Written corpora
presentation and organization

Karën Fort

karen.fort@univ-lorraine.fr / https://members.loria.fr/KFort

Institut des sciences du Digital Management & Cognition
COMPOSANTE DE L'UNIVERSITÉ DE LORRAINE

UNIVERSITÉ DE LORRAINE

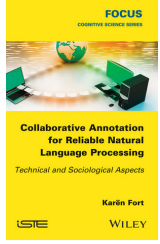
Karën Fort (she/her)

Where I talk from

See https://members.loria.fr/KFort/

- Ethics and NLP (HDR)

- Language resources creation for NLP (PhD thesis)



You?

I say what I do

and I do as I say

- everybody arrives on time
- everybody attends all the classes (please email me in case of emergency)
- do not change groups
- if you speak, speak to **us**
- no social network, no email, no (smart)phone, no food

- do not hesitate to ask questions: there are dummy questions, but I'll answer them too (once)!
- do not hesitate to tell me if:
  - I'm going too fast/not fast enough
  - you already know what I'm talking about
  - my English is not good (correct me, please)
- do not hesitate to question what I'm saying

How to contact me

- email: karen.fort@loria.fr
- office: at IDMC (215) or LORIA (B116), email before

Material

All my material is available online:

- on my website: https://members.loria.fr/KFort/idmc-nancy-from-2024/
- not on Arche
- under a CC licence

Organization of the course

- 7x2 sessions of 2h:
  - 7 lectures (CM), with me
  - 7 practice (TD), with me and Clémentine Bleuze

- evaluation:
  - (potentially) assignements: you **all** do it and send it to us (me and Ms Bleuze) and only some of you will be graded each time (it's random)
  - a presentation (more below)
  - a final exam – 50% of the final grade

Presentations

- Groups of 3 to 4 (no more than 6 groups per TD)
- Within the same TD
- Presented during the last TD session
- 15 min presentation + 5 min questions

- Subject: present an existing corpus
  - manually annotated
  - original: type of annotation, domain of the corpus, language, etc

Evaluation

Criteria:

- choice of corpus
- quality of the work (thoroughness of the analysis)
- quality of the presentation
- answers to questions

Beware: the grade will not necessarily be the same for all group members

Beware 2: no plagiarism or AI generated slides will be tolerated

Provisional agenda

1. Introduction and corpora
2. Encodings
3. Manual annotation
4. Evaluating manual annotation
5. Solutions to the annotation costs
6. Manual annotation complexity dimensions
7. *Crowdsourcing*: typologies, myths and reality, games with a purpose (*GWAP*)

Presentations: beware!