: my translation...

US English

apple, pear, grape, strawberry, blackberry, blueberry, raspberry, plum, apricot, orange, tangerine, clementine, lemon, lime, watermelon, pepper, squash, pumpkin, tomato, banana, pineapple, fig, date, mango, papaya

BR Portuguese

apple, banana, guava, pineapple, apricot, pear, watermelon, orange, lemon, cherry, tangerine, kiwi, pequi, açaí, cashew, hog plum, soursop, strawberry, raspberry, blackberry, plum, peach, passion fruit, lychee, jabuticaba

maçã, banana, goiaba, abacaxi, damasco, pêra, melancia, laranja, limão, cereja, mexerica, kiwi, pequi, açaí, caju, cajá, graviola, morango, framboesa, amora, ameixa, pêssego, maracujá, lichia, jabuticaba

Traditional Chinese

banana, apple, pineapple, guava, orange, grape, peach, cherry, blueberry, Java apple, papaya, lychee, strawberry, tomato, cantaloupe, tangerine, lemon, lime, raspberry, Japanese banana, sugarcane, watermelon, durian, sugar apple, coconut

香蕉, 蘋果, 鳳梨, 芭樂, 柳丁, 葡萄, 水蜜桃, 櫻桃, 藍莓, 蓮霧, 木瓜, 荔枝, 草莓, 番茄, 哈密瓜, 橘子, 檸檬, 茶姆, 覆盆莓, 芭蕉, 甘蔗, 西瓜, 榴蓮, 釋迦, 椰子

(Non-)Universal Aspects

Your Lists (Multilingual). Disclaimer: my translation...

US English

apple, pear, grape, strawberry, blackberry, blueberry, raspberry, plum, apricot, orange, tangerine, clementine, lemon, lime, watermelon, pepper, squash, pumpkin, tomato, banana, pineapple, fig, date, mango, papaya

---

BR Portuguese

apple, banana, guava, pineapple, apricot, pear, watermelon, orange, lemon, cherry, tangerine, kiwi, pequi, açaí, cashew, hog plum, soursop, strawberry, raspberry, blackberry, plum, peach, passion fruit, lychee, jabuticaba

maçã, banana, goiaba, abacaxi, damasco, pêra, melancia, laranja, limão, cereja, mexerica, kiwi, pequi, açaí, caju, cajá, graviola, morango, framboesa, amora, ameixa, pêssego, maracujá, lichia, jabuticaba

---

Traditional Chinese

banana, apple, pineapple, guava, orange, grape, peach, cherry, blueberry, Java apple, papaya, lychee, strawberry, tomato, cantaloupe, tangerine, lemon, lime, raspberry, Japanese banana, sugarcane, watermelon, durian, sugar apple, coconut

香蕉, 蘋果, 鳳梨, 芭樂, 柳丁, 葡萄, 水蜜桃, 櫻桃, 藍莓, 蓮霧, 木瓜, 荔枝, 草莓, 番茄, 哈密瓜, 橘子, 檸檬, 茶姆, 覆盆莓, 芭蕉, 甘蔗, 西瓜, 榴蓮, 釋迦, 椰子

(Non-)Universal Aspects

(Cross-lingual) Cultural Biases in NLP

- Disclaimer: 1 list is just an example, 100 lists start saying something
- Multilingual models trained on an asymmetric distribution of data
  - Most of it US English
  - Yes, also chatGPT :-)
- "Write and article about agriculture"

Non-Universal Aspects

Differences among Languages, effects in NLP (and MT)

- 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? Remember context is important
- 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

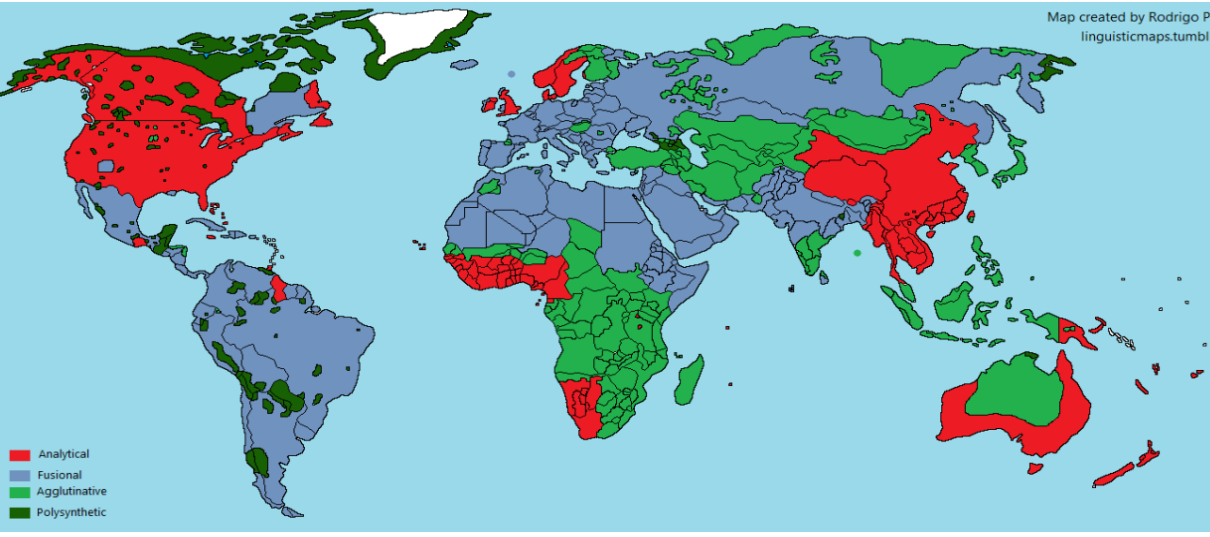
Agglutinative Turkish

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)

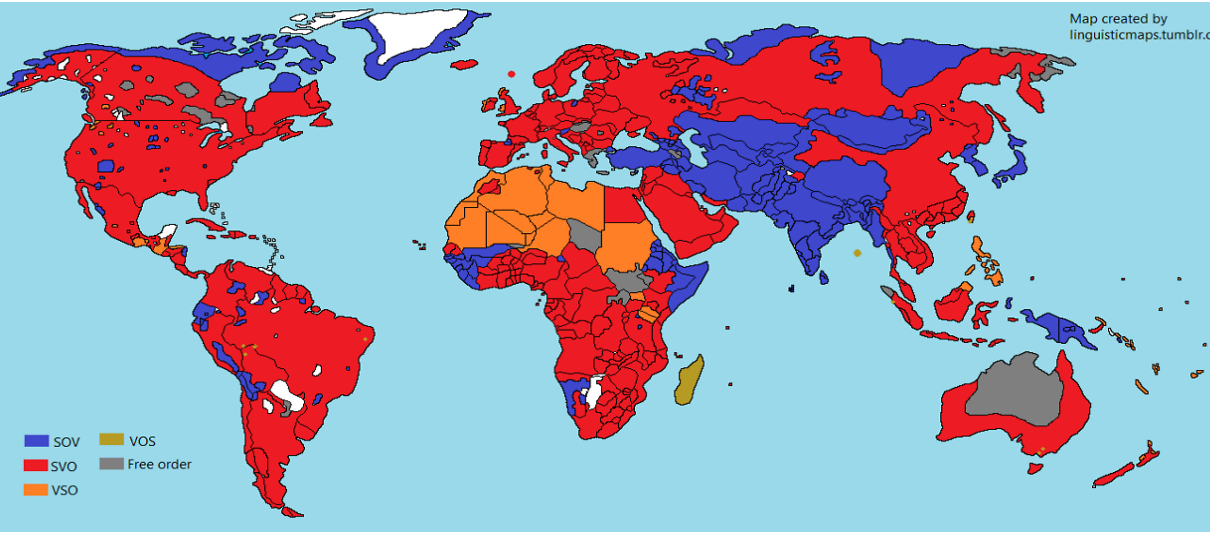
Non-Universal Aspects

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 machine translation systems

Summary

Keep in mind

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

Related keywords for later:

`reordering model`

`rule based`

`interlingua`

`WSD`

`factored model`

`translation model`

`hierarchical`

`vocabulary`

Summary

Wait!

Summary

By the way, I do:

How is German?

And your language?

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