Multilingual and Parallel Corpora
Introduction

Amir Zeldes

amir.zeldes@georgetown.edu

Organization

- Contact:
  amir.zeldes@georgetown.edu
  Poulton Hall, R. 243
- Office hours:
  Wed 3:30-5:30

Organization

- Use Canvas course site
- Credentials to corpora

More information

- Corpinfo
- GUCL
- UBAG

Requirements

- Participation, readings
- Assignments and presentations
- Final project

What's a parallel?

- Simplest case: same text in another language
- What's special about these?

https://www.youtube.com/watch?v=hYXUSHlhVHw

How about this one?

https://www.youtube.com/watch?v=6OsT0uYgw-Y

The same text in another language?

- Is *Let it Go* difficult to translate? Why?
- What makes translations more or less faithful?
- Are there recurrent themes or trends?

… Only one way to find out!

Introduction – Plan

1. Introductions and languages
2. Background and terminology
   - Corpora and Corpus Linguistics
   - Parallel corpora
   - Alignment preliminaries
3. Overview of some applications
   - Translation studies and related research questions
   - Linguistic typology and comparative linguistics
   - Machine translation and computational lexicography
   - …

Background and terminology

Introductions

- What languages do you speak?
- How did you learn them?
- What languages would you like to look at in this course?
- What would you like to find out?

What is a corpus?

- **Corpora** are digital collections of language data that are collected according to certain criteria
- **Corpus Linguistics** is the methodology dealing with
  - building
  - annotating
  - and evaluating

corpora

A *typical* corpus?

- The composition of a corpus depends on the research questions that you want to answer
  - What kind of corpus do you need in order to study how DC youths speak?
    - *Oh, yeah, yeah, that's my jont."* (LCDC, https://lcdc.georgetown.edu/)
  - To compare the use of relative clauses in spoken language and technical manuals?
    - *It's the life lessons **that** you go through **that** provide you with the wisdom to move forward.* [LCDC]
    - *it provides the man/machine interface **by which** the diving officer enters maneuver instructions and receives information [FROWN]*

Representativeness

- A corpus should be **representative** :
  - Proportions of underlying groups in the corpus should correspond to their proportion in the **population**
  - Factors whose distribution is of no interest should be made explicit
- These are prerequisites for drawing any conclusions from the corpus regarding the population!

Example

- Suppose you want to examine disagreement in university classrooms
- Research question:
  - *How do students disagree with each other and with instructors in college?*
- What does “in college” mean?
  - Proportion of course subjects corresponds to distribution in course catalogue
  - Speaker gender: not relevant (this is an assumption!)

Example

- A corpus of university classroom interaction: **MICASE** (Michigan Corpus of Academic Spoken English)

S11: yeah, i mean i i did think about that. and i guess that was my way of trying to isolate the studio. i don't want random people walking through the studio

S7: well i i appreciate, no i t- i'm i don't **disagree** with that. i want both things. [S11: yeah] i wanna somehow an answer to my first question, [S11: yeah] and

S11: yeah

S3: Speaker information restricted

S7: over, on the other side (is what you're saying.)

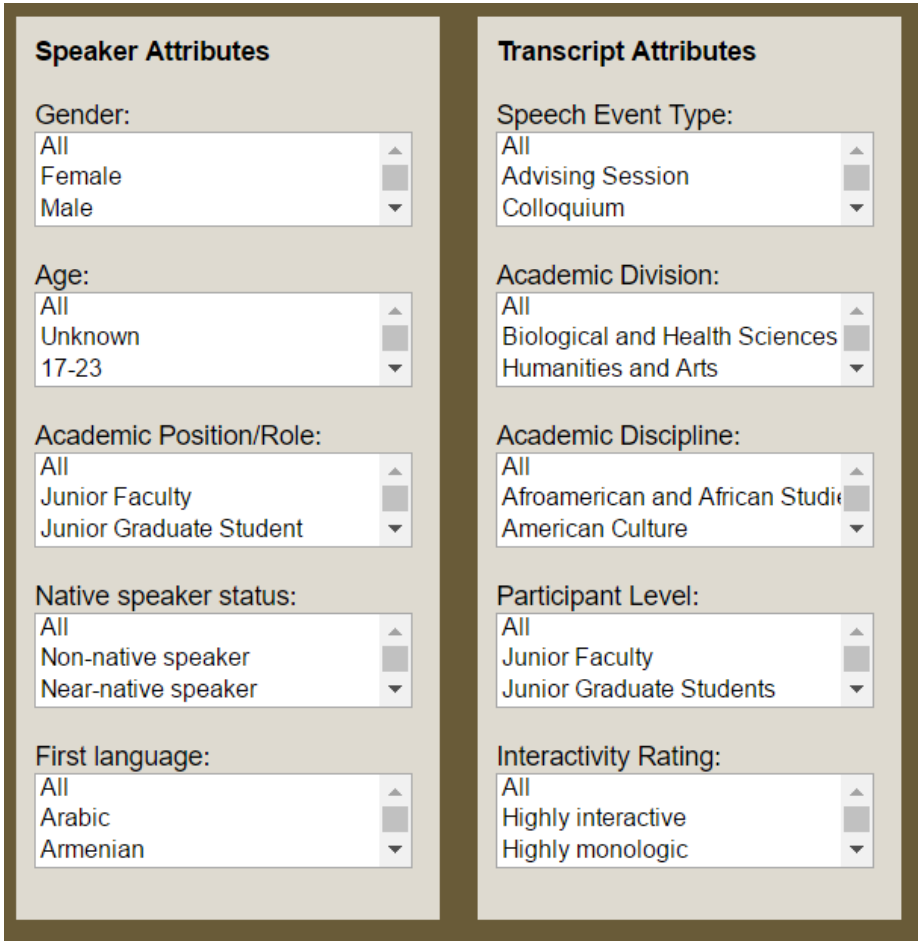
S3: Speaker information restricted

S11: well i think if you were the ty- you were the kind of person

Humanities / Architecture Critiques

Different research questions

- Suppose our research question is:
  - Do male and female students speak differently in classroom interaction?
  - If so, in what subjects?
- Now we need comparable amounts of female and male student data



Balance

- Corpora aim to be **balanced**:
  - All sub-groups should have adequate attestation
  - The decision which sub-groups to include is again based on assumptions!

- Balance is often at odds with representativeness

Example

- Question: “do CS majors differ more or less across gender than English majors?”
  - Balance: same amount of male/female speakers
  - Representativeness: the distribution of genders is not homogeneous across these genders!

Interim conclusions

- A corpus is chosen based on the research question or application that you have in mind
  - We can't always get what we want in practice
  - But we can optimize our choices with this in mind
- Assumptions should be made explicit
  - What's important?
  - What could we be missing?

What does parallel mean?

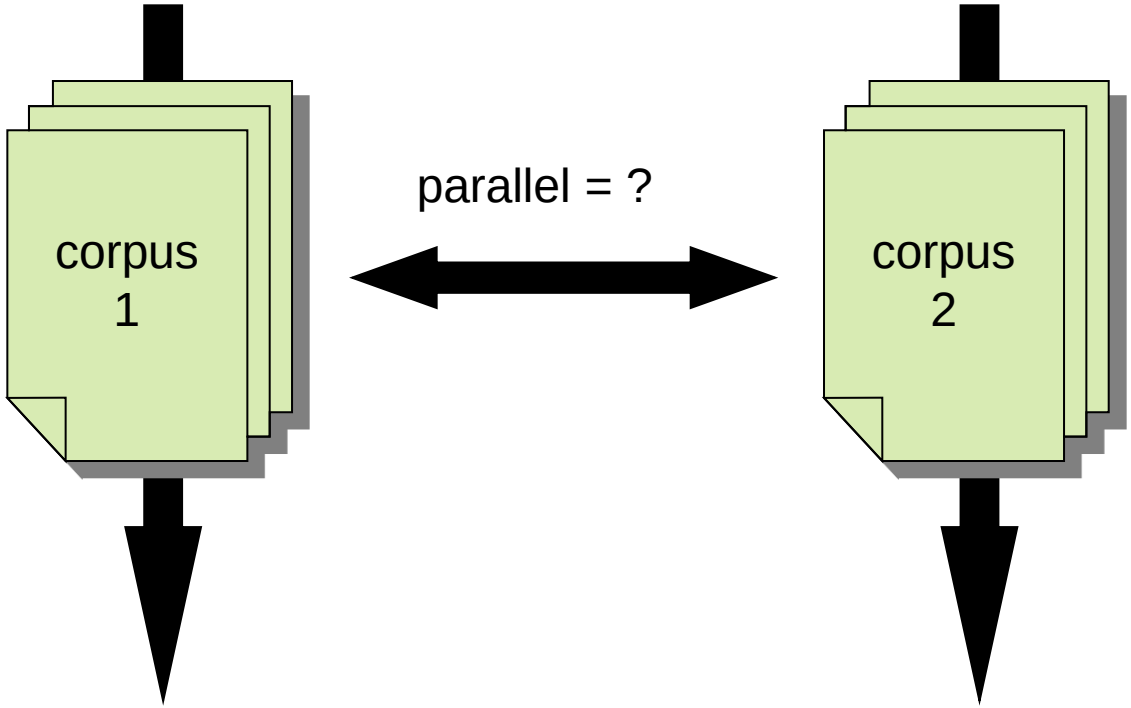
- OED: Adjective …
  - **II.** Senses relating to correspondence.
    - **7. a.** Close correspondence or analogy; a point of comparison or similarity between two people or things. Hence also: an act of drawing such correspondence or analogy; the placing of things side by side mentally or descriptively so as to show their similarity. Freq. ***to draw a parallel***.
    - 1599 *Master Broughtons Lett. Answered* vii. 22, *I craue pardon of his Grace for abasing him in paralell with such an one as thou art.*

Parallel corpora

- .. are special **corpora** containing data from multiple languages* in **parallel**
- constructed especially for corpus and computational linguistics, often out of pre-existing translations
- used in a variety of applications, including the creation of machine translation and translation memory systems

* Open question: what constitutes a language?

Parallel corpora



What do we mean by parallel?

- Some researchers use the term “parallel corpus” for all multilingual corpora and distinguish:
  - **Translation corpora –** corpora containing only translations of the same texts (each language has exactly the same content, a collection of **bi-texts**)
  - **Comparable corpus –** contains independent but comparable texts about the same topics in the same quantities

What do we mean by parallel?

- Currently the more common terminology is probably:
  - Parallel corpora: contain (usually **aligned**) translations (bi-texts)
  - Comparable corpora: contain comparable but distinct original texts in each language

For next time

- Please send me your #1 preferred **language pair** (you may also work on others later, but I will prioritize getting data for these)
- You can start reading Aijmer (2008): Parallel and Comparable Corpora (please finish by Monday, January 23)