Brussels, 27 May 2022

COST 081/22

DECISION

Subject: Memorandum of Understanding for the implementation of the COST Action "Universality, diversity and idiosyncrasy in language technology" (UniDive) CA21167

The COST Member Countries will find attached the Memorandum of Understanding for the COST Action Universality, diversity and idiosyncrasy in language technology approved by the Committee of Senior Officials through written procedure on 27 May 2022.

MEMORANDUM OF UNDERSTANDING

For the implementation of a COST Action designated as

COST Action CA21167
UNIVERSALITY, DIVERSITY AND IDIOSYNCRASY IN LANGUAGE TECHNOLOGY (UniDive)

The COST Members through the present Memorandum of Understanding (MoU) wish to undertake joint activities of mutual interest and declare their common intention to participate in the COST Action, referred to above and described in the Technical Annex of this MoU.

The Action will be carried out in accordance with the set of COST Implementation Rules approved by the Committee of Senior Officials (CSO), or any document amending or replacing them.

The main aim and objective of the Action is to reconcile language diversity with rapid progress in language technology. This will be achieved through the specific objectives detailed in the Technical Annex.

The present MoU enters into force on the date of the approval of the COST Action by the CSO.

OVERVIEW

Summary
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. Its approach is to: (i) pursue NLP-applicable universality of terminologies and methodologies, (ii) quantify inter- and intra-linguistic diversity, (iii) boost and coordinate universality- and diversity-driven development of language resources and tools. 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.

Areas of Expertise Relevant for the Action	**Keywords**
● Computer and Information Sciences: Artificial intelligence, intelligent systems, multi agent systems ● Languages and literature: Linguistics: formal, cognitive, functional and computational linguistics ● Computer and Information Sciences: Machine learning algorithms ● Languages and literature: Linguistics: typological, historical and comparative linguistics	● natural language processing ● language universals ● diversity ● idiosyncrasy ● language resources and tools

Specific Objectives
To achieve the main objective described in this MoU, the following specific objectives shall be accomplished:

Research Coordination
● To develop methods for quantifying inter- and intra-linguistic diversity.
● To develop a common understanding of language universals across 70 languages represented in the Action.
● To coordinate the diversity-driven creation, merging and enhancement of language resources unified across over 100 languages from the UD and PARSEME collections.
● To coordinate efforts towards a better coverage of inter-/intra-linguistic diversity in NLP tools.
● To raise awareness of the international community about the importance of diversity preservation in language technology.
● To disseminate the Action outcomes to stakeholders.

Capacity Building
● To create a network of experts in a large number of languages working on modelling and processing of

morphological, syntactic and semantic phenomena within a common framework.
● To foster the capacities of Young Researchers and Innovators (YRIs), with special focus on COST ITC participants.
● To coordinate and boosting universality-driven initiatives worldwide.
● To set up a long-term roadmap for the joint efforts of the universality-driven NLP community.

EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

TECHNICAL ANNEX

1. S&T EXCELLENCE

1.1. SOUNDNESS OF THE CHALLENGE

1.1.1. DESCRIPTION OF THE STATE OF THE ART

Diversity of naturally occurring phenomena is a vital heritage to be preserved in the current progress- and optimization-driven globalization era. This includes **language diversity**, understood both in terms of the number of existing languages and of the variety of linguistic phenomena within a language. At the same time, the competitiveness of European markets within the global economic landscape must rely on efficient information access, largely supported by recent advances in **data science**. The majority of the available data, especially those produced by and available to the large public, are **language data**. Human language (technically called *natural language*) is also the most natural and inclusive human-machine interface. This is why **language technology**, also called **natural language processing** (NLP), is one of the most flourishing data science branches nowadays and truly influences citizens' lives via applications such as machine translation, information retrieval, speech recognition, speech synthesis, computer-aided language learning, communication support for the impaired and many others.

INTER-LANGUAGE DIVERSITY

Language diversity, so pervasive in Europe, has often been seen as an obstacle to rapid progress in NLP. For decades, we have been witnessing an overwhelming **dominance of English** (Bender 2011) as the main language of study and, to a lesser extent, of a few other languages spoken in occidental societies. As a result, many models and methods were either suboptimally or not at all adapted to the remaining, mostly morphologically rich, languages, even those with large numbers of speakers.

This state of the art is now changing. NLP teams in many countries, especially in Europe, work on the national or regional languages. Language resources, backbones of linguistic research and of data-driven NLP methods, are being developed for an increasing number of languages. New algorithms make it possible to process very large amounts of textual data in language-agnostic ways, thus the cost of their adaptation to new languages is relatively low. Software implementations of these algorithms are publicly available and offer simple interfaces, which promote fast prototyping. Large hardware infrastructures with Big-Data-driven architectures offer unprecedented performances. Leading text processing companies provide multilingual NLP support. Medium and small enterprises also become aware of the added value from multilingual text processing and integrate NLP modules in their products. **Multilingual NLP is truly booming**, to the benefit of many end users: citizens, private and public services, etc. Despite this rapid progress in multilingual NLP, preserving language diversity remains a challenge. Firstly, most languages suffer from **data scarceness**, i.e., they are **low-resourced** or undocumented. State-of-the-art approaches are data-driven and as such rely on large amounts of data on which to train NLP tools. Therefore, these tools perform poorly for low-resourced languages (Wu & Dredze 2020) and their speakers are disadvantaged. This leads to digital extinction, increasing inequalities and switching to a well-supported dominant language, which undermines **endangered languages** even more. Secondly, diversity is also endangered within individual languages, as discussed in the next section.

INTRA-LANGUAGE DIVERSITY

Most linguistic phenomena are known to have a so-called **Zipfian distribution**, i.e., few occur frequently in texts and there is a long tail of those that rarely occur. Modern NLP algorithms strongly favour the former and often underperform in the latter, since they are conceived and tuned for optimal global performances in which the former dominate. This sensitivity of NLP to **data sparseness** creates a bias: phenomena which are suboptimally covered by technology might also be gradually abandoned by the speakers. Since such cases constitute the majority of the existing phenomena, **diversity within a language** is endangered. Very few initiatives were dedicated to **quantifying** and addressing this problem in NLP resources and tools (Derczynski et al. 2017, Ramisch 2020).

Examples of such phenomena include syntactic structures such as **unbounded dependencies**, in which syntactically close words are linearly distant in a sentence, as in *they hope to have a sheaf of documents both sides can trust*. Such utterances are fluently produced and understood by speakers but under-represented in corpora, and NLP tools have been underperforming on them (Rimell et al 2009).

Other examples of phenomena suboptimally covered by NLP are idiomatic expressions, also called **multiword expressions** (MWEs) such as *by and large* or *to take a haircut*. They are combinations of words exhibiting unexpected behaviuor, often **idiosyncratic**, i.e. specific to few objects. Most prominently, they are semantically non-compositional: their meaning cannot be deduced from the meanings of their components and from their syntactic structure in a regular way. For instance, *to take a haircut* means 'to partially forgive a debt', which has few explicit links with *take* and *haircut.* Thus, NLP applications may wrongly handle MWEs, e.g., Google translates "*lenders take 90% haircut"* into French by "*les prêteurs se coupent à 90%*" (*lenders are cut by 90%...*). If the MWE *to take a haircut* were identified in the source text and translated as a single unit of meaning, errors of this kind could be avoided. Development of MWE identifiers was recently boosted but the results show that they badly fail on rarely occurring data. This implies that account of MWE diversity in NLP technology is endangered (Savary et al. 2019).

MWEs belong to the larger phenomenon of idiosyncratic **constructions**, i.e., conventional associations of forms (whether lexical, syntactic, or pragmatic) and meanings, such as the-Adjective-the-Adjective constructions (*the more the merrier*, *the higher the better*, etc.). Idiosyncrasy in such objects is pervasive but very hard to capture, and relatively few efforts have been put into explicating their nature and covering them in NLP. Since MWEs and idiosyncratic constructions exhibit a Zipfian distribution (Williams et al., 2015), a large majority of them suffer from the diversity issues described above.

FRAGMENTATION, UNIVERSALITY AND DIVERSITY

Additional obstacles come from **fragmentation** of theories, terminologies and methods. It is more severe in linguistics and NLP than in other domains. Namely, linguistic phenomena appear “on the surface” in language-specific ways and are hard to compare across many languages, simply because an expert can only reason in terms of (a few) languages she/he is familiar with. Thus, different traditions often name similar phenomena differently, or, conversely, use the same term to name distinct phenomena. For instance, the notion of 'subject' is very difficult to pin down, both inter- and intra-linguistically (Givón, 1995). Also, theories and methods bear implicit restrictions, appropriate for some languages but not for others (e.g., free versus strict word order). As a result, revealing true commonalities is challenging and the extent of universals in language (if any) is one of the most difficult and intriguing research questions.

The study of **language universals** has a long-standing tradition (Greenberg, 1966; Chomsky, 1976), prevails in mainstream theoretical linguistics and is a central issue in **typology**. But the existence of absolute universals is a subject of a major controversy. Evans & Levinson (2009) claim that the existence of a Universal Grammar is an English-centred myth, that strong statistical tendencies ("statistical universals") should be considered instead and that linguistic research should rather use diversity as a starting point. Others argue that diversity is a surface phenomenon and its description resembles "citing a collection of exotica" (Tallerman, 2009), while universality, conversely, can be captured at the right level of abstractness. In NLP, researchers are more agnostic towards the theoretical status of language universals, rather emphasizing the usefulness of cross-linguistically consistent and applicable language descriptions. The objective of defining such descriptions will be referred to, within this Action, as ***universality***, while statistical universals will simply be called (*language*) *universals*. Three NLP initiatives have recently engaged in universality: Universal Dependencies (UD, Nivre et al. 2020), COST Action IC1207 PARSEME (Savary et al. 2018, Ramisch et al., 2020) and UniMorph (Kirov et al. 2018). They have been working relatively independently so far, jointly covering three levels of linguistic modelling: morphology, syntax and (partly) semantics. The outcomes of these efforts include:

- Multilingual **lexica**, where words and expressions receive morphosyntactic categories (noun, verb, idiom, etc.) and values (plural, past tense, modifiable, etc.), taken from unified repositories of labels.
- Multilingual **annotated corpora**, i.e. large collections of texts in which words are assigned categories and values, and particular phenomena such as MWEs are identified. If, additionally, sentences are structured into syntactic trees, such corpora are called treebanks. Annotated corpora are major operational tools for language modelling and backbones of data-driven NLP methods. UD and PARSEME have released an impressive number of 202 treebanks for 114 languages so far.

Unified modelling obviously promotes genericity, which should result not only in a better understanding of cross-language phenomena but also in more efficient, maintainable and adaptable technological solutions. At first glance, **universality** might be perceived as opposed to **diversity** because it promotes convergence of different points of view on similar phenomena. In fact, however, the contributions of universality-driven initiatives to preserving and promoting diversity are manifold.

- Their shared objective is to represent in a unified way only those phenomena which are truly similar, thus emphasizing those which are specific to particular languages. Therefore, UD and PARSEME typologies, for instance, leave room for language-specific categories and relations.
- Some idiosyncratic (i.e., rare and diverse) phenomena in a language (e.g. noun-adjective constructions in English, such as *attorney general*) are remnants of other languages' influence (here: French) and are easier to understand in a contrastive framework.
- Unified methodologies promote **inclusiveness** because they lead to the construction of shared frameworks in which new experts are easily integrated and can contribute in a grassroots manner. UD and PARSEME offer centralized collaborative infrastructures for the development of unified annotation guidelines and annotated corpora. As a result, these initiatives cover dozens of languages, many of which are low-resourced. For instance, considering the UNESCO Atlas of Languages in Danger, UD treebanks currently exist for at least 35 endangered or vulnerable languages (Akuntsu, Apurinã, Assyrian, Basque, Belarusian, Breton, Buryat, Chukchi, Erzya, Faroese, Guajajara, Irish, Ka'apor, Kangri, Karelian, Khunsari, K'iche', Komi Permyak, Komi Zyrian, Livvi, Low Saxon, Makurap, Manx, Mbyá Guaraní, Moksha, Mundurukú, Nayini, North Sámi, Scottish Gaelic, Skolt Sámi, Upper Sorbian, Warlpiri, Welsh, Western Armenian, and Yupik), 12 extinct languages (Akkadian, Ancient Greek, Classical Chinese, Coptic, Gothic, Latin, Old Church Slavonic, Old East Slavic, Old French, Old Turkic, Sanskrit, and Tupinambá), and over 20 which are underrepresented in NLP, even if demographically strong (e.g. Albanian, Armenian, Icelandic, Kazakh, and Maltese, only to cite officially recognized languages spoken in Europe).
- The lexica and corpora created by universality-driven initiatives are distributed under **open licenses**, which also promotes inclusiveness.

- Unified typologies and formats facilitate the development of **multilingual** tools, in which universal algorithms are adaptable to different languages by sheer selection of the input data. This further promotes inclusiveness, because NLP tools are created for languages with no or few NLP experts.
- Unified modelling strengthens opportunities in **cross-lingual** NLP, where the construction of tools for a given language can benefit from resources in other typologically related languages. Examples of such techniques are: (i) annotation transfer, where annotations are mapped from one language to another using parallel corpora (Yarowsky et al., 2001; Hwa et al., 2005), (ii) model transfer, where a model trained on one or more languages is applied directly to a new language, relying on universal morphosyntactic categories (Zeman and Resnik, 2008; McDonald et al., 2011; Agić et al., 2014) or on similarity of syntactic trees (Ponti et al., 2018), (iii) corpus sampling for an optimal typological variety (de Lhoneux, 2017). This contributes to overcoming data sparseness and scarceness, and strengthens inclusiveness, because tools can be created even for languages with very little or no data.
- Universality also benefits from recent deep learning models, which are trained on large quantities of multilingual **raw** (i.e., non-annotated) **texts** in language- and task-agnostic ways (Devlin et al. 2019). Such models are expected to implicitly encode knowledge about language in general (including statistical universals). These models can then be fine-tuned for a particular language and task with small quantities of annotated data, and even used for languages with no annotated data (Pires et al., 2019). Thus, low-resourced languages benefit from large multilingual data. Also, the intra-linguistic diversity is higher since raw data are by several orders of magnitude larger than annotated data.

Additionally to this strong impact of universality on diversity, the opposite influence also takes place, as stressed by Evans & Levinson (2009, p. 432): *[According to Ethnologue,] 82% of the world's 6,912 languages are spoken by populations under 100,000, 39% by populations under 10,000, [...] 8% (are) nearly extinct, and a language dies every two weeks. This loss of diversity, as with biological species, drastically narrows our scientific understanding of what makes a possible human language.*

1.1.2. DESCRIPTION OF THE CHALLENGE (MAIN AIM)

Given the above state of the art, this Action will address the following main challenge:

To reconcile language diversity with rapid progress in language technology

This challenge is very relevant and timely. Language technology is now truly booming, which brings unprecedented multilingual support to end users. But with this technological effervescence, societies run the risk of important losses in diversity. Endangered diversity is known to be a major risk in domains of life studied by biology, genetics, medicine (Forschungsverbund Berlin 2018), sociology (Phillips 2014), etc. Linguistic diversity is closely connected to these aspects and should be regarded, from a holistic perspective, as part of biocultural diversity, as put by the Terralingua initiative.

This Action adopts two original points of view on diversity. Firstly, it aims at coping with data scarceness and sparseness, i.e., at promoting both inter- and intra-language diversity, the latter being understood as the richness of linguistic phenomena exhibited within a language. Secondly, it does not consider that linguistic diversity is to be protected *against* technological progress but pursues both of these aims jointly, to their mutual benefit. This approach is now possible due to the accumulation of the critical mass of NLP outcomes (§1.1.1): (i) cross-lingually unified methodologies and resources from UD, PARSEME and UniMorph, (ii) annotation and model transfer and fine-tuning methods which allow us to capitalize both on huge quantities of non-annotated data and on decent quantities of unified annotated data available for many languages in order to boost performances for a single language, (iii) methods addressing quantitatively minor but qualitatively major linguistic phenomena, such as discontinuous syntactic and semantic dependencies (Rohanian et al. 2019). These outcomes provide a proof of

concept that universality-driven language technology simultaneously boosts language diversity and global efficiency of multilingual NLP.

However, recent deep learning models still underperform for low-resourced languages and call for more annotated data for fine-tuning (Wu & Dredze 2020). Such data are being created for an increasing number of languages, notably by UD, PARSEME and UniMorph, but their emerging standards are partly incompatible and competing. For instance, there is currently no satisfactory way to model expressions which have regular syntax but idiosyncratic semantics (i.e. a vast majority of MWEs) because a clean separation between syntactic and semantic annotation is missing due to UD and PARSEME addressing these issues separately. The UD-UniMorph comparison lists 25 differences in the two morphological vocabulary standards. UD has 3 MWE categories competing with those defined by PARSEME. To unify these standards and overcome other fragmentation issues (§1.1.1), an inclusive framework such as the one provided by a COST Action, open to the largest possible number of countries, is needed.

While Europe is the continent covering the smallest percentage of the currently existing languages (cf. Ethnologue), it hosts increasingly many speakers from other continents. In Europe multilingualism has been receiving continuous political support (see e.g. the fact sheet on Language Policy of the EU, or the Language Equality in the Digital Age report). Importantly for this Action, Europe benefits from the most multilingual NLP community (as compared e.g., to North America or China). Most European countries have their own NLP teams developing language technology for their national or minority languages (e.g., Basque and Irish). European NLP initiatives and infrastructures are dedicated to coordinated efforts in favour of multilingualism (§2.1.1). As a result, Europe is the major driving force in multilingual NLP worldwide, also to the benefit of languages spoken outside the European continent.

The challenge of reconciling language diversity with technological progress is also politically timely. UNESCO proclaimed 2019 the International Year of Indigenous Languages, and stressed the need of “a noble language policy [...] so that languages are not used for the purposes of domination, suppression and separation”. Special attention is paid to multilingualism in cyberspace (via IFAP) and to accounting for language variety under COVID. Our initiative actively contributes to these global aims.

1.2. PROGRESS BEYOND THE STATE OF THE ART

1.2.1. APPROACH TO THE CHALLENGE AND PROGRESS BEYOND THE STATE OF THE ART

The challenge of reconciling language diversity with rapid progress in language technology is approached by the Action via three measures: (i) **NLP-applicable universality of terminologies and methodologies,** (ii) **quantifying inter/intra-linguistic diversity,** (iii) **universality- and diversity-driven development of language resources and tools** for both low- and well-resourced languages (here, a low-resourced language is understood as one for which the available annotated data are judged insufficient to obtain state-of-the-art performances in the NLP task at hand, cf. Duong 2017; the targeted tasks are listed in §4.1.1, WG3). This approach is sound, despite apparent opposition between universality and diversity, because the former is, by nature, simply conditioned by the latter. Namely, verifying the universal validity of models implies looking at all (or, more realistically, as many as possible) languages and phenomena. The current assets from universality truly pave the way for the promotion of inter- and intra-linguistic diversity (§1.1.1). Also, simultaneous study of both low- and well-resourced languages is relevant. The former are direct targets of inter-language diversity efforts. The latter's large resources are more representative of language phenomena (the Zipfian tail is shorter) and better suited for characterizing intra-language diversity. However, major obstacles still remain to be overcome, in

order to achieve balance between diversity and technological progress. The Action will advance the state of the art in this field by:

- Defining **measures** of inter- and intra-linguistic diversity in language resources and tools.
- Advancing the theoretical **debate over language universals** to the level of applying empirical measures of diversity rather than just "citing a collection of exotica".
- Overcoming the **fragmentation** issues through:
  - Contrastive studies aimed at spotting similarities in unrelated or geographically remote languages;
  - Discovery and validation of hypothesized language universals, and of their NLP-applicable modelling, throughout over 70 languages covered by the Action participants;
  - Unifying the already existing terminologies and methodologies, such as annotation guidelines, morpho-syntactic tagsets and MWE categories, put forward by universality-driven initiatives;
  - Extending these terminologies and methodologies to yet uncovered phenomena and confronting them with a larger number of languages;
  - Developing better criteria for applying unified guidelines to specific languages;
  - Increasing the cross-lingual consistency of the already existing language resources;
  - Unifying the resources which currently separately model morphology, syntax and MWEs;
  - Developing tools for faster corpus annotation, unification and merging.
- Increasing **inter-language diversity** via a better NLP support for low-resourced languages:
  - Enlarging language resources for at least 24 low-resourced languages (Akuntsu, Chukchi, Basque, Bulgarian, Frisian, Greek, Guajajara, Hindi, Hungarian, Irish, Javanese, Ka'apor, Lithuanian, Makurap, Maltese, Manx, Mundurukú, Odia, Romanian, Serbian, Slovene, Swedish, Tagalog, Turkish) from 17 language genera (according to WALS);
  - Creating language resources for at least 14 not yet covered endangered/extinct languages (Abaza, Cusco Quechua, Georgian, Hittite, Kabyle, Karo, Ligurian, Maghrebi Arabic, Neapolitan, Occitan, Old Irish, Laz, Xibe, Yakut);
  - Designing evaluation scenarios which favour tools performing well on low-resourced languages;
  - Developing high-quality NLP tools for low-resourced and endangered languages, based on transfer/fine-tuning of annotations/models from well-resourced ones.
- Increasing the **intra-linguistic diversity** in NLP resources and tools through:
  - Measuring the intra-linguistic diversity in the UD and PARSEME resources (for over 100 languages);
  - Making better use of the existing resources, by sampling or split, based on their estimated diversity;
  - Improving the selection of new data to be annotated, so as to favour intra-linguistic diversity;
  - Defining more accurate cross-linguistically valid annotation principles for diverse phenomena;
  - Developing NLP tools which advance the state of the art in terms of diversity;
  - Driving the NLP community towards diversity-based evaluation, e.g., via evaluation campaigns.

This approach is original for several reasons. Firstly, it considers both intra- and inter-linguistic diversity. Secondly, it engages in NLP-applicable universality at an unprecedented scale, since morphological, syntactic and (part of) semantic phenomena are covered in the same framework. Thirdly, it pays special attention to idiosyncrasy, which crosses all these levels of linguistic modelling, and it undertakes modelling of idiosyncratic constructions, which have been rarely addressed, especially in multilingual frameworks. Finally, it questions the mainstream viewpoint on evaluation of NLP performances, in that statistically underrepresented phenomena, crucial for diversity, are brought to light.

1.2.2. OBJECTIVES

1.2.2.1 Research Coordination Objectives

To respond to its Challenge, the Action will pursue the following research coordination objectives.

- [RCO1] To develop methods for quantifying inter- and intra-linguistic diversity.
- [RCO2] To develop a common understanding of language universals across 70 languages represented in the Action.
- [RCO3] To coordinate the diversity-driven creation, merging and enhancement of language resources unified across over 100 languages from the UD and PARSEME collections.
- [RCO4] To coordinate efforts towards a better coverage of inter-/intra-linguistic diversity in NLP tools.
- [RCO5] To raise awareness of the international community about the importance of diversity preservation in language technology.
- [RCO6] To disseminate the Action outcomes to stakeholders (§2.2.2).

1.2.2.2 Capacity-building Objectives

Language universality, by nature, can only be pursued in a multilingual, inclusive and coordinated framework. Therefore, the capacity-building objectives of this Action include the following:

- [CBO1] To create a network of experts in a large number of languages working on modelling and processing of morphological, syntactic and semantic phenomena within a common framework.
- [CBO2] To foster the capacities of Young Researchers and Innovators (YRIs), with special focus on COST ITC participants.
- [CBO3] To coordinate and boosting universality-driven initiatives worldwide (also outside Europe).
- [CBO4] To set up a long-term roadmap for the joint efforts of the universality-driven NLP community.

2. NETWORKING EXCELLENCE

2.1. ADDED VALUE OF NETWORKING IN S&T EXCELLENCE

2.1.1. ADDED VALUE IN RELATION TO EXISTING EFFORTS AT EUROPEAN AND/OR INTERNATIONAL LEVEL

A number of current and recent initiatives are relevant to the Action.

1. International networks and initiatives

- Universal Dependencies (UD) – open community effort with over 300 contributors who address cross-linguistically consistent grammatical annotation and have already produced more than 200 treebanks in over 100 languages; unfunded (except for local funding of the individual contributors).
- UniMorph – a collaborative effort (coordinated by US experts) to improve how NLP processes complex morphology in the world's languages, particularly those which are low-resourced; unfunded.
- NewNLP – US-EU initiative for supporting NLP technology in low-resourced languages and dialects.
- LinGO Grammar Matrix and CoreGram – projects focused on implementing common core grammar components for various languages in order to empirically test language universals.

2. European networks and projects

- IC1207 PARSEME (2013-2017) – COST Action on MWEs and parsing, with 200 researchers from 31 countries.
- CA18209 NexusLinguarum (2019 – 2023), CA18231 Multi3Generation (2019 – 2023), CA19102 LITHME (2020 – 2024) - ongoing COST Actions on interoperable multilingual data, language generation, and linguistics-technology dialogue.
- ELEN and Elexis - European Language Equality Network and H2020 lexicographic infrastructure.
- Embeddia (2019 – 2022) – European H2020 project dedicated to the use of cross-lingual embeddings coupled with deep neural networks to allow existing monolingual resources to be used across languages.
- European Language Grid (2019-2020) – European H2020 platform for language resources and tools, drawing notably upon META-NET and CLARIN.
- Gourmet (2019-2021) – European H2020 project aiming at using and improving neural machine translation for low-resourced language pairs and domains.
- Prêt-à-LLOD (2019-2022) – European H2020 project dedicated to combining language technologies with Linguistic Linked Open Data (LLOD), and to creating ready-to-use multilingual data.
- AELAW (2015-2019) – COST Action dedicated to ancient European languages and writings; it notably produced a database of existing ancient inscriptions.
- ReLDI and CLASSLA – cross-border network of researchers and professionals, and a CLARIN knowledge centre, dedicated to south-Slavic language technology.
- LEAD-ME (2020-) - COST Action on media accessibility in Europe.

3. National networks and projects

- Czech Republic: LINDAT/CLARIN – research infrastructure for sharing language data and tools; it has a large open repository of corpora in over 120 languages, including the UD and PARSEME corpora; it also provides on-line services for corpus query, syntactic parsing and translation.
- France: PARSEME-FR (2016-2021), SELEXINI (2022-2026), AUTOGRAMM (2022-2026) and CREAM (2021-2025) – projects on MWEs and semantic lexicons for French, tools and resources for low-resourced languages.
- France: LIFT (Linguistique Informatique, Formelle et de Terrain) – a special interest group on NLP, formal linguistics and linguistic fieldwork; covers language universals and low-resourced languages.
- Germany: TuLaR – an initiative for the development of language resources for 79 Tulpian languages.
- Greece: Filotis (2021-2023) – EU- and Greece-funded project on creating language technology for Pomak, a minority language in Bulgaria, Greece and Turkey.
- Norway: INESS – Norwegian corpus infrastructure with: (i) online corpus querying in 80 languages, including the UD corpora, (ii) documentation on how MWEs are annotated and queried in treebanks.
- Portugal: VOC.CPLP – lexicons for 9 varieties of Portuguese, including the African dialects.
- Slovenia: New Slovene Grammar and DSDE with WPs on morphology and MWEs.

4. International events

- UnLId – Dagstuhl Seminar on the universals of linguistic idiosyncrasy in multilingual computational linguistics, (30-31 Aug 2021), Dagstuhl, Germany
- MWE – Workshop on Multiword Expressions, co-located with major NLP venues; 17th ed. in 2021
- UDW – Universal Dependencies Workshop co-located with international venues; 5th edition in 2021
- TLT – Workshop on Treebanks and Linguistic Theories, mainly in Europe, 20th edition in 2021
- SLTU/CCURL - Workshop on computational techniques for low-resourced languages
- CLTW - Celtic Language Technology Workshop

- Evaluation campaigns (called shared tasks) for NLP tools organized by:
  i. UD: CoNLL 2017 and 2018 Shared Task on Multilingual Parsing from Raw Text to Universal Dependencies; 71 languages addressed in 2018
  ii. PARSEME: PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions in 2017, 2018 and 2020; 20 languages addressed in 2018
  iii. UniMorph: CoNLL-SIGMORPHON Shared Task in 2016-2021; 100 languages since 2019.

Despite this recent proliferation of initiatives dedicated to multilingual NLP, universality, diversity and idiosyncrasy issues are not jointly addressed in any of them. Thus, this Action will bring about an added value to this rich body of research and networking activities by:

- Drawing upon the experience of Universal Dependencies and PARSEME, whose main outcomes (annotation guidelines, annotated corpora, tagsets, publications, etc.) are openly available.
- Providing unified versions of the UD and PARSEME annotation guidelines, which are currently partly incompatible. To this aim, results of the UnLId Dagstuhl Seminar will be exploited.
- Taking steps towards unifying morphological annotation principles with UniMorph.
- Using NewNLP instruments and distributing its own outcomes via NewNLP infrastructure.
- Supporting LinGO Grammar Matrix and CoreGram with experimental data.
- Organizing shared tasks for joint parsing and MWE identification and discovery.
- Supporting the objectives of ELEN with technological solutions.
- Collaborating with Elexis for the design of unified lexicon-corpus interfaces.
- Using the UDW, MWE, TLT, SLTU/CCURL and CLTW workshops as dissemination instruments.
- Addressing fundamental scientific issues orthogonal to technological progress, which are not addressed by the on-going and starting COST Actions and European projects: revealing and understanding of language universals and their computationally tractable modelling; focusing on underrepresented language phenomena, challenging for NLP and crucial for diversity; extending unified terminologies and methodologies to cover a broader range of constructions.
- Addressing objectives similar to those of NexusLinguarum but in a complementary way. While this Action deals with unified modelling of diverse linguistic regularities and idiosyncrasies in typologically diverse languages, NexusLinguarum's focus is not on the language phenomena themselves but on how to represent them in interoperable formats, easy to interlink with the Semantic Web.
- Contributing to the European Language Grid by releasing its outcomes under open licenses via the CLARIN infrastructure. In particular, it will enrich the LINDAT/CLARIN and INESS platforms with new versions of annotated corpora, based on unified syntactic and semantic categories. This will ensure availability and sustainability of the Action's outcomes beyond its funding duration.
- Ensuring liaison with on-going and starting COST Actions and European projects by: regularly informing the respective coordinators on the Action's programme and progress; co-organising common events such as workshops within international conferences of the domain; promoting the Action's Short-Term Scientific Missions (STSMs) and Training Schools among Young Researchers and Innovators included in these projects and Actions.
- Inviting members from AUTOGRAMM, Filotis, INESS, LIFT, New Slovene Grammar, SELEXINI, TuLaR ReLDI, and VOC.CPLP. It will bring international networking added value to these nationally funded research initiatives. It will also trigger and endorse new national projects.

2.2. ADDED VALUE OF NETWORKING IN IMPACT

2.2.1. SECURING THE CRITICAL MASS AND EXPERTISE

Universality, pursued by the Action, requires a participation of experts of as many languages as possible. The network will ensure a close-to-optimal critical mass that could possibly be achieved within an

international funded research framework. The Action will benefit from the initial network of proposers, which included over 70 experts from 30 COST countries, 3 Near-Neighbour Countries (NNC), and 4 International Partner Countries (IPC). Together, their skills covered over 70 distinct languages. This is a strong basis for networking and collaboration, given that joint knowledge of a rare language, linguistics and technology is a niche expertise. Additional experts will join Action during its lifetime, with the aim of covering additional languages (endangered or low-resourced), as well as involving experts in typology, who have a global understanding of large groups of languages, and experts from industry (large companies and SMEs) who explicitly address multilingualism and low-resourced languages. Contributions, especially for low-resourced languages, from PhD students supervised by the Action's members are also envisaged.

This substantial knowledge of experts in dozens of languages gathered around a common framework is still far from covering all the existing languages, whose number is estimated at over 7,000 according to Enthnologue, over 600 of which are also written, and whose majority is spoken outside Europe. Note, however, that the outcomes of the Action include centralized, collaborative and open infrastructures in which external experts can be integrated remotely. Therefore, the Action will experiment with open calls for participation addressed to experts of yet uncovered languages, to whom training and technical support will be offered. Video-conference or video recording will be used to give such remote members opportunities of attending Working Group meetings and Training School courses.

As to the critical mass of the input data, the Action builds upon two previous initiatives, Universal Dependencies and PARSEME, and reaches out to a third (mainly US-coordinated) one, UniMorph. These three groups produce and maintain the largest multilingual resources with unified annotation to date, on the level of morpho-syntax and of MWEs. These resources will be merged and enlarged within the Action, and made fully available under open licenses.

Finally, concerning the necessary software, there is a rich updated body of works in cross-lingual annotation or model transfer and fine-tuning (§1.1.2). They provide a proof of concept of the mutual benefit between universality and diversity in language technology. The Action will benefit from these (nationally funded) assets since it involves experts developing precisely these kinds of software.

2.2.2. INVOLVEMENT OF STAKEHOLDERS

The Action will directly involve the following categories of stakeholders:

- **Experts in theoretical linguistics** and **typology**, thanks to the unified terminologies and methodologies for language modelling, will gain a better understanding of language universals, and of truly language-specific phenomena. They will also be given testbeds for empirically assessing inter-linguistic similarity and understanding if it stems from common origins or mutual influence.
- **Experts in NLP** will benefit from the opening of the NLP domain to inter- and intra-linguistic diversity, and from increased inclusiveness. This will be achieved via a direct integration of the researchers in the Action and via the Action's dissemination.
- **Young Researchers and Innovators** (YRIs). The objectives and the intended workplan require long-standing efforts and efficient coordination. This implies that the transfer of competences between senior and young researchers has to be ensured. Therefore, YRIs will be given important roles in the network. Each Working Group will be coordinated by a leader and a vice-leader, one of which will be an YRI, so as to receive intensive training in coordination tasks. For Training Schools and Short-Term Scientific Missions priority will be given to YRIs. The interests of the YRIs will also be explicitly represented within the Action Core Group, which will include an elected YRI Representative.

- **Researchers working on low-resourced and endangered languages**, in and outside Europe, will benefit from: (i) direct interactions within the Action, (ii) centralized methodological and discussion platforms, with direct access to expertise from well-resourced languages, (iii) shared data validation, release, licensing, distribution and storage infrastructure, (iv) centralized unified documentation in which new languages will be relatively easy to integrate, (v) opportunities to make the unified guidelines evolve via the inclusion of language-specific phenomena, (vi) open-access releases of all resources and tools. Their interests will be directly represented by WG4 (§4.1.1).
- **Speakers of low-resourced and endangered languages** will gain technological support from NLP. Their languages will be documented by native experts in a unified multilingual framework. Multilingual NLP tools will more easily cover low-resourced and endangered languages, thanks to the unified categories. Data scarcity in these languages will be addressed by transferring knowledge gained from typologically close well-resourced languages. The awareness of language diversity will be increased due to joint distribution of outcomes for both well- and low-resourced languages. The interests of these speakers will be represented by the relevant researchers in WG4.
- **Experts in empowering low-resourced languages**, such as the Erasmus+ Digital Language Diversity Project (2015-2018), will assess and report on the "digital language vitality" progress of these languages. The risk of digital extinction is real for many languages and involvement from such experts is crucial. Members of such initiatives will be directly informed.
- **Professionals in language technology** software will be able to provide a better coverage of diverse linguistic phenomena and of a larger number of languages in their products, which until now have been uncovered. Researchers affiliated to such companies will be invited to participate.
- **Professionals in language teaching** will be able to use the diversity measures defined by the Action to adapt texts used in language teaching to the competence level of their students. The Action already includes researchers who are language teachers and will represent interests of this group.

Other stakeholders, not directly involved in the network, are targeted by the exploitation plan (§3.2.2). They include the **larger research community**, **European and worldwide institutions** (notably the European Commission and UNESCO), **speakers of low-resourced and endangered languages**, and **citizens** in and beyond Europe.

2.2.3. MUTUAL BENEFITS OF THE INVOLVEMENT OF SECONDARY PROPOSERS FROM NEAR NEIGHBOUR OR INTERNATIONAL PARTNER COUNTRIES OR INTERNATIONAL ORGANISATIONS

Since the approach taken by the Action is based on universality of terminologies and methodologies, collaboration with experts from NNC and IPC countries is an evident mutual need. On the one hand, the network cannot efficiently reason about linguistic properties and build resources for a given language without the help of experts of this language, so including them helps expand the diversity of the languages covered. On the other hand, these experts and their communities can benefit from the integration of their languages into a unified framework, notably for the reasons of inclusiveness, more direct interactions among experts of different languages (leading to more insightful linguistic descriptions) and cross-lingual performance boosting (§1.1.1.3). Some experts from NNC or IPC study languages which are already covered by European members of the network, but their participation is crucial for the excellence and visibility of the Action, since they are internationally renowned in the topics to be addressed (dependency syntax, dependency parsing, MWEs, idiosyncratic constructions, and typology). The benefit of these experts from participating in the Action is to be seen notably in extending their contributions and impact to new languages and communities. Last but not least, science as a whole will benefit from the closer dialogue established across geographical and political borders.

3. IMPACT

3.1. IMPACT TO SCIENCE, SOCIETY AND COMPETITIVENESS, AND POTENTIAL FOR INNOVATION/BREAKTHROUGHS

3.1.1. SCIENTIFIC, TECHNOLOGICAL, AND/OR SOCIOECONOMIC IMPACTS (INCLUDING POTENTIAL INNOVATIONS AND/OR BREAKTHROUGHS)

The **scientific and technological impact** expected from the Action is manifold:

- Better understanding of language universals and providing a unified practical framework which allows implementing and testing theoretical advances in this field;
- Development of language technology models able to overcome the challenges posed by the sparseness of "Zipfian tail" phenomena (§1.1.1.2);
- Establishment of *de facto* standards and best practices for the development, maintenance, reproducibility and evaluation of language resources and computational models derived from them;
- Creation of national or regional projects and spin-off initiatives coordinated by Action's participants to pursue more specific goals (e.g., the development of a corpus or tools for a particular language);
- Breakthrough results in theoretically interesting but computationally hard problems in NLP, such as automatic identification of strongly discontinuous syntactic dependencies and of MWEs;
- Long-lasting collaboration among multilingual and interdisciplinary experts from many countries;
- Driving mentalities and practices of the scientific NLP community towards:
  - Considering diversity on par with statistical efficiency, while evaluating language technology;
  - Accepting that the validation of scientific hypotheses and benchmarking of technological solutions should be systematically performed within multilingual frameworks.

The expected middle- and long-term **socio-economic impact** includes the following:

- Enhancing the quality of language technology used by the general public in everyday life, in synergy with the goals of the Digital Europe Programme on supporting a Digital Single Market;
- Enhancing technological support of multilingualism in Europe and beyond, thereby strengthening the cultural identity of the users;
- Avoiding digital extinction of low-resourced and endangered languages and increasing equality in information access by language minorities;
- Raising the awareness of inter- and intra-linguistic diversity as a common heritage to be promoted;
- Demonstrating that technological support for low-resourced languages can hardly be achieved at the national level but requires specialised degrees and programmes at the European/international level;
- Fighting against social fragmentation involving language, such as linguistic prejudice and inequality.

See also §2.2.2, where details of the Action's impact on particular stakeholders is described.

3.2. MEASURES TO MAXIMISE IMPACT

3.2.1. KNOWLEDGE CREATION, TRANSFER OF KNOWLEDGE AND CAREER DEVELOPMENT

The Action is expected to contribute to creating new knowledge in language universals at the levels of morphology, syntax and semantics. It will also demonstrate how this common understanding and modelling of language phenomena benefits both theoretical linguistics and language technology.

Dissemination and exploitation (§3.2.2) will be the main instruments of knowledge transfer towards the scientific community, institutions, industrials and the large public. Knowledge transfer will also be ensured within the Action itself, along four axes. Firstly, it will occur across various languages. Additional language experts will join the Action, benefit from centralized collaborative tools and contribute their own expertise to the unified framework. Secondly, scientific assets and soft skills will be transferred between senior researchers and YRIs (§2.2.2). Thirdly, the impact of the Action beyond its funding duration will be secured by the fact that all its major outcomes will be openly available via sustainable platforms. Fourthly, the Action expects follow-up initiatives within the Digital Europe Programme or H2020 funded projects.

The Action's impact on career development will be crucial for YRIs, due to their strong integration in the network (§2.2.2). Also, senior researchers are expected to gain new scientific and soft skills thanks to their roles in the Action. Their visibility in the community will grow and, as a result, promotion of several researchers in a middle stage of their career is expected. The Action's activities such as Training Schools will also be open to life-long training, which may positively influence some other careers.

3.2.2. PLAN FOR DISSEMINATION AND/OR EXPLOITATION AND DIALOGUE WITH THE GENERAL PUBLIC OR POLICY

The Action's **dissemination** plan aims to ensure that the Action's scientific results are disseminated among the research community. The target **audiences**, additionally to the stakeholders included in the Action (§2.2.1), include: (i) Master's and PhD students, as well as candidates to life-long training, affiliated in the research centres where Training Schools will be organised, (ii) members of research projects, other COST Actions and networks dedicated to issues related to the Action (§2.2.1), (iii) other researchers in the fields of NLP, computational linguistics, linguistics and typology.

The dissemination **means** include:

- The Action's **website**, which will be the main dissemination tool. It will contain: a description of the scientific issues, a detailed description of the objectives, the scientific programme, a regularly updated list of Action's members, the description of WGs, links to research events related to the Action's topics, links to related nationally funded initiatives, an easy contact point for potential new partners interested in joining the network, and links to all Action's outcomes (§4.1.2).
- The Action's internal **mailing lists** – one for the entire network, and one for each WG.The Action's **blog** or T**witter**– for fast spontaneous communication with external followers.
- The Action's **activities open to a wider scientific public**: Training Schools for young researchers and life-long training candidates; STSMs addressed mainly to YRIs (also outside the Action); Workshops co-located with top-tier international conferences; PhD topics related to the Action, etc.
- Electronically published open-access **proceedings** of workshops organised by the Action.
- Scientific **publications** in peer-reviewed international journals and conferences in NLP and linguistics.
- Master's and doctoral **trainings** in NLP carried out by the Action's members.
- European **infrastructures** for the dissemination of language **resources and tools**, notably CLARIN (via national nodes such as LINDAT or INESS, §2.1.1).

Invited talks at Action's meetings given by NLP **industry** experts.

The Action **exploitation** plan concerns the impact of the results of the Action on science, industry, society and governments mostly after its end. It is described in the table below.

The main trait of these dissemination and exploitation plans is **Open Access**. All major outcomes will be distributed under open licenses, which increase inclusiveness and enable derivative work, maximising applicability and sustainability. The training material, annotation guidelines, annotated corpora and lexicons will use various flavours of Creative Commons (CC). Workshop proceedings, publications and slides published via the ACL Anthology will adopt CC-BY-4.0, Wikipedia entries will comply with CC-BY-SA. The annotation and diversity support software will mostly adopt the GNU General Public License. The licenses for the other software (parsers, MWE identifiers, etc.) will be chosen by their authors, but the Action will strongly promote open-source release.

Stakeholders	Results	Means	Impact
Research community	Theoretical and practical findings of the Action	Journals, conferences, workshops, shared tasks	Post-Action research activities and publications containing new results
Experts in yet uncovered languages	Theoretical and practical findings of the Action	Collaborative online infrastructures which will remain operational beyond the Action's duration. Remote training.	Inclusion of new languages into the unified framework. Larger validation of the framework.
Professionals in language technology	Language resources in unfied formats. Methods for solving technological tasks in many languages.	Open source releases of resources, tools and models via sustainable infrastructures (CLARIN, Github). Injection of expertise via European Digital Innovation Hubs.	Better coverage of diverse linguistic phenomena and of a larger number of languages in their products
European institutions	Coordinated expertise on technological support for language diversity	Coordinated responses to EC public consultations on future research directions in research and innovation programmes. Action's website.	Better account of inter- and intra-language diversity in public research and innovation funding programmes.
European and worldwide institutions	Measures of intra-linguistic diversity	Injection of expertise via European Digital Innovation Hubs. Action's website. Press releases announcing the Action's topics, objectives and achievements	Enhancing inclusiveness of disadvantaged social groups, e.g. by estimating and reducing complexity of texts via the diversity measures.
UNESCO	Action's outcomes summary. Links to related and spin-off language safeguiarding projets.	Online forms for comments to the UNESCO Atlas of the World'sLanguages in Danger	Improvement of the data presented in the Atlas of Endangered Languages
Professionals in language teaching	Diversity measures defined by the Action and tools to apply them to texts.	Online versions of the diversity measurement tools. Commercial software based on the Action's proposals.	Adapting texts used in language teaching to the competence level of students.
Speakers of low-resourced/endangered languages	Documentation of these languages in the unified framework. Knowledge transfer from well-resourced languages.	Open source releases of resources, tools and models. End-user applications built upon the Action's outcomes	Better protection of endangered languages. More justice in information access (due to native language suupport).
Citizens in and beyond Europe	Language resources and tools for many languages. Tools boosted for covering diverse linguistic phenomena. Action's results summarized in accessible ways.	End-user applications built upon the Action's outcomes. Articles in Wikipedia on the Action's topics. Demos at the European Researchers' Night and the European Day of Languages.	Better technology support for native languages. Finer support of diverse linguistic phenomena. Increasing the awareness of inter- and intra-linguistic diversity.

4. IMPLEMENTATION

4.1. COHERENCE AND EFFECTIVENESS OF THE WORKPLAN

4.1.1. DESCRIPTION OF WORKING GROUPS, TASKS AND ACTIVITIES

To ensure the achievement of its objectives, the Action will be structured into 4 Working Groups (WGs):

WG1: CORPUS ANNOTATION

Annotated corpora constitute the Action's major operational tools for NLP-applied universality. Therefore, WG1 will be 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: (i) defining the division of labour between morpho-syntactic and semantic annotation, (ii) addressing

hard or weakly covered syntactic phenomena (syntactically irregular structures, relative clauses, coordination, pronoun inclusivity, etc.), (iii) covering new MWE categories (nominal, adjectival and functional MWEs), (iv) paving the way for unified annotation guidelines for idiosyncratic constructions;

- Coordinate the development and maintenance of centralized **software** for universality-based corpus construction: (i) online spaces for community discussion and editing annotation guidelines, (ii) tools for automatic pre-annotation, annotation transfer and manual annotation of corpora, (iii) tools for corpus merging, validation, curation, statistics, conversion and release. The software development itself will be funded at national levels;
- Defining **file formats** for corpora annotated according to the unified guidelines;
- Construction of annotated **corpora**: (i) adapting the existing corpora to the enhanced guidelines, (ii) 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 (§1.1.1.2). 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 will be dedicated to:

- Cross-language **unification of lexical features**: (i) harmonizing the definition of a “syntactic word” across languages, (ii) harmonizing lemmatization rules (for words and MWEs) and lexical features across languages, (iii) standardizing lists of lexemes for auxiliaries, pronouns and determiners;
- **Design** of a lexicon-corpus **interface** aiming at: (i) interlinking MWE lexicon entries with their occurrences in corpora, (ii) 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 will be dedicated to NLP coordinating the development of tools leveraging universality and promoting diversity:

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

The tools themselves will be funded at the national level. WG2 will bring the federating effect to these activities, notably by organizing multilingual **evaluation campaigns** on parsing and MWE identification. Diversity-based evaluation measures from WG4 will be promoted. The outcomes should validate the computational tractability of the terminologies unified in WG1.

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

4.1.2. DESCRIPTION OF DELIVERABLES AND TIMEFRAME

To ensure the achievement of the objectives (§1.2.2) and to measure the Action's progress, the following deliverables will be produced (cf. timeframe in §4.1.4):

[D1] A scientific publication describing **measures** of inter- and intra-linguistic diversity in language resources and tools.
[D2] Unified, enhanced and enlarged versions of existing **annotation guidelines** for lexical features, morphology, syntax and MWEs, together with the criteria for applying these unified guidelines to specific languages.
[D3] Centralized documentation of the nationally funded **software** infrastructures coordinated in WG1 and WG4, to support universality and diversity in language resources.
[D4] Centralized documentation of unified **file formats** and conventions for corpora and lexica.
[D5] Centralized documentation of (new or enhanced) annotated **corpora** for at least 100 languages.
[D6] Documentation of prototypes of NLP-applicable **lexica** of MWEs and idiosyncratic constructions.
[D7] Centralized documentation of multilingual and cross-lingual **NLP tools** (§1.1.1.3) coordinated in WP3: syntactic and semantic parsers, MWE discovery tools, MWE identifiers, prototypes of identifiers of idiosyncratic constructions.
[D8] Diversity **benchmarks** for NLP: diversity-driven evaluation scenarios for NLP resources and tools; infrastructure for evaluation campaigns of NLP tools; evaluation results of at least 2 evaluation campaigns and focused on inter/intra-linguistic diversity in 100 languages.
[D9] Website: describing the Action's organisation, and gathering the links to its events and outcomes.
[D10] Other **dissemination material**: reports from STSMs; material from training schools; proceedings of workshops; joint papers in Open Access journals, conferences and books; dissemination material dedicated to a large audience (e.g., demonstrations of tools and Wikipedia entries about diversity in NLP, MWEs and idiosyncratic constructions, and interesting syntactic phenomena).
The table below shows the measurability of the objectives (§1.2.2) in terms of the deliverables.

		Deliverables									
		1	2	3	4	5	6	7	8	9	10
Research coordination objectives (RCO)	1	x		x	x				x		
	2		x	x	x	x		x		x	x
	3	x	x	x	x	x	x			x	x
	4	x			x	x		x	x	x	x
	5								x	x	x
	6		x	x	x	x	x	x	x	x	x
Capacity-building objectives (CBO)	1		x	x	x	x		x	x	x	x
	2			x	x			x	x	x	x
	3		x	x	x	x	x	x	x	x	x
	4		x	x			x		x	x	x

4.1.3. RISK ANALYSIS AND CONTINGENCY PLANS

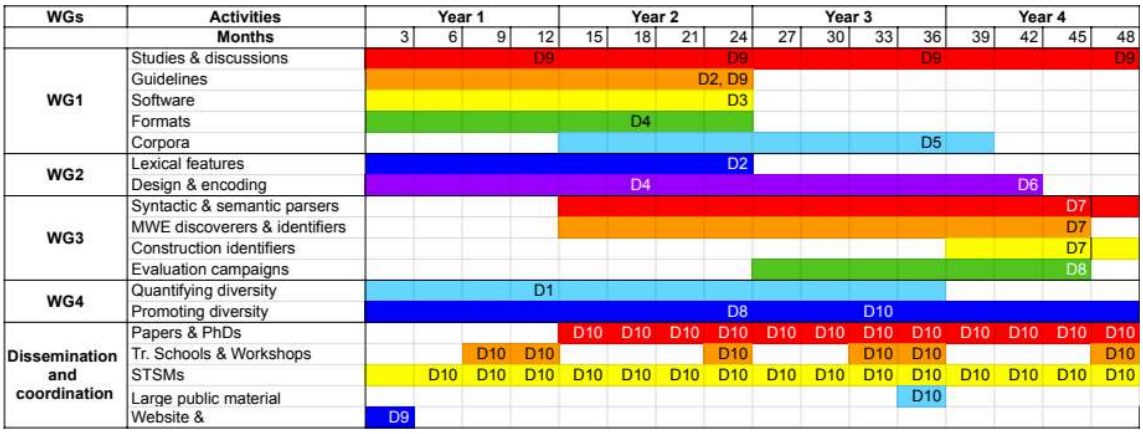
The risks related to the Action's Work Plan stem from three main factors. First, the activities in favour of NLP-applicable universality of terminologies and methodologies, which the Action is to coordinate, are very ambitious, and partly achievable in practice. This is because universality requires validating the Action's hypotheses against *all* existing languages (many of them not written and spoken by few individuals) and language phenomena (most of which occur rarely). Second, the fragmentation issues are particularly severe in language modelling (§1.1.1.3), and even a very open and inclusive discussion forum may still have a moderate success due to varying linguistic traditions. Third, for a large majority of existing languages, there are few or no linguistic/NLP experts who could be integrated into this collaborative effort. Also, many of those who are relevant might not be easily included due to weak financial support at the national level.

Therefore, the only realistic approach is to come as close as possible to universality, that is, to unify and validate terminologies and methodologies across as many languages and phenomena as possible. As a consequence, the first contingency plan is to adapt the number of languages and phenomena covered by the Action 's deliverables to those which can be addressed in practice, given the available expertise, and the degree of agreement between experts. Those phenomena for which the experts do not agree, can remain handled differently for various languages, i.e., with language-specific categories. Secondly, the Action will promote national spin-off initiatives. If some goals prove particularly hard to fulfil by the whole network and require national research funding, they will be delegated for pilot studies on a smaller scale. Thirdly, worldwide open evaluation campaigns (D8.3) will involve a large international community in achieving the Action's goals. If the tools coordinated by the Action (D7) do not achieve the state-of-the-art results, the authors of the best-scored tools will be prompted to openly publish their prototypes and reference them via the Action infrastructure. Fourthly, in case of languages with reduced or no national support, the Action will prioritise mobility funds from/to the relevant countries, and support co-authorship with publication fees covered by the Action. Finally, the open access release of all the existing outcomes will allow the network to pursue its goals, if the achievement of some of them is delayed, also beyond the duration of the Action. An external risk is due to COVID-19. If on-site meetings and cross-border travels are restricted in some periods, the Action will organize online events.

4.1.4. GANTT DIAGRAM

The following diagram illustrates the Action's schedule in terms of WGs, their activities and deliverables. Task dependency was accounted for, e.g., unified guidelines and file formats come before enhanced and extended corpora and lexica The tasks in WG4 are interlinked with the 3 other WG, due to their

orthogonality. Finally, dissemination and coordination follow classical instruments and budgets of a COST Action.



WGs	Activities	Year 1				Year 2				Year 3				Year 4			
	Months	3	6	9	12	15	18	21	24	27	30	33	36	39	42	45	48
WG1	Studies & discussions				D9				D9				D9				D9
	Guidelines								D2, D9								
	Software								D3								
	Formats						D4										
	Corpora												D5				
WG2	Lexical features								D2								
	Design & encoding						D4								D6		
WG3	Syntactic & semantic parsers															D7	
	MWE discoverers & identifiers															D7	
	Construction identifiers															D7	
	Evaluation campaigns															D8	
WG4	Quantifying diversity				D1												
	Promoting diversity								D8			D10					
Dissemination and coordination	Papers & PhDs					D10	D10	D10	D10	D10	D10	D10	D10	D10	D10	D10	D10
	Tr. Schools & Workshops			D10	D10				D10			D10	D10				D10
	STSMs		D10	D10	D10	D10	D10	D10	D10	D10	D10	D10	D10	D10	D10	D10	D10
	Large public material												D10				
	Website &	D9															

REFERENCES

- Agić, Tiedemann, Merkler, Krek, Dobrovoljc, Može (2014): Cross-lingual Dependency Parsing of Related Languages with Rich Morphosyntactic Tagsets. Proc. LT4CloseLang Workshop
- Bender (2011): On achieving and evaluating language-independence in NLP, in LILT, 6(3):1–26
- Berdicevskis, Çöltekin, Ehret, von Prince, Ross, Thompson, Yan, Demberg, Lupyan, Rama, Bentz (2018), Using Universal Dependencies in cross-linguistic complexity research, Proc. UDW Workshop
- Chomsky (1976): Reflections on Language, London: Temple Smith.
- Derczynski, Nichols, van Erp, Limsopatham (2017): Results of the WNUT2017 shared task on novel and emerging entity recognition. Proc. 3rd Workshop on Noisy User-generated Text
- Devlin, Chang, Lee, Toutanova (2019): BERT: Pre-training of deep bidirectional transformers for language understanding. In Proc. NAACL
- Duong (2017): Natural Language Processing for Resource-Poor Languages, PhD thesis, The University of Melbourne
- Evans, Levinson (2009): The myth of language universals: Language diversity and its importance for cognitive science. Behavioral and Brain Sciences 32
- Forschungsverbund Berlin (2018): Genetic diversity helps protect against disease. ScienceDaily
- Givón (1995). Functionalism and grammar. John Benjamins.
- Greenberg (ed.) (1966): Universals of language. MIT Press.
- Hwa, Resnik, Weinberg, Cabezas, Kolak (2005): Bootstrapping Parsers via Syntactic Projection across Parallel Texts. Natural Language Engineering, 11(3)
- Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, Monojit Choudhury (2020) The State and Fate of Linguistic Diversity and Inclusion in the NLP World, Proc. of ACL 2020.
- Kirov, Cotterell, Sylak-Glassman, Walther, Vylomova, Xia, Faruqui, Mielke, McCarthy, Kübler, Yarowsky, Eisner, Hulden (2018): UniMorph 2.0: Universal Morphology, Proc. LREC 2018

- de Lhoneux, Stymne, Nivre (2017): Old School vs. New School: Comparing Transition-Based Parsers with and without Neural Network Enhancement. Proc. TLT Workshop
- McDonald, Petrov, Hall. 2011. Multi-Source Transfer of Delexicalized Dependency Parsers. Proc. EMNLP
- Nivre, de Marneffe, Ginter, Goldberg, Hajič, Manning, Pyysalo, Schuster, Tyers, Zeman (2020): Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection. Proc. LREC 2020
- Nivre, Rimell, McDonald, Gomez-Rodríguez (2010): Evaluation of Dependency Parsers on Unbounded Dependencies, Proc. COLING 2010.
- Phillips (2014): How Diversity Makes Us Smarter. Scientific American, October
- Pires, Schlinger, Garrette (2019): How Multilingual is Multilingual BERT? Proc. ACL-2019
- Ponti, Reichart, Korhonen, Vulić (2018): Isomorphic Transfer of Syntactic Structures in Cross-Lingual Natural Language Processing. Proc. ACL 2018.
- Edoardo Maria Ponti, Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, Anna Korhonen (2019) Modeling Language Variation and Universals: A Survey on Typological Linguistics for Natural Language Processing, in Computational Linguistics 45(3).
- Ramisch, Savary, Guillaume, Waszczuk, Candito, Vaidya, Barbu Mititelu, Bhatia, Iñurrieta, Giouli, Güngör, Jiang, Lichte, Liebeskind, Monti, Ramisch, Stymne, Walsh, Xu (2020): Edition 1.2 of the PARSEME Shared Task on Semi-supervised Identification of Verbal Multiword Expressions, Proc. MWE-LEX.
- Rimell, Clark, Steedman (2009): Unbounded Dependency Recovery for Parser Evaluation, Proc. EMNLP.
- Rohanian, Taslimipoor, Kouchaki, Ha, Mitkov (2019): Bridging the Gap: Attending to Discontinuity in Identification of Multiword Expressions, Proc. NAACL.
- Savary, Candito, Barbu Mititelu, Bejček, Cap, Čéplö, Cordeiro, Eryiğit, Giouli, van Gompel, HaCohen-Kerner, Kovalevskaitė, Krek, Liebeskind, Monti, Parra Escartín, van der Plas, QasemiZadeh, Ramisch, Sangati, Stoyanova, Vincze (2018): PARSEME multilingual corpus of verbal multiword expressions, in Markantonatou et al. (Eds.) Multiword expressions at length and in depth. Language Science Press.
- Savary, Cordeiro, Ramisch (2019) "Without lexicons, multiword expression identification will never fly: A position statement", Proc. of MWE-WN 2019.
- Maggie Tallerman (2009): If language is a jungle, why are we all cultivating the same plot?, Behavioral and Brain Sciences 32(5)
- Williams, Lessard, Desu, Clark, Bagrow, Danforth, Dodds (2015): Zipf's law holds for phrases, not words. Scientific Reports, 5
- Wu, Dredze (2020): Are All Languages Created Equal in Multilingual BERT? In Proceedings of the 5th Workshop on Representation Learning for NLP
- Yarowsky, Ngai, Wicentowski. 2001. Inducing Multilingual Text Analysis Tools via Robust Projection Across Aligned Corpora. Proc. HLT
- Zeman, Resnik. 2008. Cross-Language Parser Adaptation between Related Languages. Proc. IJCNLP