

# Assignment – in the small groups and for the next time

- Creation of a small corpus in your mother tongue:
  - 3 sentences with [i, a, u] in your mother tongue plus [uRu] [iRi] and [aRa]  
example in French « Le loup arriva sur le lit de URU. »
  - Design the corpus
  - Record it in the supine, sitting and standing position.
  - Phonetic annotation (by hand or forced alignment followed by corrections by hand)
  - Exploitation via a Python program or a Praat script (compute formants, duration of closure... or any other relevant phonetic feature)
  - Statistical test (check that the formant values are different for the 3 postures ...)
  - Report due to May 25

# Corpora for Speech Copora pour la parole

Yves Laprie  
CNRS / Université de Lorraine  
[yves.laprie@loria.fr](mailto:yves.laprie@loria.fr)

# Before anything else

- To build a corpus or not?
- Build means:
  - Design the corpus
  - Design the setup
  - Check the ethical and legal aspects
  - Find and choose subjects
  - Record the corpus
  - Annotate the corpus and monitor the annotation task
  - Save the corpus
- And finally exploit the corpus
  - Disseminate it
- **Think about all these points before starting the adventure**

# General point of view about corpora

- Databases in speech processing are called corpus (and corpora in the plural).
  - First expected behavior:
    - Constructing a corpus is expensive and the work is generally dramatically underestimated → Reuse existing corpora
    - Explore catalogs of corpora at LDC Linguistic Data Consortium (<https://www ldc upenn edu/>) or ELDA-ELRA European Language Resources - Evaluations and Language resources Distribution Agency (<http://www elra info>)
    - A number of national web sites collecting corpora (Clarin-D for German, Ortolang...) or list of corpora.
- Explore and contact authors

# What makes the price of corpora

- Price: from 0 to more than 10 000 € depending on the nature, size and the status of the buyer
- Should be related to the efforts spent to collect and annotate the corpus
- Raw data cannot be used without annotations (orthographic, phonetic, prosodic, geometric...):
  - Good annotations → enables exploitation of the corpus

# What kind of data about speech?

- Acoustic signal (recorded with one or several microphones)
- Visual data: speaker's face geometry
- Articulatory gestures (many acquisition techniques more or less invasive)
- Vocal folds (several acquisitions more or less invasive)
- Aero-acoustic parameters (pressure at several points in the vocal tract or below the glottis, airflow)

# Recording

- Either the internal microphone of your computer
- Or:
  - an external audio card for PC (Presonus AudioBox USB 96 25th Anniv Ed, M-Audio Fast Track USB mk2, ...)
  - a good microphone AKG C-520 (do not forget to use the phantom alimentation!)

# What kind of data about speech?

- Acoustic signal (recorded with a microphone)
  - **Praat, Transcriber 1.5.1**  
(<https://transcriber.en.uptodown.com/windows/download>)
- Visual data: speaker's face geometry
- Articulatory gestures (many acquisition techniques more or less invasive)
  - **Visartico, MITK, ItkSnape**
- Vocal folds (several acquisitions more or less invasive)
- Aero-acoustic parameters (pressure at several points in the vocal tract or below the glottis, airflow)



# Annotation software

Some examples:

- Xtrans speech signal → orthographic annotations with possibly several speakers
- Force alignment tools:
  - French <http://astali.loria.fr>
  - Others. Look at <https://github.com/pettarin/forced-alignment-tools>
  - <https://clarin.phonetik.uni-muenchen.de/BASWebServices/interface/WebMAUSBasic>
  - Montreal Forced Aligner  
[https://montreal-forced-aligner.readthedocs.io/en/latest/getting\\_started.html](https://montreal-forced-aligner.readthedocs.io/en/latest/getting_started.html)  
<https://mfa-models.readthedocs.io/en/latest/dictionary/index.html#dictionary>
- Praat → a standard in phonetics (but require to be aware of signal analysis limits)
- Visartico → electromagnetographic data
- MITK, ItkSnape → MRI data

# General Data Protection Regulation (GDPR)

- Règlement général sur la protection des données (RGPD)
- Linked to personal data: “[A]n identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person.”
- Principles of the GDPR ([https://ec.europa.eu/info/law/law-topic/data-protection/reform/rules-business-and-organisations/principles-gdpr\\_en](https://ec.europa.eu/info/law/law-topic/data-protection/reform/rules-business-and-organisations/principles-gdpr_en))
  - explain in clear and plain language why you need the data, how you’ll be using it, and how long you intend to keep it.
  - indicate those purposes to individuals when collecting their personal data
  - use of coding techniques (e.g. pseudonymisation, cryptography or anonymisation technics), the use of protected servers against external threats
  - this concerns primarily trade, but science as well

# GPRD in sciences

- How is consent for processing in scientific research obtained?
- [https://ec.europa.eu/info/law/law-topic/data-protection/reform/rules-business-and-organisations/legal-grounds-processing-data/grounds-processing/how-consent-processing-scientific-research-obtained\\_en](https://ec.europa.eu/info/law/law-topic/data-protection/reform/rules-business-and-organisations/legal-grounds-processing-data/grounds-processing/how-consent-processing-scientific-research-obtained_en)

“Some flexibility in relation to the degree of specification and granularity of consent is allowed in the context of scientific research. When collecting personal data, researchers might not be able to fully identify the purposes for processing. In those cases they can ask individuals to give consent for certain areas of scientific research or parts of research projects. In any case, consent must keep its core elements, that is to say it must be freely given, informed, sought via clear affirmative action and specific to the extent allowed by the research in question. Researchers must make sure they also comply with the ethical and methodological standards required in their field.”

# Ethical issues

- In speech (and in many other domains) a key question is how invasive a acquisition technique is.
- Invasive is linked to:
  - The device is inside the speaker's body (fibroscopy for instance), or its physical principle involves consequences inside the speaker's body (X-ray for instance).
  - The device can indirectly alter the speaker's health because it has not been properly sanitized.
- As soon as a medical technique is used an ethical agreement is mandatory:
  - In France CPP ("Comité de Protection des Personnes", projects evaluated at the national level)

# Crowdsourcing and corpora

- A collaborative internet approach that can be used for the annotation of corpora.
- The process of distributing tasks to an open, unspecified population via the internet.
- Several solutions: games with a purpose, volunteer-based platforms, paid-for marketplaces such as Amazon Mechanical Turk (AMT) and CrowdFlower (CF)
- Four steps:
  1. Project definition (design the objective and the crowdsourcing task)
  2. Data preparation (Collect and pre-process corpus, Build and/or reuse annotator management interfaces, Run pilot studies)
  3. Project execution (Recruit contributors, Train and select them, Monitor crowdsourcing work)
  4. Data evaluation and aggregation (Evaluate and aggregate annotations and guarantee the overall consistency)

# Crowdsourcing and corpora

- Tasks must be sufficiently simple and intuitive
  - Do not overestimate annotators or recruit appropriate people (with training in the target area)
- Split the work in smaller parts... but risk of inconsistencies
- Quality question: how the quality can be guaranteed?
  - At least it is necessary to include a set of data (annotated by experts) common to all people involved and discard their annotations if they are far from the expected level.
- Determine how much to award
- Legal and ethical questions
  - how to protect data which are annotated;
  - how to properly acknowledge contributions;
  - how to ensure contributor privacy and wellbeing;
  - and how to deal with consent and licensing issues



# Preparation of a spoken corpus (1/5)

## General questions:

- Source of speech:
  - Radio (simple but copyrights and preparation to remove non speech signal)
  - Audio book (a good speaker for a reasonably long duration)
  - Archived speech (longitudinal control of a speaker)
  - Recorded speech
- Annotations:
  - Manual or automatic (→ impact on the quality)
  - Orthographic annotation (→ automatic speech recognition)
  - Phonetic annotations : sound segmentation, prosody, pronunciation errors → enables fine exploitation but requires experts)
  - Other annotations (articulatory, geometrical...)

# Preparation of a spoken corpus (2/5)

## General questions:

- Type of microphones
  - One microphone (directive or non directive)
  - Several microphones (record environment + one or several speakers)
  - Internal laptop microphone
  - Acquisition card or not (to get a more control sampling frequency and quality)
- Quality of the speech signal:
  - from an anechoic room  to
  - ... quiet room with non flat walls, school room,
  - ... smartphone in a noisy place, 
  - ... a cocktail party noisy room
  - ... IP voice (skype...)



# Preparation of a spoken corpus (3/5)

## General questions:

- number of speakers → linked to speaker variability. The more the better but the more also means more annotations
- Kind of speakers
  - children ... to elderly people
  - native versus non native speakers
  - normal or pathological voices
  - professional speakers (actor, journalist) or not
- Speech style:
  - Spontaneous speech (no control at all)
  - Elicited speech (to guide speakers)
  - Read speech (requires some software for presenting sentences and recording)
  - Read speech with imposed phonetics (synthesis)
  - Pseudo words (to study some precise phonetic issue)

# Preparation of a spoken corpus (4/5)

- Phonetically oriented questions:
  - Carrier sentence (waste time), list of words (list effect on the last word)
  - Pay attention to the boundaries (lexical or prosody) which may alter the expected phenomenon.
- Onion strategy when the acquisition technique reliability is not guaranteed (unexpected sensor removal, MRI session too short, technical problem...):
  - Incorporate the possibility of failure within the design of the corpus
  - One first set of data corresponding to the very heart of the corpus which enables a first level of exploitation
  - An additional layer which covers the whole objective.

# Preparation of a spoken corpus (5/5)

General questions:

- Sampling frequency and compression:
  - 22050 Hz is a good compromise
  - 16000 Hz as well
  - Less means that fricatives frication noise is lost.
- Compression (MP3 for instance) or not? : **No compression because it partially destroys the signal.**

# Final devoicing of stops in languages

- Some examples:
  - French: *bague* / *bac* (ring, tray or tank) /bag/ vs. /bak/
  - English: *bag* : *bag* / *back* /bæg/ vs. /bæk/
  - German *Bad* /ba:t/, *Zug* /tʃu:k/
  - Russian...
- Which parameters or tools could be used to analyze devoicing?
  - F0, segmentation (duration of vowels and fricatives)...
  - Perception experiments (ask native speakers)
  - Transformations of speech signals (and then perception)
  - Articulatory parameters
  - Vocal folds parameters

- Literature review
- Which corpora?
  - Spontaneous speech
  - Controlled speech (with a carrier sentence)
  - EPGG or articulatory data
- Pilot study:
  - Record examples (simple contrasts, several languages)
  - Test perception (several languages)
  - Design a small corpus
  - Annotation with Praat
  - Explore parameters

# Corpus organization

**Principle: do not test the system on data used for training**

**Strategy → split the corpus in three parts:**

- **Training.** The biggest part for training the models (DNN, HMM...)
  - **Development.** A smaller part (10 times smaller) for adjusting weights between the different components (acoustic, language...)
  - **Test.** The smallest part intended to evaluate the system.
- Example in hours : 150, 10, 5
- Assumption: the corpus is big enough

# Compensating for insufficient corpus sizes

## K-fold Cross-Validation

- Split the corpus in K parts (so as to not run the test on data used for training).
- For  $k = 1$  to the K parts do
  - train models on the corpus minus part k
  - evaluation on part k
- endfor
- compute the average evaluation

In the worst case each part can be a corpus item (a sentence for instance)

Increase a lot the evaluation task!

# Even smaller corpus

- When the acquisition technique imposes constraints on the subjects, the corpus may be limited to a only few subjects.
- One subject: no risk, or at least the limit is clear
- Two subjects: do not say the the subjects are very different! (2 points are always very different)
- Above two: use statistical tests before claiming anything.



# Website for creating corpora

- **Recording a corpus** JCorpusRecorder :  
<https://members.loria.fr/VColotte/j-corpus-recorder/>
- **Creating an open speech recognition dataset for (almost) any language:** <https://medium.com/@k Clintcho/creating-an-open-speech-recognition-dataset-for-almost-any-language-c532fb2bc0cf>
- Tools: aeneas
  - <https://www.readbeyond.it/aeneas/>
  - Requires python 2.7, ffmpeg, espeak
- A web site for corpus construction <https://www.phonetik.uni-muenchen.de/forschung/BITS/TP1/Cookbook/>
- Processing corpora with Praat:  
<http://www.praatscriptingtutorial.com/filesExtendedExample>

01101100

01101111

01110010

01101001

01100001

01101100

01101111

01110010

011010010...

111000010110

11100100110

1000010110

1111110

01101100  
01101111  
0110010  
01101001  
01100001  
01101100  
01101111  
0110010  
01101001  
1100001011  
1100100111  
1000101111  
111111

Loria

Laboratoire lorrain de recherche  
en informatique et ses applications

Exploiting corpora

# Scripting Praat for exploiting corpora

- A scripting language is involved in Praat but it is not very to use (debugging is not easy).
- Many tutorials exist:
  - <https://eleanorchodroff.com/tutorial/PraatScripting.pdf>
  - <https://praatscripting.lingphon.net/environments-2.html>
- A big repository with existing scripts
  - <http://phonetics.linguistics.ucla.edu/facilities/acoustic/praat.html>

# Developing scripts in Praat

- Several solutions:
  - The “lazy” but reasonable and efficient solution: reuse an existing script and modify it if need be.
  - Minimize the development of scripts inside Praat. Use scripts only to get measurements and then use Python or any other language.
  - Use “history” to save simple pieces of scripts.
  - Develop a new script.

# A very small survey of Praat scripting

- Principle:

- “picking up” results from Praat → **All** the objects and functions can be called with **exactly** the same name as in the Praat menus.

Example listing the wav files in a folder

```
resultList = Create Strings as file list... "resultList", "C:\Users\ylaprie\Corpus\*.wav"
```

- a minimal language to exploit those results
  - scalars (starting with a small letter) : `frequency = 16000.`
  - strings (starting with a small letter and finishing with \$): `filename$ = "record.wav"`
- Control structures:
  - `if then else endif` (= for testing equality)
  - `for endfor`
  - `procedure OneProcedureName endproc`
- Some important commands
  - writing `writeInfoLine: "the result is", blablaVariable`
  - writing in a file `fileappend 'resultfile$' 'resultline$'`

# One example with your files

- Display F1 and F2 in F1-F2 plane to check the impact of the position (standing, sitting or supine).
- The “lazy” solution: reuse an existing script (don’t remember where I found it).
- Before running the script change:

#change to set the folder you want

```
directory$ = "C:\Users\ylaprie\Documents\Corpus\Corpus3Positions"
```

....

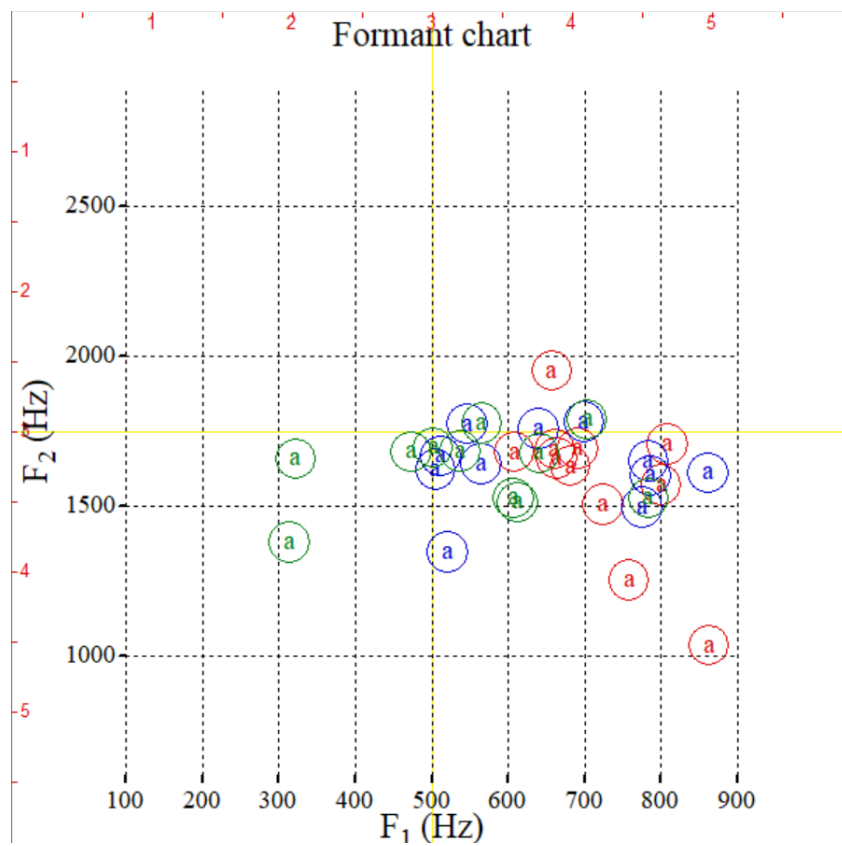
```
textgridFilename$ = directory$ + "/" + basename$ + ".textgrid" --> check whether this is  
TextGrid or texgrid for your files
```

...

```
#by default "a" because it should exist in all the languages and because it is a vowel  
segment_label$ = "a"
```

```
#the first tier (1) but it could be another one (here it is 2 with the examples).  
tierNumber = 2
```

# F1 F2 for the three positions



# Testing two populations

Here, have the two series of formants for the sitting and the supine positions the same mean value for formant F1?

→ test t (Student)

- With several possibilities
  1. The same vowels in two positions (paired values since the sentences are the same)
  2. Two independent sets of vowels (for instance two texts)

The hypothesis  $H_0$  is that the two series have the same mean value.

If the test value is outside the  $[-t, t]$  provided by the student table for a given risk,  $H_0$  is rejected.

See:

- With Python

[https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest\\_ind.html](https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest_ind.html)

- In French <http://www.sthda.com/french/wiki/test-de-student-formules>



# 1) Two paired series

Two sets of vowels (the same vowels for two positions)

$$t = \frac{m}{s/\sqrt{n}}$$

$m$  and  $s$  are the mean of the difference of the series of measurements (F1 formant for instance) and standard deviation of the difference of the series of measurements and  $n$  the size of the series.

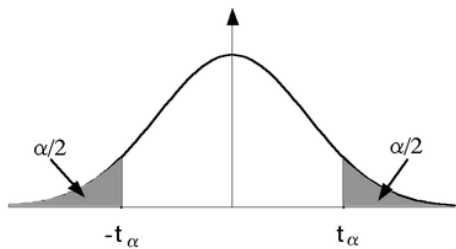
The dof (degree of freedom) is  $dof = n - 1$

read in the table the value corresponding to the risk  $\alpha = 5\%$  considering the degree of freedom.

In Excel :

=T.TEST(A1:A10;B1:B10;2;1)    2 bilateral/unilateral

Type 1 (paired), 2 (equal variance), 3 (different variances)



Exemple : avec d.d.l. =10, pour  $t=2.228$  la probabilité est  $\alpha=0.05$

d.d.l./ $\alpha$	0.9	0.5	0.3	0.2	0.1	0.05	0.02	0.01	0.001
1	0.158	1	2	3.078	6.314	12.706	31.821	64	637
2	0.142	0.816	1.386	1.886	2.92	4.303	6.965	10	31.598
3	0.137	0.765	1.25	1.638	2.353	3.182	4.541	5.841	12.929
4	0.134	0.741	1.19	1.533	2.132	2.776	3.747	4.604	8.61
5	0.132	0.727	1.156	1.476	2.015	2.571	3.365	4.032	6.869
6	0.131	0.718	1.134	1.44	1.943	2.447	3.143	3.707	5.959
7	0.13	0.711	1.119	1.415	1.895	2.365	2.998	3.499	5.408
8	0.13	0.706	1.108	1.397	1.86	2.306	2.896	3.355	5.041
9	0.129	0.703	1.1	1.383	1.833	2.263	2.821	3.25	4.781
10	0.129	0.7	1.093	1.372	1.812	2.228	2.764	3.169	4.587
11	0.129	0.697	1.088	1.363	1.796	2.201	2.718	3.106	4.437
12	0.128	0.695	1.083	1.356	1.782	2.179	2.681	3.055	4.318
13	0.128	0.694	1.079	1.35	1.771	2.16	2.65	3.012	4.221
14	0.128	0.692	1.076	1.345	1.761	2.145	2.624	2.977	4.14
15	0.128	0.691	1.074	1.341	1.753	2.131	2.602	2.947	4.073



<b>16</b>	0.128	0.69	1.071	1.337	1.746	2.12	2.583	2.921	4.015
<b>17</b>	0.128	0.689	1.069	1.333	1.74	2.11	2.567	2.898	3.965
<b>18</b>	0.127	0.688	1.067	1.33	1.734	2.101	2.552	2.878	3.922
<b>19</b>	0.127	688	1.066	1.328	1.729	2.093	2.539	2.861	3.883
<b>20</b>	0.127	0.687	1.064	1.325	1.725	2.086	2.528	2.845	3.85
<b>21</b>	0.127	0.686	1.063	1.323	1.721	2.08	2.518	2.831	3.819
<b>22</b>	0.127	0.686	1.061	1.321	1.717	2.074	2.508	2.819	3.792
<b>23</b>	0.127	0.685	1.06	1.319	1.714	2.069	2.5	2.807	3.767
<b>24</b>	0.127	0.685	1.059	1.318	1.711	2.064	2.492	2.797	3.745
<b>25</b>	0.127	0.684	1.058	1.316	1.708	2.06	2.485	2.787	3.725
<b>26</b>	0.127	0.684	1.058	1.315	1.706	2.056	2.479	2.779	3.707
<b>27</b>	0.137	0.684	1.057	1.314	1.703	2.052	2.473	2.771	3.69
<b>28</b>	0.127	0.683	1.056	1.313	1.701	2.048	2.467	2.763	3.674
<b>29</b>	0.127	0.683	1.055	1.311	1.699	2.045	2.462	2.756	3.649
<b>30</b>	0.127	0.683	1.055	1.31	1.697	2.042	2.457	2.75	3.656
<b>40</b>	0.126	0.681	1.05	1.303	1.684	2.021	2.423	2.704	3.551
<b>80</b>	0.126	0.679	1.046	1.296	1.671	2	2.39	2.66	3.46
<b>120</b>	0.126	0.677	1.041	1.289	1.658	1.98	2.358	2.617	3.373
<b>Infini</b>	0.126	0.674	1.036	1.282	1.645	1.96	2.326	2.576	3.291

## 2) Two independent series

Two independent sets of vowels (for instance two texts)

$m_A$   $m_B$  are the two mean values (F1 for the two sets of vowels)

$$t = \frac{m_A - m_B}{\sqrt{\frac{s^2}{n_A} + \frac{s^2}{n_B}}}$$

$$s^2 = \frac{\Sigma(x - m_A)^2 + \Sigma(x - m_B)^2}{n_A + n_B}$$

To know whether the difference is significant read in the table the value corresponding to the risk  $\alpha = 5\%$  considering the dof (degree of freedom).

$$\text{dof} = n_A + n_B - 2$$

In Excel :

=T.TEST(A1:A10;B1:B10;2;2) 2 bilateral/unilateral

Type 1 (paired), 2 (equal variance), 3 (different variances)