Data Statements for the REFreSD dataset

- ✓ **Dataset name:** **R**ationalized-**E**nglish **Fre**nch **S**emantic **D**ivergences
- ✓ **Citation:** Detecting Fine-Grained Cross-Lingual Semantic Divergences without Supervision by Learning to Rank, *In EMNLP 2020*
- ✓ **Data statements developers:** Eleftheria Briakou & Marine Carpuat
- ✓ **Data statement author:** Eleftheria Briakou
- ✓ **Others** who contributed to this document: Tommaso Caselli & Valerio Basile
- ✓ **Link to dataset:** https://github.com/Elbria/xling-SemDiv/tree/master/REFreSD

INTRODUCTION

This document presents Data Statements for the Rationalized English-French Semantic Divergences (REFreSD) dataset, published at EMNLP 2020. The following list was collected at the **LREC 2020 Workshop on Data Statements**, organized by: Emily M. Bender, Batya Friedman, and Angelina McMillan-Major.

This project's objective is to advance our fundamental understanding of the computational representations and methods to compare and contrast text meaning **across languages**. Currently, much cross-lingual work in Natural Language Processing relies on the assumption that sentences drawn from parallel corpora are equivalent in meaning. Yet, content conveyed in two distinct languages is rarely exactly equivalent. We assess the ability of computational methods to detect such meaning mismatches by comparing their predictions with human judgments on our REFreSD dataset. Human annotators are asked to read text excerpts in two languages (e.g., one in English and another in French) and we collect their assessment of the meaning differences they observe via sentence-level divergence judgments and token-level rationales.

A. CURATION RATIONALE

Examples are drawn from the English-French section of the **WikiMatrix** corpus. We choose this resource because (1) it is likely to contain diverse, interesting divergence types since it consists of mined parallel sentences of diverse topics which are not necessarily generated by (human) translations, and (2) Wikipedia and WikiMatrix are widely used resources to train semantic representations and perform cross-lingual transfer in NLP. We exclude obviously noisy samples by filtering out sentence-pairs that a) are too short or too long, b) consist mostly of numbers, c) have a small token-level edit difference.

B. LANGUAGE VARIETY/VARIETIES

The dataset is built on top of the English-French **Wikimatrix** dataset—a corpus of mined parallel sentences from English and French Wikipedia articles—for which information on language varieties is not available.

C. SPEAKER DEMOGRAPHIC

The original corpus (Wikimatrix) is extracted from Wikipedia articles in English and French. Some content of Wikipedia articles has been (human) translated from existing articles in another language while others have been written or edited independently in each language. Therefore, information on how the original text is created is not available.

D. ANNOTATOR DEMOGRAPHIC

This dataset includes annotations from 6 participants recruited from the University of Maryland, College Park (UMD) educational institution. Recruitment for this project is driven by emails and flyers circulated via relevant UMD units, as well as through personal contacts in these units. Participants range in age from 20–25 years, including one man and five women. For each participant, we ensure they are proficient in both languages of interest: three of them self-report as English native speakers, one as a French native speaker, and two as bilingual English-French speakers. All non-bilingual L1-English speakers are graduate students in French Translation Studies while the L1-French participant pursues a degree in Linguistics at UMD.

E. SPEECH SITUATION

N/A

F. TEXT CHARACTERISTICS

The annotated English-French sentence-pairs are **randomly** sampled from Wikimatrix corpus. Therefore, they are expected to cover many topics.

G. RECORDING QUALITY

N/A

H. PROVENANCE APPENDIX

WikiMatrix: Mining 135M Parallel Sentences in 1620 Language Pairs from Wikipedia