MULTINCI [WIP] – A MULTILINGUAL NOUN COMPOUND IDIOMATICITY DATASET

THOMAS PICKARD
UNIVERSITY OF SHEFFIELD
tmrpickard1@sheffield.ac.uk

B. MĂDĂLINA ZGREABĂN
UNIVERSITEIT UTRECHT
b.zgreaban@uu.nl

ALINE VILLAVICENCIO
UNIVERSITY OF SHEFFIELD
a.villavicencio@sheffield.ac.uk

NCTTI

- The Noun Compound Type and Token Idiomaticity dataset [7]:
  280 English (en) and 180 Portuguese (pt) nominal compounds (NCs).
- Human annotated in three context sentences (type and token-level annotation).
- PROs: effects of context on annotation judgements, comparison for language models.

MULTINCI – OBJECTIVES

- Extended NCTTI dataset having core NCs common cross-linguistically, as well as language-specific compounds.
- Include languages with limited MWE resources.
- Increase cross-lingual applications by having correspondences across languages.

LANGUAGES

- **English** (en):
  - Cleaned & updated context sentences
  - Extended compound list to increase potentially idiomatic expressions
- **Romanian** (ro):
  - Test case for protocol
  - 260 NCs
  - 36 directly equivalent to en; 39 exclusive to Romanian, and 185 that have en translations (not part of the original NCTTI)
- **Georgian** (ka), Irish (ga):
  - Initial work underway (funded by UniDvie STSMs)
- **Modern Greek** (el), **Ukrainian** (uk), **Brazilian Portuguese** (pt-br):
  - Potential collaborations identified

FUTURE WORK

- **Protocols** to be refined and completed.
- **Data collection**, translation and its annotation for in-progress and planned languages.
- **Annotations** from human volunteers.
- Extend MultiNCI to **more languages** and their varieties
  - Collaborations welcome
  - See UniDive STSM call

Funded by the European Union

UniDive

cost
EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

REFERENCES