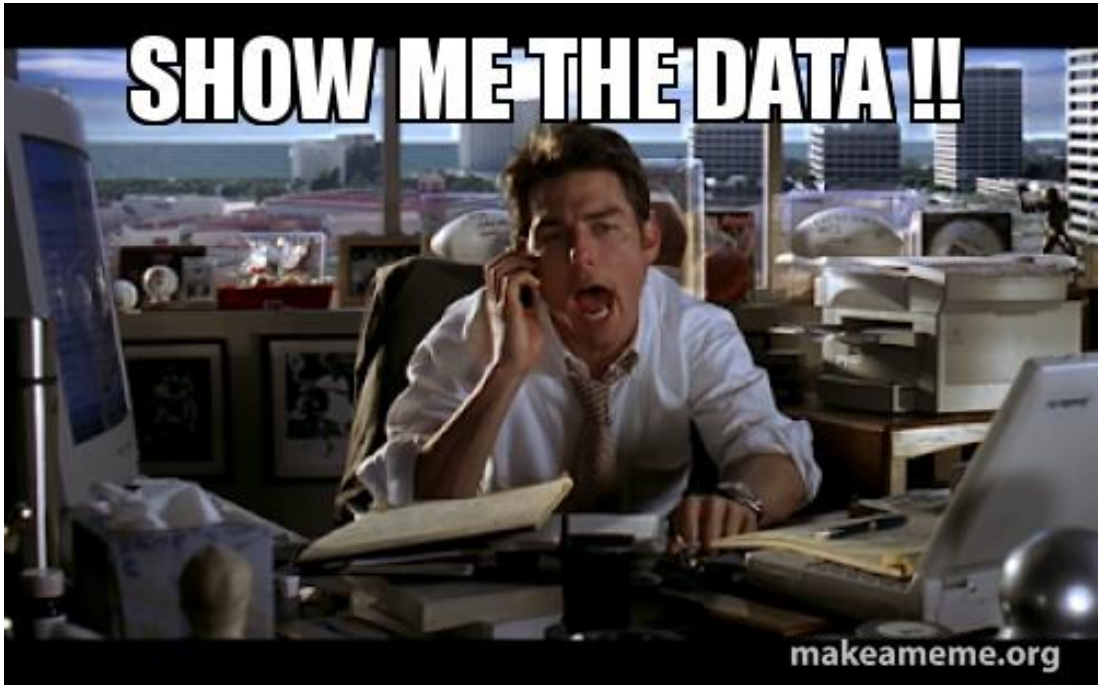
The datasets



The datasets

- 10+ sets for the entire course (and the exam!)
- (And the text book's data as "gries_data")
- Real-world data from research in the Humanities
- Mostly "small", but hopefully both relevant, diverse and exciting
- Many of them collected and analysed at our university!
- They'll come back around: gradually more complex analyses

Datasets folder

Subfolder per dataset, containing:

- Dataset itself as a table (.csv, .xsl, .tsv, .xlsx, ...)
- “README”: detailed plain-text description of the dataset’s contents, variables and variable levels, publications based on the data, ...

→ **Familiarize yourself with some of the datasets through the homework exercises**

Datasets: Overview

1. Arthur
2. Chat
3. Chat_repeated
4. Correlaciones
5. Fanfiction
6. Federalist
7. Queer
8. Social
9. Spelling
10. (Gries’s data)
11. And more

1. Arthur

- Origin: http://www.arthurianfiction.org/
- Database of medieval works from Arthurian narratives and the **books** (manuscripts) in which they survive
  - Multiple regions/languages from medieval Europe
  - Detailed information on dimensions of the books:
    - How large (page and text area)?
    - How many illustrations, columns, pages?
    - Paper or parchment?
    - Date (approximation)?
    - From which region do they originate?

BL, Additional 10292, f. 100

2. Chat

- **Sociolinguistic** dataset
- Collected and analyzed at CLiPS (Hilte, Vandekerckhove, Daelemans)

- Summary of **Flemish teenagers' chat conversations** (WhatsApp, Facebook):
  - Socio-demographic info: the teenagers' "profile"
    - Gender, educational track
  - Linguistic info: number of chatspeak features in the text
    - Emoticons, non-standard Dutch words

2. Chat

nr_tokens	subject_ID	gender	education	emoticons	nonstd_Dutch
1487	1	male	general	49	292
1859	2	male	technical	49	298
178	3	male	general	0	32
258	4	female	general	0	53
48	5	male	vocational	0	3
314	6	female	vocational	1	43

3. Chat_repeated

- Sociolinguistic dataset, **larger version of "Chat"**
- Collected and analyzed at CLiPS (Hilte, Vandekerckhove, Daelemans)

- Summary of **Flemish teenagers' chat conversations** (WhatsApp, Facebook):
  - Socio-demographic info: the teenagers' "profile"
    - Gender, educational track, **age category**
  - Linguistic info: number of chatspeak features in the data
    - Emoticons, non-standard Dutch words

ADDITION: contains "repeated measurements"
(= multiple observations for 1 subject - see later)

3. Chat_repeated

nr_tokens	subject_ID	gender	education	age_cat	emoticons	nonstd_Dutch
502	1499	female	vocational	young_teen	50	92
1470	2211	female	vocational	young_teen	2855	492
1536	website 658	male	vocational	young_teen	231	222
676	123	male	vocational	young_teen	0	124
5494	1327	male	general	young_teen	365	827
219	website 689	male	vocational	young_teen	2	11

4. Correlaciones

- Dataset with metadata on well-known Spanish Literature form the Silver Age
- Collected by Dr. José Calvo Tello (morethanbooks.eu) for PhD
- Information on authors:
  - Life span
  - Gender
  - How many works/novels published
  - Etc.
- Information on works:
  - When published?
  - How long is the story (in days)?
  - How “canonical” is it?
  - Etc.

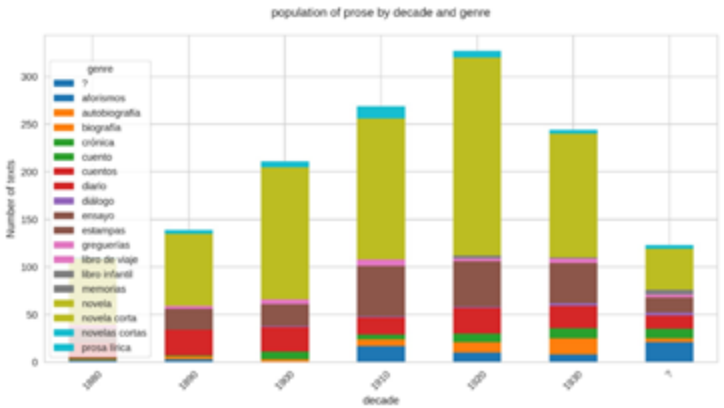


Figure 12: Number of texts over decades differentiating genres

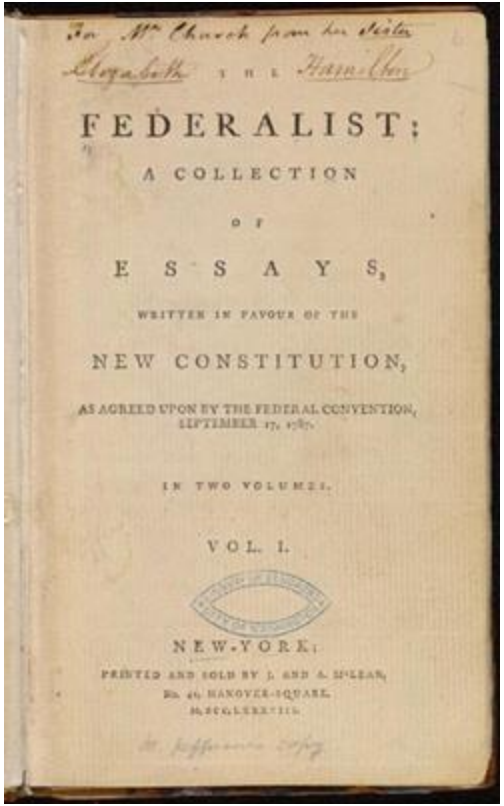
"The Novel in the Spanish Silver Age: A Digital Analysis of Genre through Machine Learning" [2020]

5. Fanfiction

- “Fan fiction” or “transformative literature”
- Fiction by non-professional authors inspired by cultural “fandoms”, such as J.K. Rowling or Sherlock Holmes
- Huge (!) and global phenomenon
- Dataset with (user-provided) metadata on “fics”
  - Intended audience?
  - When published/finalized?
  - Which fandom?
  - Which characters?
  - How many “likes” and “views”?
- (Warning: often “explicit” in nature...)

6. Federalist

- Federalist papers: collection of papers (ca. 1788)
- Promoting the ratification of US constitution
- Very famous in US history, but anonymous
- All published under pseudonym “Publius”
- Famous candidates: Hamilton? Jay? Madison?
- “Bag-of-words” table with word frequencies in each essay
  - Each row = essay
  - Each column = word
  - Each cell = absolute frequency of word in text
- Statistical authorship attribution



7. Queer

- Famous question from sociophonetics (mentioned in Gries)
- “Is there a difference in **pitch** between speech homosexual/heterosexual men?”
- Dataset from recent paper (Suire et al. 2020)
- Various pitch measurements and sociological variables for set of Spanish speakers

Archive of Sexual Behavior (2020) 49:2575–2588
https://doi.org/10.1007/s10508-020-01665-3

ORIGINAL PAPER

Speech Acoustic Features: A Comparison of Gay Men, Heterosexual Men, and Heterosexual Women

Alexandre Suire[1] · Arnaud Tognetti[2,3] · Valérie Durand[1] · Michel Raymond[2] · Melissa Barkat-Defradas[1]

Received: 7 September 2018 / Revised: 14 February 2020 / Accepted: 18 February 2020 / Published online: 31 March 2020

Abstract
Potential differences between homosexual and heterosexual men have been studied on a diverse set of social and biological traits. Regarding acoustic features of speech, researchers have hypothesized a feminization of such characteristics in homosexual men, but previous investigations have so far produced mixed results. Moreover, most studies have been conducted with English-speaking populations, which calls for further cross-linguistic examinations. Lastly, no studies investigated so far the potential role of testosterone in the association between sexual orientation and speech acoustic features. To fill these gaps, we explored potential differences in acoustic features of speech between homosexual and heterosexual native French men and investigated whether the former showed a trend toward feminization by comparing theirs to that of heterosexual native French women. Lastly, we examined whether testosterone levels mediated the association between speech acoustic features and sexual orientation. We studied four sexually dimorphic acoustic features relevant for the qualification of feminine versus masculine voices: the fundamental frequency, its modulation, and two understudied acoustic features of speech, the harmonics-to-noise ratio (a proxy of vocal breathiness) and the jitter (a proxy of vocal roughness). Results showed that homosexual men displayed significantly higher pitch modulation patterns and less breathy voices compared to heterosexual men, with values shifted toward those of heterosexual women. Lastly, testosterone levels did not influence any of the investigated acoustic features. Combined with the literature conducted in other languages, our findings bring new support for the feminization hypothesis and suggest that the feminization of some acoustic features could be shared across languages.

Keywords Speech · Voice · Acoustics · Sexual orientation · Testosterone levels · Gender atypicality

Introduction

The gender atypicality hypothesis suggests that gender atypical traits in homosexuals could be used as cues to indicate sexual orientation. Differences between heterosexual and homosexual individuals have thus been studied on a diverse set of traits such as face (e.g., Freeman, Johnson, Ambady, & Rule, 2010; González-Álvarez, 2017; Lyons, Lynch, Brewer, & Bruno, 2014; Rieger, Linsenmeier, Gygax, Garcia, & Bailey, 2010; Skorska, Geniole, Vryzas, McCormick, & Bogaert, 2015; Wang & Kosinski, 2018), olfaction (e.g., Sergeant, Dickins, Davies, & Griffiths, 2007), behavior (e.g., Ambady, Hallahan, & Conner, 1999; Rieger, Linsenmeier, Gygax, & Bailey, 2008; Valentova, Rieger, Havlicek, Linsenmeier, & Bailey, 2011), cognition (e.g., Neave, Menaged, & Weightman, 1999; Xu, Norton, & Rahman, 2017), and voice (e.g., Gaudio, 1994; Munson, McDonald, DeBoe, & White, 2006b; Pierrehumbert, Bent, Munson, Bradlow, & Bailey, 2004; Rendall, Vasey, & McKenzie, 2008). In addition to the fact that homosexuals exhibit traits that differ from those of heterosexuals, it has been shown that some of them, such as specific neural processes (LeVay, 1991; Savic, Berglund, & Lindström, 2005) or specific childhood behaviors (Alanko et al., 2010; Bailey & Zucker, 1995), displayed values shifted toward those of the opposite sex, i.e., a feminization in homosexual men and a masculinization in homosexual women (Pierrehumbert et al., 2004). Moreover, studies have shown that both men and women are able to accurately assess sexual orientation from both sexes from various features such as the face or body movements (Ambady

Arnaud Tognetti
arnaud.tognetti@gmail.com

[1] CNRS, IRD, EPHE, Institut des Sciences de l'Evolution, University of Montpellier, Montpellier, France

[2] Institute for Advanced Study in Toulouse, 21 Allée de Brienne, 31015 Toulouse, France

[3] Department of Clinical Neuroscience, Karolinska Institutet, Stockholm, Sweden

8. Social

- **Sociological** dataset
- Collected and analyzed at CLiPS (Hilte, Vandekerckhove, Daelemans)

- **Flemish teenagers' socio-demographic and social class profiles**
  - Gender
  - Educational track in secondary school
  - Parents' socio-economic class
  - Home language

8. Social

subject_ID	gender	education	language	parents_class
1	female	technical	Dutch_only	1
2	male	general	Dutch_only	1
3	female	general	Dutch_only	1
4	female	general	Dutch_only	1
5	female	technical	Dutch_only	1
6	male	vocational	Dutch_only	1

9. Spelling

- **Sociolinguistic** and **psycholinguistic** dataset
- Collected and analyzed at CLiPS (Surkyn, Vandekerckhove, Sandra)

- Summary of **verb spelling errors** in **Flemish teenagers' chat conversations** (WhatsApp, Facebook):
  - Socio-demographic info: the teenagers' "profile"
    - Gender, age, educational track, grade
  - Linguistic info: verb spelling errors
    - Verb lemma, error/no-error

  Contains "repeated measurements"
  (= multiple observations for 1 subject and for 1 lemma - see later)

9. Spelling

subject_ID	lemma	gender	age	education	grade	error
325	aanbidden	male	17	general	Gr3	0
563	aanbidden	female	18	technical	Gr3	0
742	aanbieden	male	18	vocational	Gr3	0
899	aanbieden	male	17	vocational	Gr3	0
625	aanbieden	female	16	general	Gr2	0
1277	aanbieden	female	18	general	Gr3	0