On the Intra- and Inter-linguistic Challenges of Multilingual Silver-Data Creation and Disambiguation Biases

*Edoardo Barba, Niccolò Campolungo, **Simone Tedeschi** and Roberto Navigli*

- **Silver-data creation** [1, 2] is a **powerful, fast and cheap tool** that can be used to **tackle both inter- and intra-linguistic challenges** in NLP by producing training data for:
  - **Low-resource languages**
  - A **variety of tasks**, including those involving **figurative language**
- Lexical-semantic **disambiguation biases** strongly **affect NLP systems** [3]
  - Analyses on the DiBiMT benchmark show that **MT models are still far from correctly handling infrequent senses**
- **Relevant WGs: 1, 3 and 4**

[1] Tedeschi et al. (2021) *"WikiNEuRal: Combined neural and knowledge-based silver data creation for multilingual NER."*
[2] Tedeschi et al. (2022) *"ID10M: Idiom Identification in 10 Languages."*
[3] Campolungo et al. (2022) *"DiBiMT: A novel benchmark for measuring Word Sense Disambiguation biases in Machine Translation."*

Babelscape
Sapienza Università di Roma
elexis european lexicographic infrastructure
Project No. 731015
PRIN MIUR
PerLIR Project
erc
MOUSSE ERC Consolidator Grant No. 726487
cost European Cooperation in Science & Technology
Cost Action CA21167 (UniDive)
UniDive 1st general meeting