WP 5

Semantic lexicon at the service of diversity

Participants

- Partners in charge: LISN (A. Savary)
- Involved partners:
  - ATILF (M. Constant),
  - LIFAT (C. De Runz, A. Soulet),
  - LIS (C. Ramisch)

Motivation

- Diversity of linguistic phenomena - a **heritage** to be preserved
- Zipf's law: **few frequent** items; **long tail** of **rare** items
- **Models** and **performance measures** often favour the former and underperform in the latter
- In benchmark-based evaluation **generalisation** and **robustness** are rarely assessed
- Diversity important for the quality of NLP **applications**
  - parsing (Narayan & Cohen 2015)
  - QA (Yang et al. 2018)
  - dialog systems (Palumbo et al. 2020)
- ... but largely neglected in building and evaluating them.

Objectives

- **Quantify** linguistic diversity (on the example of MWEs)
- Define **evaluation scenarios** which favor diversity in MWE identification
- Assess the contribution of the **semantic lexicon** to increasing the diversity in MWE identification

WP 5.1: Quantifying diversity of multiword expressions in a corpus and in system predictions

- Why MWE?
  - A phenomenon we control and understand
  - Diversity due to idiosyncrasy (Constant et al. 2017)
  - Critical hardness of generalisation over unseen data (Savary et al. 2019)
- Dimensions of diversity
  - variety = number of **types**
  - balance = evenness of the distribution of **items** in types
  - disparity = **distances** between types
- Links with (corpus, sentence, ...) **complexity**

Types and items

- **MWE** lemmas (*commit theft*) and occurrences (*committed theft, thefts committed*)
  - features for disparity: vocabulary, morphology, syntactic dependencies, contexts of MWE occurrences
- Semantic **slots** (Agent, Patient) and their realizations
- Semantic **frames** (*steal*, *fly*) and their occurrences

Representativeness

- Aim: estimate how representative a corpus is of diversity in language
- Focusing on rare MWEs
  - **Good-Turing test** & **Benford's law**, previously applied to knowledge bases (Soulet et al. 2018; Yan et al. 2018)
  - estimating how many types **unseen** in a corpus exist in language, based on rarely seen types

Evaluation scenario

- Task: MWE identification
- Methods
  - diversity-driven, corpus split, over-sampling and augmentation
  - diversity measures applied to system outcomes
  - favour MWE identifiers performing well on rare and diverse phenomena (Ramisch et al. 2020),
  - across possibly many languages
- Framework: PARSEME shared task on automatic identification of MWEs (Savary et al. 2017, Ramisch et al. 2018, Ramisch et al. 2020)

WP5.2: Diversity-oriented extrinsic evaluation of the semantic lexicon

- Hypothesis: the induced lexicon to be more representative of linguistic diversity than both handcrafted lexicons (Wiktionary) and manually annotated corpora (PARSEME corpora)
  - known MWEs (WP1) linked to **new corpus occurrences**
  - new MWEs discovered from **outlier frames** (WP3.3)
- Evaluation scenario
  - Extend the PARSEME corpus with the SELEXINI corpus with MWEs
  - Assess the resulting joint corpus for representativeness (WP5.1)
  - Train and evaluate MWE identifiers for diversity

Deliverables

- MWE-annotated **corpus** with gold PARSEME data and pseudo-gold occurrences,optimally split for diversity
- **Lexicon assessment** in terms of diversity

Ongoing work

- **Adam Lion-Bouton** - PhD on MWE lexicon format and diversity
- New PhD topic defined with Arnaud Soulet, Cyril De Runz (LIFAT) and Thomas Lavergne (LISN)
- Links with
  - Dagstuhl seminar on « Universals of Linguistic Idiosyncrasy in Multilingual Computational Linguistics », August 2021, May 2023
  - CA21167 COST action UniDive « Universality, diversity and idiosyncrasy in language technology » (2022-2026)