WP1 - Corpus preparation, lexicon design, corpus-lexicon interface

WP1 subtasks

- WP1.1: Corpus collection, documentation and preprocessing
- WP1.2: Creation of a model for contextual representations
- WP1.3: Extraction and mapping of the pre-lexicon
- WP1.4: Design of the final semantic lexicon format

Resources required for the project

- A (very) large **corpus** of French raw texts (for induction)
- **Pre-processing** tools to apply on the raw corpus
  - POS, parsing, deep parsing, MWEs
- A **pre-lexicon** extracted from Wiktionary

- Less crucial
  - Pre-trained LMs: rely on huggingface? Further pre-training on SELEXINI raw corpus?
  - Evaluation sets for intrinsic evaluation of lexicon induction: Wiktionary, CALOR, Asfalda, FRSemCor

1. Raw corpus: criteria

- **Open** licences, shareable
- **Diverse** (registers, domains) and more or less balanced
- **Deduplicated**, clean
- **Size** matters?
- Keep order of sentences in documents?
  - Ideally, yes
- **Compatible** with other WPs' evaluation domains (include sentences)
  - WP2: FRSemCor, Asfalda, Calor, SemEval…
  - WP4: FQuad, CALOR-QUEST…
  - WP5: PARSEME…

1. Raw corpus: some options

- FRWaC https://wacky.sslmit.unibo.it/
- OSCAR https://oscar-corpus.com/
- CommonCrawl https://commoncrawl.org/
  - CoNLL ST 2017 UD-parsed version
    - https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-1989
- Orfeo
  - https://hal.archives-ouvertes.fr/hal-01449600
- BigScience
  - https://bigscience.huggingface.co/blog/building-a-tb-scale-multilingual-dataset-for-language-modeling
- Corpus used to train FlauBERT https://arxiv.org/abs/1912.05372
- Corpus used to train CamemBERT https://aclanthology.org/2020.acl-main.645/

2. Preprocessing tools: criteria

- In-house tools (more control) vs. off-the-shelf tools (SOTA, less hassle)
- How far do we go?
  - Strict requirements
    - Tokenisation
    - MWEs (all categories)
      - Using Travis.mono (?) for PARSEME-FR or FTB
      - Using lexicon (Wiktionary) projection/matching
    - POS - UD is enough
    - Lemmas - investigate - UDPipe on French UD merged?
    - Deep syntactic arguments DeepSequoia
  - Would be nice
    - Have an idea of the quality (for release)
    - Syntax
    - Morphological features
- Formalism
  - UD, SUD, enhanced UD, deep Sequoia…

2. Preprocessing tools: some options

- Parsing and POS tagging
  - UDPipe https://ufal.mff.cuni.cz/udpipe
  - Macaon https://aclanthology.org/2021.iwpt-1.3/
  - Off-the-shelf: stanza, CoreNLP, spacy…
- MWE identification
  - Transition https://aclanthology.org/W17-1717/
  - MTLB-Struct https://aclanthology.org/2020.mwe-1.19/
  - LSR https://aclanthology.org/2021.mwe-1.6/
  - Systems must be trained on PARSEME-FR corpus https://hal.archives-ouvertes.fr/hal-03016721
- Named entity recognition
  - Same as MWE identification?
- Develop from scratch?

3. Pre-lexicon: criteria

- Open licences
- Coverage, quality
- Ease of manipulation, extraction
- What information to extract/keep?
  - Lexeme (lemma+POS)
  - Senses (hierarchy?)
  - Example sentences (WP2)
  - Definitions (WP3)
- Format of shared pre-lexicon

3. Pre-lexicon: some options

- Wiktionary
  - DBNary http://kaiko.getalp.org/about-dbnary/
  - GLAFF http://redac.univ-tlse2.fr/lexiques/glaff.html
  - Wiktextract https://github.com/tatuylonen/wiktextract
  - Raw dumps https://dumps.wikimedia.org/
- Other lexicons (maybe?)
  - TLFi http://atilf.atilf.fr/
  - BabelNet etc. https://babelnet.org/

A proposal

- Three working groups in parallel
  - Raw corpus
  - Pre-processing
  - Pre-lexicon
- Goal: prepare a survey table on resources available for French
  - Select criteria to record (table columns) ~15min
    - Reference (URL, paper) 🔗
    - Strengths
    - Limitations 👎
    - Size, licence, domains, access, format…
  - Select resources to consider (table rows) ~30min
  - Fill in the comparative table by noting the criteria values for the resources (table cells) ~45min
  - Present the resulting table to the whole group

To go further...

- Corpus access interface
  - SQL DB
  - NoSketch engine
- Lexicon-corpus matching
  - MWEs