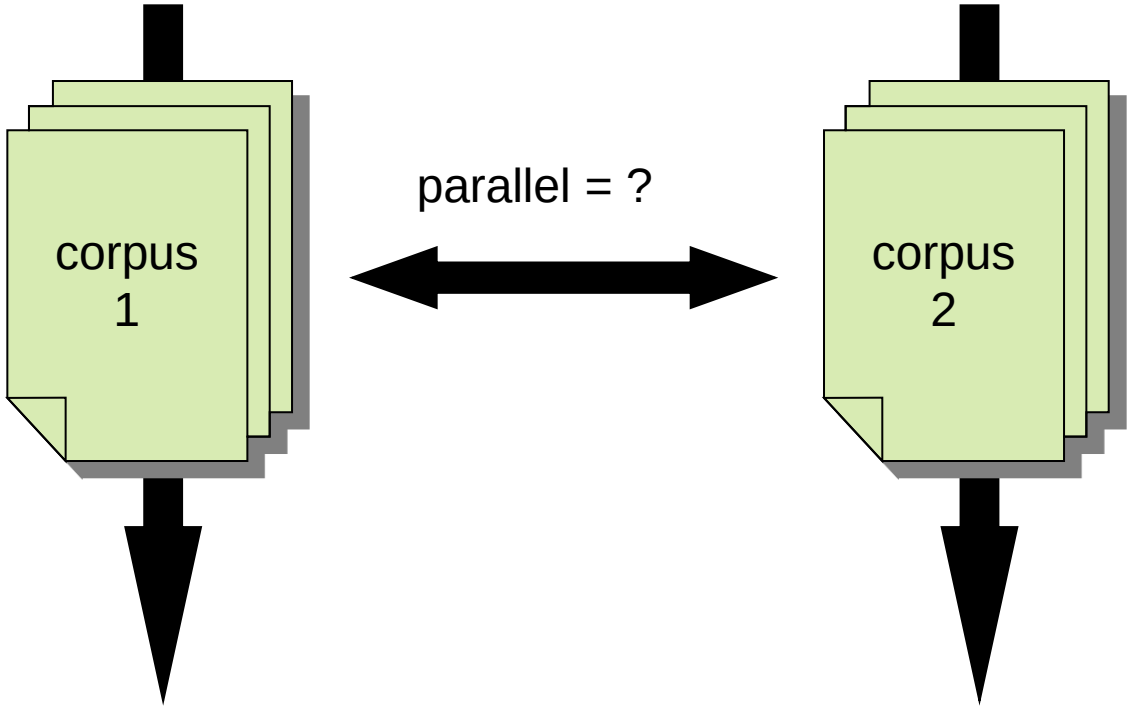
Multilingual and Parallel Corpora

Introduction (ctd.)

Amir Zeldes

amir.zeldes@georgetown.edu

Parallel corpora



What do we mean by parallel?

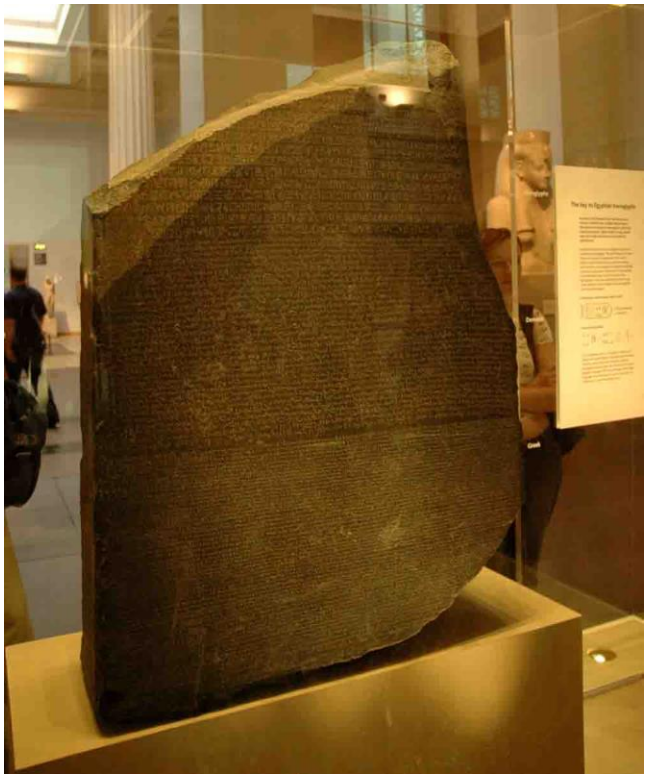
- Parallel corpora: contain (usually **aligned**) translations (bi-texts)
- Comparable corpora: contain comparable but distinct original texts in each language

The Rosetta Stone

- Classic example of a bi-text (or tri-text, depending)
- In itself, not a **corpus**

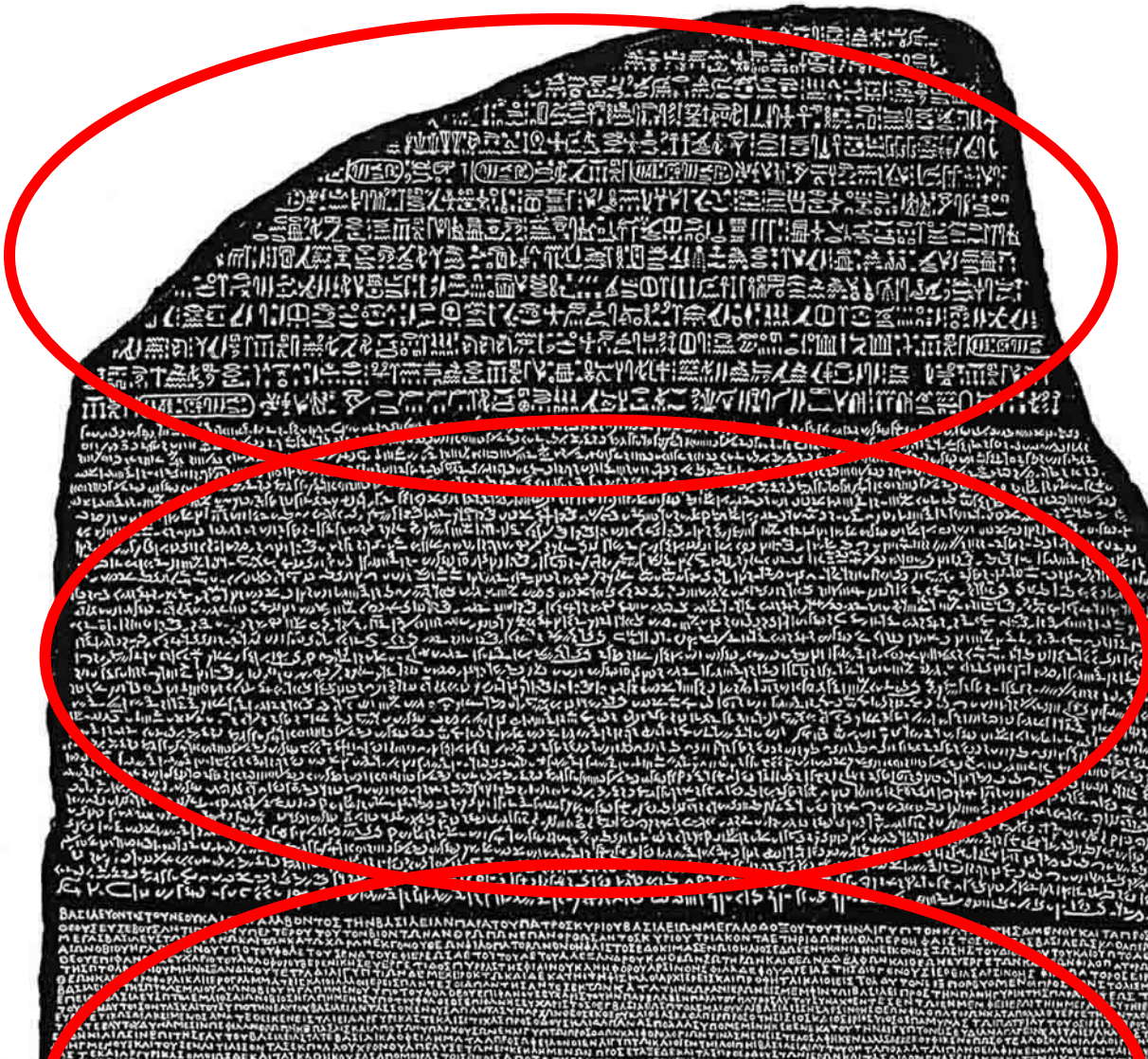
➢ **Text ≠ Corpus!**

➢ A corpus is a sample of a certain language type
➢ What would it take for the Rosetta Stone to be a corpus?
➢ What kind of corpus would it be?



What does a bi-text tell us?

- Multiple instances of the ‘same’ text
- Same content
- What corresponds to what?

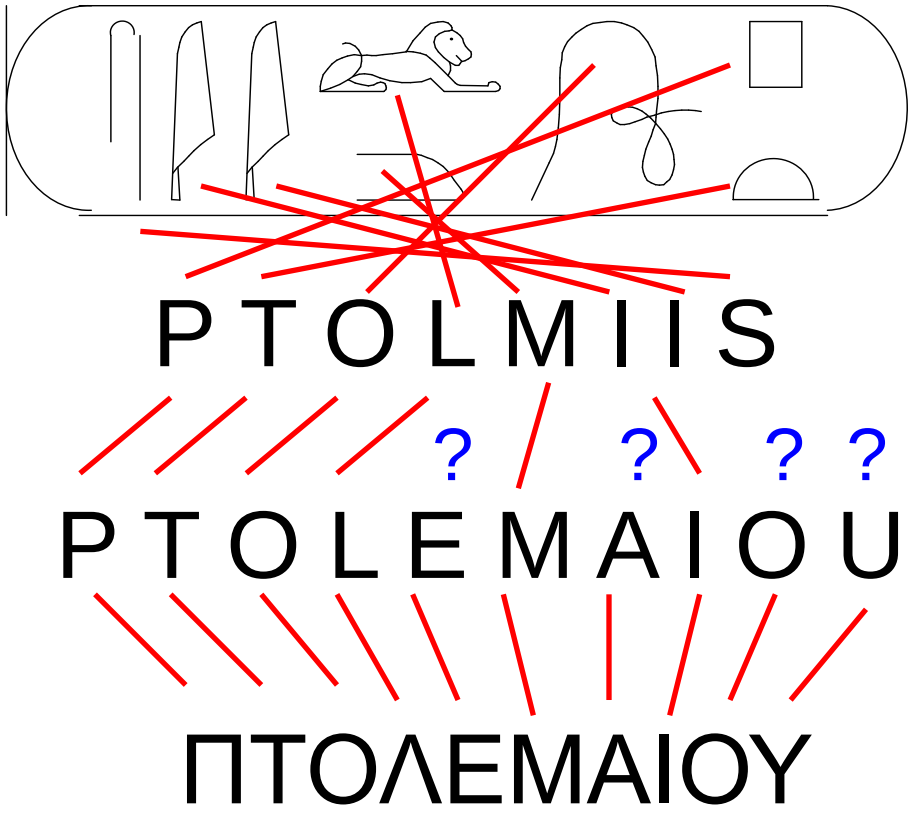


What does a bi-text tell us?

- Multiple instances of the ‘same’ text
- Same content
- What corresponds to what?

What does a bi-text tell us?

- Multiple instances of the ‘same’ text
- Same content
- What corresponds to what?



Parallel and comparable at the same time?

- A corpus can contain many different texts, if this fits our research question (different lengths, genres, etc.)
- And we don’t have to use only texts that we have translations of
- We might often need **parallel** and **comparable** corpora

Translations and reference corpora

- Most “national” corpora (e.g. BNC, ANC, COCA…) contain a variety of text types
  - News
  - Spoken languages
  - Fiction
  - …
- But virtually none contain translations…
- Why?

Translations and reference corpora

- Translations are considered ‘bad’ or ‘unnatural’ English (or German, or…)
- Linguists are interested in “real” language
- Do we need this in a description of the English language?

http://tvtropes.org/pmwiki/pmwiki.php/Main/BlindIdiotTranslation

Counter arguments

- Translations are a legitimate, autonomous 'genre'
  - Also: translations are an autonomous genre **in each language** (Spanish translations)
  - Translations can have **sub-genres** (technical manual translations)
  - Individual **language-pair** genres (English language Chinese restaurant menus)…

Counter arguments

- Translations are a language-independent, naturally occurring interesting phenomenon
  - Regularities and rules should be explored in their own right
  - Essential to bilingual dictionary writing
- Translations are valuable for **comparing** languages
  - What can we learn about differences between Spanish and English from:
    - Can't hold it back anymore : *No puedo ocultarlo más*
  - Or about Dutch and English from:
    - *Zullen wij een sneeuwpop maken…*

Language Comparison

Basic research question type:

- What does language B do in the same situation?
  - How do you say *As if!* In Hebrew?

➢ My guess: *hayita met* “you would die”

➢ What is ‘the same situation’?

Language Comparison

We examine a translation:

Who cares?

Parallel aligned text in the translation =

"the same situation"?

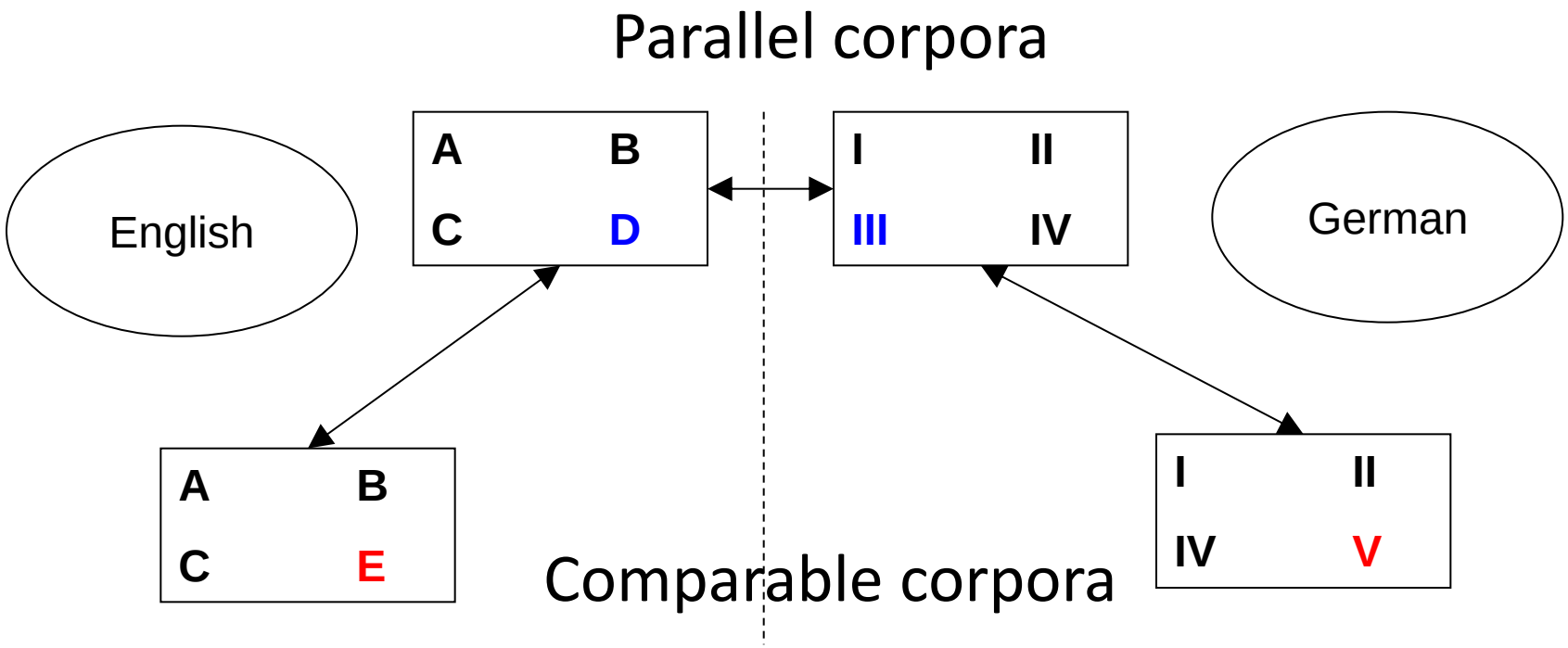
- Assumption: Reference and context independent of linguistic realization

Language Comparison

- Would a speaker of Language B have said that outside of a translation?
- How can we check?

- Comparable corpora are original and independent
  - Does this form appear with the desired meaning 'in the wild'?
  - Why does the corpus need to be comparable?

Mutually complementary



Other applications

- Applied
  - Multilingual lexicography / terminology
  - Machine translation / assistive technology (TMs)
  - Translation pedagogy
- Theoretical
  - Translation studies
  - Comparative linguistics, dialectology
  - Language typology
  - Theoretical and historical linguistics
  - …

Multilingual lexicography

- Parallel corpora allow a direct search: what corresponds to what? How often?
- Example database – real examples of translations
- Are we covering all possible translations
- Especially relevant for terminological dictionaries
  - language for special purposes (e.g. ESP)
  - search by genre/text type
  - detecting variation and standardization (think of the EU, NASA, …)

Exercise

- Can we cover the meanings?
  - How many translations can you think of for …
    - want
    - give
    - take
  - Will translation work both ways?
- Very preliminary check:
  - https://corpling.uis.georgetown.edu/paravoz/

My example – eng:fra

- Want in French:
  - To French (Hound of the Baskervilles)
    - But I **want** to know why the word ' moor ' should have been written ?
    - Mais je **voudrais** bien savoir pourquoi le mot « lande » a été écrit à la main
    - You will tell him that you **want** to see the waste-paper of yesterday .
    - Vous lui direz que vous **voulez** voir les papiers mis hier au rebut
    - We have provided him with all that he can **want**
    - Nous lui avons fourni tout ce dont il **avait besoin**

My example – eng:fra

- Want in French:
  - From French (Three Musketeers)
    - Je n' **ai** pas **besoin** de vous
    - I do not **want** you
    - où , par caprice , par mécontentement ou par **défaut** de fortune , ils avaient endossé
    - which , from caprice , discontent , or **want** of fortune , they had donned

Quantities

- From French
  - voul* -> want: 18/76
  - besoin -> want: 12/76
  - veux -> want: 4/76
  - voudr* -> want: 3/76
- To French
  - want -> besoin: 2/8
  - want -> voudr*: 1/8
  - want -> voul*: 1/8
  - want -> veux: 0

Teaching computers what to do

- How can computers translate natural language automatically?
- How can we teach them what is aligned with what?
- (How do we know this ourselves?)
- What do computers do when there is no example of a similar translation?

➢ Put one of your test sentences into Google Translate

What's good and what not?

- Vous lui direz que vous voulez voir les papiers mis hier au rebut
- Google:
  - *You will tell him you want to see the papers put yesterday scrapped*
- Actual:
  - *You will tell him that you want to see the waste-paper of yesterday .*

For Monday

- Assignment for class (not graded)
  - Look at translations for *want*, *give* or *take* in your language
  - Break down the proportions of each option
  - See if you can get Google Translate to produce all of them using simple sentences
    - What triggers Google Translate's decisions? (try negating, changing tense, adding adverbs, word order …)
    - Are the decisions correct?
    - Do you get any errors?