



COST Action CA21167: Universality, diversity and idiosyncrasy in language technology (UniDive)

Proposal to host short-term scientific mission (STSM)

December 2022

HOST INFORMATION

1. Host institution

The proposed STSM would be hosted in the United Kingdom at the [University of Sheffield](#):

Department of Computer Science
The University of Sheffield
Western Bank
Sheffield
UNITED KINGDOM
S10 2TN

[Google Maps](#)

1.1 Sheffield

[Sheffield](#) is a medium-sized city in the North of England, with a vibrant character, industrial heritage the strong sense of identity typical of Yorkshire. The city has lots of green space, access to several National Parks and transport links which connect it to major cities of the UK.

2. Contact information

Primary contact: [Professor Aline Villavicencio](#), Chair in Natural Language Processing
a.villavicencio@sheffield.ac.uk

Secondary contact: Mr Thomas Pickard
tmrpickard1@sheffield.ac.uk

STSM DETAILS

1. Duration and Dates

The proposed STSM would ideally take place between the **10th June** and the **1st September 2023** (to align with non-teaching periods at the host institution), for a duration of **3 to 4 weeks**. However, there is scope for flexibility in this if different timing better suits the needs of participants.

If more than one participant takes part in the STSM, their activity dates would ideally overlap, in order to maximise the efficacy of the activity.

2. Description

The purpose of the proposed STSM is to expand on previous work undertaken to develop the [NCTTI dataset](#) [Garcia et al. 2021] and subsequent expansion leading to the dataset used in the [SemEval 2022 Task 2](#) [Tayyar Madabushi et al. 2021, 2022] on Multilingual Idiomaticity Detection.

The NCTTI dataset consists of potentially idiomatic nominal compounds in English and Brazilian Portuguese, context sentences and compositional paraphrases. These were used to collect compositionality judgements from native speakers, providing a source of valuable information for researchers interested in the phenomenon of idiomaticity.

The proposed STSM would seek to build on this work by extending the NCTTI dataset to cover additional languages. In particular, we would like to focus on languages for which existing resources on idiomaticity are limited. Opportunities also exist to expand the scope of annotation for new and existing languages, e.g. by incorporating dependency parsing information.

Detailed activity would be guided by the visiting researcher(s) and is likely to be influenced by the availability of existing resources such as lexica of idiomatic expressions and corpora of context sentences. An outline activity plan is likely to include the following:

- Collation of target compounds for the language(s) of interest
- Collection of suitable context sentences containing the target compounds
- Development / translation of annotation guidelines in the target language
- Testing and deployment of the annotation system

- Recruitment of annotators
- Monitoring of annotations and addressing issues which might arise
- Analysis of results and outputs

It is likely that some of this activity will extend beyond the core dates of the STSM, especially as it may take some time to collect annotations. Ongoing remote collaboration between participants and the host institution will enable these activities to be completed.

The STSM activity is intended to generate an updated and expanded version of the NCTTI dataset, which will be made publicly available via an open data repository, and is likely to lead to a corresponding publication.

This will contribute directly to the objectives of UniDive WG1 (Corpus Annotation) by expanding the availability of annotated corpora for additional languages, and by capturing dependency parsing information.

These resources will also be beneficial to ongoing efforts to develop language technologies which better handle idiomatic language, in particular the [Modelling Idiomaticity in Human and Artificial Language Processing](#) project which is primarily based at the STSM host institution and seeks to develop cross-lingual representations of idiomaticity. This aligns with UniDive WG3's focus on cross-lingual and multilingual language technology.

References

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 *Proc. ACL-IJCNLP 2021*, ACL.

Harish Tayyar Madabushi, Edward Gow-Smith, Carolina Scarton, and Aline Villavicencio. 2021. [AStitchInLanguageModels: Dataset and Methods for the Exploration of Idiomaticity in Pre-Trained Language Models](#). In *Findings EMNLP 2021*, ACL.

Harish Tayyar Madabushi, Edward Gow-Smith, Marcos Garcia, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2022. [SemEval-2022 Task 2: Multilingual Idiomaticity Detection and Sentence Embedding](#). In *Proc. SemEval-2022*, ACL.