Funded by the European Union

cost EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

MULTINCI – A MULTILINGUAL NOUN COMPOUND IDIOMATICITY DATASET

UniDive

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

THE NCTTI DATASET

- The Noun Compound Type and Token Idiomaticity dataset (NCTTI, [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.

PROBLEM:

- Idiomaticity differs cross-linguistically [12], [4], [3].

SOLUTION: MULTINCI

- Extended NCTTI dataset having core NCs common cross-linguistically and language-specific compounds.
- PROs: Balanced for idiomaticity and translated to en.

OBJECTIVES

- Include languages with limited MWE resources (WG4).
- Increase cross-lingual applications by having correspondences across languages (WG3).
- Include varied contexts [8].
- Human plausibility ratings for target items [WG1].

WORK TO DATE

- Developing a protocol for:
  - Collecting NCs and context sentences for new languages.
  - Developing web-platforms for annotation [6], [7].
  - Methodological incorporation of literality judgements [5] after [10].
  - Methodological incorporation of contextual judgements for potentially idiomatic NCs.

ENGLISH

- NCTTI: context sentences from ukWaC [2] and brWaC [11].
- Problems: Incomplete sentences; undesired elements; typographical errors; offensive or biased content.
- Solution: NCTTI cleaned and supplemented with sentences from EnTenTen20 and EnTenTen21 [9].
- Extended compound list to increase potentially idiomatic items [8].

ROMANIAN: NCS

- Test case for protocol.
- En NCs translated to Romanian: 222 translations.
- Translations excluded if they:
  - used prepositions or had a different meaning in Romanian;
  - were borrowed in their en form;
  - were part of another MWE category;
  - were not likely to be used by speakers;
- Excluded translations replaced by Romanian NCs from Wiktionary.
- Final Romanian dataset: 109 non-compositional, 88 partially compositional and 89 fully compositional NCs.
- 36 directly equivalent to en; 39 exclusive to Romanian, and 185 that have en translations (not part of the original NCTTI).

(1) A crescut știind la perfecție limb-a germană, ca **limbă maternă**. (ro)
have.PST.3SG grow.PST.3SG know.GER of perfection language-the German as **language maternal**

lit. 'He grew up knowing German perfectly, as his **maternal language**.'
'He grew up knowing German perfectly, as his **mother tongue**.'
(RoTenTen21 (Jakubíček et al., 2013); our translation and gloss)

(2) Mă simteam **oai-a neagră** a concursu-l-ui. (ro)
myself feel.PST.IPFV.1SG **black sheep**-the of contest-the-GEN

'I was feeling like the **black sheep** of the contest.'
(RoTenTen21 (Jakubíček et al., 2013); our translation and gloss)

(3) Nici n-ar zice omul că sunteți niște **coate-goale**. (ro)
neither NEG-would say.PRS.3 man-the that be.PRS.2PL some **elbows-naked.PL**

lit. 'Neither would he say that you are **elbows-naked**.'
'Nor would he say that you are a **poor person**.'
(RoTenTen21 (Jakubíček et al., 2013); our translation and gloss)

Figure 1: Examples of Romanian target items in context.

ROMANIAN: CONTEXT SENTENCES

- For token-level annotation: from roTenTen [9] and fewer from CoRoLa [1].
- Sentences minimally modified for clarity or if they provided definitions of NCs.

ROMANIAN: CHALLENGES

- NCs in Romanian contain genitive nouns, e.g. ochii minții (lit. 'eyes of the mind') 'mind's eye', or change their form when in the genitive.
- Solutions: new annotation guidelines introduced, modifications to web scripts for annotation.

OTHER LANGUAGES

- UNDERWAY: Georgian (ka) and Irish (ga) (funded by short-term scientific missions),
- UNDERWAY: potential collaborations for modern Greek (el), Ukrainian (uk) and Brazilian Portuguese (pt-br).

FUTURE WORK

- PROTOCOL to be refined and completed.
- Data collection, translation and its annotation for in-progress and planned languages.
- Annotations from human volunteers.
- Extend MultiNCI by more languages and their varieties (COLLABORATIONS WELCOME).

REFERENCES