Universality, Diversity and Idiosyncrasy in Language Technology

Gediminas Navickas, Gražina Korvel

Institute of Data Science and Digital Technologies
Vilnius University

Background

Efficient access to the constantly growing quantities of data, especially of language data, largely relies on advances in data science. This domain includes natural language processing (NLP), which is currently booming, to the benefit of many end users. However, this optimization-based technological progress poses an important challenge: accounting for and fostering language diversity.

The UniDive Action takes two original stands on this challenge. Firstly, it aims at embracing both inter- and intra-language diversity, i.e. a diversity understood both in terms of the differences among the existing languages and of the variety of linguistic phenomena exhibited within a language. Secondly, UniDive does not assume that linguistic diversity is to be protected against technological progress but strives for both of these aims jointly, to their mutual benefit.

UniDive builds upon previous experience of European networks and projects which provided a proof of concept for language modelling and processing, unified across many languages but preserving their diversity.

The main benefits of the Action will include, on the theoretical side, a better understanding of language universals, and on the practical side, language resources and tools covering, in a unified framework, a bigger variety of language phenomena in a large number of languages, including low-resourced and endangered ones.

About UniDive

UniDive, is an interdisciplinary scientific network devoted to universality, diversity and idiosyncrasy in language technology.

Its main objective is to **reconcile language diversity with rapid progress in language technology**.

It embraces both inter- and intra-language diversity, i.e. a diversity understood both in terms of the differences among the existing languages and of the variety of linguistic phenomena exhibited within a language

It gathers about **250 interdisciplinary experts** (linguists, computational linguists, computer scientists, psycholinguists, and industrials) from almost 40 COST countries (36 of which are COST countries)

It represents dozens of languages from many different language genera.

Basic Data

- Duration:
  23 September 2022 – 22 September 2026
- Official COST webpage:
  https://www.cost.eu/actions/CA21167/
- Detailed objectives and workprogram: to be found in the Memorandum of Understanding
- 4 Working Groups
- **We welcome new members**

Ways of Participation

- Become UniDive member
- Training schools
- STSM: Short Term Scientific Missions
- Joint publications
- Grants for young researchers

Working Groups

WG1: Corpus Annotation

Annotated corpora constitute the Action's major operational tools for NLP-applied universality. Therefore, WG1 is dedicated to the following activities:

- **Studies** and community **discussions** in language typology and language universals at the level of morphology, syntax and semantics, with special attention paid to idiosyncrasy at all these levels.
- Unification and enhancement of cross-lingual annotation **guidelines** for morpho-syntax and MWEs.
- Coordinate the development and maintenance of centralized **software** for universality-based corpus construction.
- Defining **file formats** for corpora annotated according to the unified guidelines.
- Construction of **annotated corpora**: adapting the existing corpora to the enhanced guidelines, creating new annotated corpora following the enhanced guidelines.

WG2: Lexicon-corpus interface

In the context of a quest for diversity, electronic lexica are complementary to corpora because they aim at holistic language modelling, describing possibly many linguistic objects, whereas in corpora many phenomena occur rarely or never. Lexica can also be useful in unifying terminologies, e.g., when a category can be described as a closed word list. In this context WG2 is dedicated to:

- Cross-language **unification of lexical features**:
  - harmonizing the definition of a “syntactic word” across languages,
  - harmonizing lemmatization rules (for words and MWEs) and lexical features across languages,
  - standardizing lists of lexemes for auxiliaries, pronouns and determiners.
- **Design** of a lexicon-corpus **interface** aiming at:
  - interlinking MWE lexicon entries with their occurrences in corpora,
  - cross-lingually unified lexicography of idiosyncratic **constructions**.
- Proof-of-concept lexical encoding of MWEs following the above design.

WG3: Multilingual and cross-lingual language technology

Unified modelling helps solve NLP tasks with higher accuracy and better awareness of diversity. Therefore, this WG is dedicated to NLP coordinating the development of tools leveraging universality and promoting diversity:

- Multilingual and cross-lingual **syntactic parsers** which:
  - pay attention to hard and underrepresented phenomena (unbounded dependencies, MWEs)
  - leverage transfer of annotations or models in order to cope with data scarceness.
- Prototypes of multilingual and cross-lingual **semantic parsers** which:
  - derive bi-lexical semantic dependencies from syntactic trees,
  - resolve idiosyncrasies in the syntax-semantics interface.
- Multilingual **MWE discovery** tools which:
  - exploit large non-annotated data to compensate the sparseness of MWEs in annotated corpora,
  - are coupled both with lexicons and MWE identifiers.
- Multilingual **MWE identifiers** which:
  - are coupled with MWE discovery and lexica to better handle unseen data,
  - pay attention to underrepresented phenomena, e.g., discontinuity, variability of MWEs.
- Prototypes of tools for automatic identification of **idiosyncratic constructions**.

WG4: Quantifying and promoting diversity

This WG is transversal to WGs 1-3 and will focus on how the Action serves inter- and intra-linguistic diversity. Its activities will overlap with the 3 other WGs in:

- **Networking** for diversity:
  - bringing together pre-existing groups dedicated to NLP-applicable universality, integrating experts of notably low-resourced languages not yet covered by these groups, integrating experts in linguistic typology.
- **Quantifying** diversity:
  - designing measures of inter- and intra-linguistic diversity in language resources and tools, using these measures to quantify diversity in UD and PARSEME corpora.
- **Promoting** diversity:
  - procedures for **better use of the existing resources**, based on their estimated diversity, **selecting new data** to be annotated, designing **evaluation scenarios** which favour tools performing well on rare and diverse phenomena and low resourced languages,
  - **validating** the unified annotation **guidelines** (WG1) **and lexicon formats** (WG2) against newly included languages and defining new language-specific categories and extensions,
  - coordinating the creation and enhancement of annotated **corpora** and **lexica** for low-resourced languages,
  - **discovering** and analysing **rare** linguistic **phenomena**, and describing them in resources and tools,
  - coordination of the development of NLP **tools** (WG3) for low-resourced and endangered languages.

Funded by the European Union

Contact: Gediminas Navickas, gediminas.navickas@mif.vu.lt