MultiNCI - A Multilingual Noun Compound Idiomaticity Dataset

Thomas Pickard
University of Sheffield / Sheffield, UK
tmrpickard1@sheffield.ac.uk

B. Madalina Zgreabăn
Universiteit Utrecht / Utrecht, NL
b.zgreaban@uu.nl

Aline Villavicencio
University of Sheffield / Sheffield, UK
a.villavicencio@sheffield.ac.uk

Relevant UniDive working groups: WG1, WG3, WG4

1 Introduction

The Noun Compound Type and Token Idiomaticity dataset (NCTTI, Garcia et al., 2021), produced in 2021, is a collection of 280 English (en) and 180 Portuguese (pt-br) nominal compounds (NCs) across a range of compositionalities (including fully transparent entries), with three naturally-occurring context sentences included for each item. These are annotated with human judgements of compositionality at type level and within each context sentence. These annotations allow for evaluation of context-dependent effects on human judgements, as well as providing ratings which can be compared with language model outputs for the same stimuli. However, as idiomatic realisations differ from language to language and from culture to culture (Yağiz and Izadpanah, 2013; Boers et al., 2004; Boers and Demecheleer, 2001), to enable a multilingual study of diversity of idiomatic realisation, we created **MultiNCI**, a dataset containing a common core of NCs shared across languages, in addition to language-specific compounds. The dataset is well-balanced for idiomaticity at a language level, and contains not only the idiomatic equivalents of the NCs, but also the translations for each of the NCs into English.

The present outline describes work-in-progress which will form the basis of a modified and expanded version of the NCTTI dataset, with the following objectives:

- Expanding the languages represented in the dataset, particularly to include languages for which MWE resources are limited *(WG4)*.
- Mapping correspondences between target items across languages, making the dataset more valuable for multi- and cross-lingual applications *(WG3)*.
- Including varied contexts for potentially idiomatic nominal compounds (PIEs, Haagsma et al., 2020), and expanding the dataset to include more of these.
- Capturing ratings from participants on the literal plausibility of target items included in the data *(WG1)*.

2 Work to Date

General (non-language-specific) work to date has focused on development of a protocol for collation of a set of target nominal compounds and context sentences for new languages, deployment of a web-based system for annotation collection (based on those used in Cordeiro et al., 2019 and Garcia et al., 2021), and expansion of the methodology to incorporate gathering of literality judgements (adapted from Bulkes and Tanner, 2017, after Libben and Titone, 2008) and in-context judgements for literal instances of potentially idiomatic NCs.

2.1 English

The NCTTI dataset uses context sentences extracted from ukWaC (Baroni et al., 2009) and brWaC (Wagner Filho et al., 2018). In many cases, however, these sentences are incomplete or contain undesirable elements such as non-linguistic or nonsensical 'web text', distracting typographical errors, and potentially offensive or biased content. In order to alleviate these issues, the English (en) context sentences have been manually cleaned and supplemented with new instances taken from the EnTenTen20 and EnTenTen21 corpora (Jakubíček et al., 2013) and other web sources or constructed by the authors where literal instances could not be readily found in corpora. The compound list has also been expanded to increase the number of potentially idiomatic items (PIEs; Haagsma et al., 2020) while maintaining balance across the compositionality classes.

(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.

2.2 Romanian

The development of a set of target items, context instances, and annotation collection tools for **Romanian** (ro) has acted as a test case for our collation protocol.

Starting with the list of English nominal compounds, we translated them to find Romanian equivalents[1]. Of the 280 English compounds, 222 had Romanian translations. However, not all of these were suitable for use – if a translation included prepositions, had a different meaning in Romanian than the original one, was borrowed in its English form, was part of another category of MWEs (e.g. verbal MWEs), or was not likely to be used by speakers, it was excluded from the dataset.

Replacements for excluded compounds and those without Romanian translations were sought, using Wiktionary as a source while matching compositionality ratings with the original dataset.

The final Romanian target list contains a balanced number of nominal compounds, namely 109 non-compositional, 88 partially compositional and 89 fully compositional. 36 of these are directly equivalent to items on the English list (e.g. Figure 1 item 1), 185 were not on the English NCTTI list but do have English translations (e.g. Figure 1 item 2), and 39 are specific to Romanian and do not have direct equivalents in English (e.g. Figure 1 item 3).

The context sentences for token-level annotation were taken mostly from roTenTen (Jakubíček et al., 2013), with a few additions from CoRoLa (Barbu Mititelu et al., 2018). The sentences required few modifications – where needed, changes were minimal, motivated either for clarity or so as not to give away the meaning of the compound.

Several challenges were encountered when adapting the annotation guidelines and processes from English, mainly due to the rich morphology of Romanian. For example, many of the Romanian nominal compounds contain nouns in genitive forms, such as **ochii minții** (lit. 'eyes of the mind') 'mind's eye', or change their form to the genitive when used in a sentence. Thus, new categories in the annotation guidelines had to be introduced so as to cover the possible morphological forms of the nominal compounds.

2.3 Other Languages

Similar activity to that described above for Romanian is underway for **Georgian** (ka) and planned for **Irish** (ga), in both cases supported by UniDive-funded short-term scientific missions. Potential collaborators have been identified who may be interested in working on **modern Greek** (el) and **Ukrainian** (uk), and on expanding the existing coverage of **Brazilian Portuguese** (pt-br).

[1]While this approach is likely to introduce an element of bias into the dataset, it also allows us to maximise instances of parallelism and MWEs with equivalent semantics.

3 Future Work

Future activities planned for MultiNCI prior to publication of the dataset include:

- Refinement and completion of the data collection protocol.
- Target data collation, translation and deployment of the annotation system for the in-progress and planned languages.
- Recruitment of and collection of annotations from volunteer annotators.

We also plan to continue expanding the coverage of MultiNCI by incorporating additional languages and language varieties, and welcome input and contributions from UniDive members.

4 Acknowledgements

This work was in part supported by the Centre for Doctoral Training in Speech and Language Technologies (SLT) and their Applications funded by UK Research and Innovation [grant number EP/S023062/1].

References

Verginica Barbu Mititelu, Dan Tufi\textcommabelows, and Elena Irimia. 2018. The Reference Corpus of the Contemporary Romanian Language (CoRoLa). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

F Boers and M Demecheleer. 2001. Measuring the impact of cross-cultural differences on learners' comprehension of imageable idioms. *ELT Journal*, 55(3):255–262.

Frank Boers, Murielle Demecheleer, and June Eyckmans. 2004. Cross-cultural Variation as a Variable in Comprehending and Remembering Figurative Idioms. *European Journal of English Studies*, 8(3):375–388.

Nyssa Z. Bulkes and Darren Tanner. 2017. "Going to town": Large-scale norming and statistical analysis of 870 American English idioms. *Behavior Research Methods*, 49(2):772–783.

Silvio Cordeiro, Aline Villavicencio, Marco Idiart, and Carlos Ramisch. 2019. Unsupervised Compositionality Prediction of Nominal Compounds. *Computational Linguistics*, 45(1):1–57.

Marcos Garcia, Tiago Kramer Vieira, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2021. Assessing the Representations of Idiomaticity in Vector Models with a Noun Compound Dataset Labeled at Type and Token Levels. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2730–2741, Online. Association for Computational Linguistics.

Hessel Haagsma, Johan Bos, and Malvina Nissim. 2020. MAGPIE: A large corpus of potentially idiomatic expressions. In *Proceedings of the 12th language resources and evaluation conference*, pages 279–287, Marseille, France. European Language Resources Association.

Miloš Jakubíček, Adam Kilgarriff, Vojtěch Kovář, Pavel Rychlý, and Vít Suchomel. 2013. The TenTen Corpus Family. *Proceedings of the 7th International Corpus Linguistics Conference CL 2013*.

Maya R. Libben and Debra A. Titone. 2008. The multidetermined nature of idiom processing. *Memory & Cognition*, 36(6):1103–1121.

Jorge A. Wagner Filho, Rodrigo Wilkens, Marco Idiart, and Aline Villavicencio. 2018. The brWaC Corpus: A New Open Resource for Brazilian Portuguese. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Oktay Yağiz and Siros Izadpanah. 2013. Language, Culture, Idioms, and Their Relationship with the Foreign Language. *Journal of Language Teaching and Research*, 4(5):953–957.