COST Action
Progress Report at 24 months
(23/09/2022 to 23/09/2024)

DRAFT

CA21167: Universality, diversity and idiosyncrasy in language technology

The Action was approved by the Committee of Senior Officials (CSO) on 27-5-2022 and has the MoU reference COST 081/22.

This report shows the data entered into e-COST to enable the Action Chair to verify the completeness and accuracy of the report with the MC prior to submitting the report via e-COST in fulfilment of the rules for COST Action Management, Monitoring and Final Assessment.

COST Association AISBL
Avenue du Boulevard - Bolwerklaan 21, box 2 | 1210 Brussels, Belgium
T +32 (0)2 533 3800 | office@cost.eu | www.cost.eu

Funded by the European Union

Action leadership and participants

Leadership positions

Position	Name	Contact details	Country*
Chair	Prof Agata Savary	agata.savary@universite-paris-saclay.fr +33169158003	France

Position	Name	Contact details	Country*
Action Vice-Chair	Dr Daniel Zeman	zeman@ufal.mff.cuni.cz +420-951-554-225	Czechia

Working groups

#	WG Title	# of participants	WG Leader	Country*
1	WG1: CORPUS ANNOTATION	235	Mr Bruno Guillaume Bruno.Guillaume@loria.fr	France
2	WG2: LEXICON-CORPUS INTERFACE	180	Dr Verginica MITITELU vergi@racai.ro	Romania
3	WG3: MULTILINGUAL AND CROSS-LINGUAL LANGUAGE TECHNOLOGY	241	Prof Joakim Nivre joakim.nivre@lingfil.uu.se	Sweden
4	WG4: QUANTIFYING AND PROMOTING DIVERSITY	149	Dr Marie-Catherine de Marneffe marie-catherine.demarneffe@uclouvain.be	Belgium

DRAFT

Other key leadership positions

Position	Name	Contact details	Country*
Science Communication Coordinator	Ms Olesea Caftanatov	olesea.caftanatov@math.md	Moldova
GH Scientific Representative	Alina Wróblewska	alina@ipipan.waw.pl	Poland

* The country displayed is:

- for the Action Chair, the country that nominated that person to the Management Committee before they were elected Action Chair;
- for the Action Vice-Chair the country that nominated the person as a Management Committee Member,
- for all other leadership positions, if the person is a MC Member the country displayed is the country of nomination, otherwise it is the country of the person's primary work affiliation.

Participants

COST members having accepted the MoU

AL	22/06/2022	AM	09/01/2023	AT	22/06/2022	BE	22/06/2022	BA	22/06/2022
BG	22/06/2022	HR	22/06/2022	CY	22/06/2022	CZ	22/06/2022	DK	22/06/2022
EE	22/06/2022	FI	22/06/2022	FR	22/06/2022	GE	22/06/2022	DE	22/06/2022
EL	22/06/2022	HU	22/06/2022	IS	22/06/2022	IE	22/06/2022	IL	22/06/2022
IT	22/06/2022	LV	22/06/2022	LT	22/06/2022	LU	22/06/2022	MT	22/06/2022
MD	22/06/2022	ME	22/06/2022	NL	22/06/2022	MK	22/06/2022	NO	22/06/2022
PL	22/06/2022	PT	22/06/2022	RO	22/06/2022	RS	22/06/2022	SK	22/06/2022
SI	22/06/2022	ZA	22/06/2022	ES	22/06/2022	SE	22/06/2022	CH	22/06/2022
TR	22/06/2022	UA	22/06/2022	UK	22/06/2022				

Other participants

Institution Name	Country

DRAFT

Summary

The main aim and objective of the Action is to

reconcile language diversity with rapid progress in language technology

During its first two years the Action progressed the achievement of this as described below

The main aim of the Action is to reconcile language diversity with rapid progress in language technology. This aim 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.

The Action's progress to date in addressing this main aim is substantial. Two MoU objectives have a high level of progress. Others have a level of progress of 26-50%. The network is large and active, with many ongoing tasks defined in a grassroot manner. Many IRIs play important leadership roles. We already took a number of dissemination and exploitation steps. A number of members reported on the impact of UniDive on their careers, and we trust that this impact is actually much larger.

Concerning the deliverables, we coordinated the publication of several new universality-driven language corpora (UD and PARSEME treebanks) in 32 new languages, most of them low-resourced. We have 25 joined peer-reviewed publications and several software pieces, including those for novel ways of diversity quantification. We have ongoing tasks dedicated to several universality issues addressing: annotation guidelines, data formats, the wordhood notion, corpus annotation infrastructures, etc.

We have started setting up an evaluation framework which will showcase the measurement of the main objective. Namely, we achieved understanding of previous efforts in diversity quantification from scientific domains (such as ecology) where diversity has been largely theorized. We have first implementations of these measures and we study their applicability in NLP. We are working towards NLP evaluation campaigns in which: (i) the universality of data formats is a guiding principle, (ii) the data have a high inter- and -intra-linguistic diversity, (iii) systems solving an NLP task are evaluated both for performance and for diversity.

The Action will implement the following measures in the coming two years to overcome any issues identified in this report as potentially endangering the achievement of the objectives of the Action

The mailing lists have been newly implemented at the University of Chisinau, Moldova, by the institution of the Science Communication Coordinator. The usage of this lists is still unstable, likely due to aggressive antispam policies from some universities against content stemming from Moldova. Some configuration issues of these new lists still need to be solved.

Action website

https://unidive.lisn.upsaclay.fr/

Achievement of MoU objectives, deliverables and additional outputs/ achievements

MoU objectives

Please self-assess and describe the level of achievement of each MoU objective. For any MoU objective that is 25% or less achieved, please add an explanation.

Mou objective	To develop methods for quantifying inter- and intra-linguistic diversity.
Type of objective	1.a Development of a common understanding/definition of the subject matter 1.d Comparison and/or performance assessment of a theory, model, methodology, technology or technique
Level of progress	26 - 50%
Description of progress with achieving the MoU objective	At least three scientific articles (Estève et al. 2024, Ploeger et al. 2024a, Ploeger et al. 2024b) describe particular inter- and intra-linguistic diversity measures, applied in NLP evaluation. WG4 is also developing new, more generic measures. On the one hand, it is evaluating how different existing diversity measures capture lexical diversity; on the other hand, it is developing a measure to be used in the shared-task organized by WG3 to assess the diversity of morphosyntactic analysers.

Mou objective	To develop a common understanding of language universals across 70 languages represented in the Action.
Type of objective	1.a Development of a common understanding/definition of the subject matter
Level of progress	26 - 50%
Description of progress with achieving the MoU objective	WG3 is organizing a shared task on morphosyntactic parsing, using representations that abstract over differences in morphosyntactic structure, in particular along the analytic-synthetic scale. This will help us better understand what is really universal across typologically different languages. WG1 extended the universalist terminology and annotation guidelines for MWEs to non-verbal categories of MWEs; this paves the way towards modeling all types of MWEs in 26 languages along the same lines, and in the mid-term perspective, to study language universals concerning idiomaticity, in these 26 languages. WG1 prepared a survey about annotation practices among UniDive community members; the goal is to determine ways in which linguistic theory (particularly typology) can help improve UD annotation guidelines. WG1 designed a construction-oriented documentation of the UD guidelines which will enable a clearer overview of the differences and similarities of syntactic strategies across languages. (Task 1.3) WG1 prepared a survey on annotating spoken data within the UD and SUD annotation schemes to enhance the cross-lingual comparability of this under-resourced and under-explored type of data (Task 1.5) WG2 has two tasks in which language universals wrt lemmatization and tokenization (of words and multiword expressions) are investigated, mainly by surveys sent to treebanks developers and maintainers and by interpreting their responses in a multilingual context.

DRAFT

	WG2 also carries on a survey on the notion of wordhood, so as to unify the understanding of the notion of a "syntactic word" in very many languages.

Mou objective	To coordinate the diversity-driven creation, merging and enhancement of language resources unified across over 100 languages from the UD and PARSEME collections.
Type of objective	1.a Development of a common understanding/definition of the subject matter 1.b Coordination of information seeking, identification, collection and/or data curation 1.e Development of knowledge needing international coordination, pertaining to a new or improved theory, model, methodology, technology or technique
Level of progress	26 - 50%
Description of progress with achieving the MoU objective	WG3 coordinates the creation of unified annotated corpora for the shared task on morphosyntactic parsing. WG3 contributes indirectly to this objective by exploring, evaluating and visualizing language tools and resources as documented in existing repositories such as CLARIN and ELG, thereby clarifying what needs exist for different languages and resource types. WG2 coordinates a task in which a parallel corpus (initially of 10 languages) is enriched with several new ones and annotated at a series of linguistic levels, including lexical, morph-syntactic and semantic (word senses and multiword expressions). WG1 coordinates and provides training for the creation of new manually annotated corpora, especially within the UD, SUD and PARSEME annotation schemes. Since the beginning of the Action, 32 new languages joined the UD collection of treebanks, and 22 of them were created by UniDive members. Additionally, for 17 languages which already had some treebanks in UD, new treebanks were added. The PARSEME corpus release 1.3 extended the universalist approach in PARSEME to 26 languages (Arabic and Serbian are new to this version). The corpus is now fully compatible with UD, which increases universality. WG1 conducted a survey on UD annotation practices and challenges to inform the development of a manual providing support to new annotators. WG1 conducted a survey on tools for manual corpus annotation to identify and promote tools supporting diversity-driven corpus annotation.

Mou objective	To coordinate efforts towards a better coverage of inter-/intra-linguistic diversity in NLP tools.
Type of objective	1.c Coordination of experimentation or testing 1.d Comparison and/or performance assessment of a theory, model, methodology, technology or technique 1.g Input to stakeholders (e.g. standardization body, policy-makers, regulators, users), excluding commercial applications
Level of progress	0 - 25%
Description of progress with achieving the MoU objective	WG4 is contributing to this objective by surveying what is currently done in the NLP field with regard to diversity. What is considered diverse? How is diversity accounted for and measured? Which NLP applications and at which pipeline stages are concerned by diversity quantification?

DRAFT

Mou objective	To raise awareness of the international community about the importance of diversity preservation in language technology.
Type of objective	1.g Input to stakeholders (e.g. standardization body, policy-makers, regulators, users), excluding commercial applications 1.j Dissemination of research results to stakeholders (excluding specific input in view of knowledge application)
Level of progress	0 - 25%
Description of progress with achieving the MoU objective	New diversity-driven evaluation measures will be used in campaigns of NLP tasks, open to a wide international community. One such evaluation campaign is in the pipeline in WG3 and a call for more campaigns was recently published.

Mou objective	To disseminate the Action outcomes to stakeholders.
Type of objective	1.i Dissemination of research results to the general public 1.j Dissemination of research results to stakeholders (excluding specific input in view of knowledge application)
Level of progress	26 - 50%
Description of progress with achieving the MoU objective	See the details in the "Dissemination" section of this progress report.

Mou objective	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.
Type of objective	2.a Building a community around a topic of scientific and/or socio-economic relevance, allowing for knowledge exchange and the development of a joint research agenda 2.c Bridging separate fields of science/disciplines to achieve breakthroughs that require an interdisciplinary approach 2.e Building capacity in the demographic inclusiveness of networks in science and technology, including representation of newly established research groups, Early-Career Investigators, the under-represented gender and teams from countries/regions with less capacity in the field of the Action
Level of progress	76 - 100%
Description of progress with achieving the MoU objective	The Action currently gathers about 380 members from some 40 countries, also outside Europe. The network is large and active, due to several instruments. Firstly, the Action is popular and new experts submit WG applications every month. An inclusiveness-driven policy was defined by the Core Group to handle these applications. Most candidates, even those whose expertise is not central to the Action, are admitted, provided that they can contribute, for instance by building language resources for a new under-resourced language or disseminating the outcomes to stakeholders. The Working Groups are large and active since the beginning of the action. In the 1st Grant Period each WG published calls for tasks related to the WG topics and for task leaders. We currently have: 5 tasks in WG1, 3 tasks (with several subtasks) in WG2, 3 tasks in WG3, and 3 tasks in WG4. Tasks are grassroot initiatives which ensure high implication of the network members and extend its impact. These (sub)tasks have several co-leaders each. Each task and subtask has its own online meetings, where concrete progress is achieved. Each WG also has regular online meetings for information, coordination and reporting (10 meetings were organized by WG1, 12 by WG2, 10 by WG3 and 6 by WG4). The minutes of all the

DRAFT

	tasks are available on the Action website (https://unidive.lisn.upsaclay.fr/doku.php?id=working_groups). Concerning onsite events, we have 1 large yearly General Meeting for the whole action, and 1 yearly meeting for a selected WG. During theseevents we make progress on the ongoing tasks, but we also integrate new members, not necessarily active in tasks, as well as external experts, by inviting them to submit and present (peer reviewed) posters on their scientific activity. We had 1 training school after which new young researchers joined the Action and/or achieved a better understanding of the Action's objectives and methods. Additionally, WG3 contributes to this objective through the organization of the shared task on morphosyntactic parsing, which involves the creation of data within a common framework by experts in a (large) number of languages. The PARSEME network existed before UniDive, with 24 language teams. In UniDive. WG1 brought 5 new language leaders to the network (Arabic, Serbian, Dutch, Ancient Greek, Ancient Egyptian). New annotators joined the already covered languages. Several meetings reinforced the convergences between the PARSEME and UD modelling practices. In the CLiB 2024 conference two members of WG2 organized a tutorial on MWEs for the Bulgarian researchers community, as well as for other interested participants. As a conclusion, the network is large and active. Many members, also beyond the Core Group, play active roles in the tasks. The meetings are frequent, and show overall systematic progress.

Mou objective	To foster the capacities of Young Researchers and Innovators (YRIs), with special focus on COST ITC participants.
Type of objective	2.e Building capacity in the demographic inclusiveness of networks in science and technology, including representation of newly established research groups, Early-Career Investigators, the under-represented gender and teams from countries/regions with less capacity in the field of the Action
Level of progress	26 - 50%
Description of progress with achieving the MoU objective	Three YRI play leadership roles in the Extended Core Group (https://www.cost.eu/actions/CA21167/#tabs+Name:Main%20Contacts%20and%20Leadership): Olesea Caftanatov (Moldova), Kaja Dobrovoljc (Slovenia) and Atul Kumar Ojha (Ireland). Twelve YRIs are (sub)task leaders: • In WG1: André Coneglian (Brazil), Flavio Massimiliano Cecchini (Belgium), Carlos Ramisch (France), Atul Kr. Ojha (Ireland), František Forgáč (Slovakia), Kaja Dobrovoljc (Slovenia) • In WG2: Kilian Evang (Germany), Jaka Čibej (Slovenia) • In WG3: Omer Goldman (Israel), Leonie Weissweiler (Germany) • In WG4: Abigail Walsh (Ireland), Louis Estève (France) The 1st UniDive Training School hosted over 35 YRIs, and helped them develop new skills and become part of the community. YRIs were the beneficiaries of STSM and ITC conference grants: 12 out of 17 in Grrat Period 1, 14 out of all 27 in Grant Period 2. 2 out of 5 on-site events were organized in ITCs (Turkey, Moldova) and hosted not only reimbursed participants, but also local, mainly YRI, participants.

Mou objective	To coordinate and boosting universality-driven initiatives worldwide.

DRAFT

Type of objective	2.a Building a community around a topic of scientific and/or socio-economic relevance, allowing for knowledge exchange and the development of a joint research agenda 2.e Building capacity in the demographic inclusiveness of networks in science and technology, including representation of newly established research groups, Early-Career Investigators, the under-represented gender and teams from countries/regions with less capacity in the field of the Action
Level of progress	26 - 50%
Description of progress with achieving the MoU objective	The UD workshop and the MWE workshop are long-standing venues for two universality-driven communities, Universal Dependencies and PARSEME, underlying UniDive. Even if coordinated mainly the UniDive members, these venues have a worldwide scope. The boosting effect was achieved in 2024, when UniDive took the initiative to merge these venues into a joint event: the MWE-UD workshop in Torino. Its co-location with a major NLP conference, LREC-COLING 2024, opened the event to a larger NLP public. As previously mentioned, UniDive boosted the universality-driven development of language resources by attracting 22 new languages to UD and 5 to PARSEME treebank collections. In WG3, the upcoming shared task on morphosyntactic parsing also brings a converging effect between UD and UniMorph (a universality-driven initiative dedicated to modeling morphology).

Mou objective	To set up a long-term roadmap for the joint efforts of the universality-driven NLP community.
Type of objective	2.a Building a community around a topic of scientific and/or socio-economic relevance, allowing for knowledge exchange and the development of a joint research agenda 2.d Acting as a stakeholder platform or trans-national practice community, pertaining to a certain area of socio-economical or societal application, or to a certain market sector 2.e Building capacity in the demographic inclusiveness of networks in science and technology, including representation of newly established research groups, Early-Career Investigators, the under-represented gender and teams from countries/regions with less capacity in the field of the Action
Level of progress	26 - 50%
Description of progress with achieving the MoU objective	A white paper (Savary et al. 2023: PARSEME Meets Universal Dependencies: Getting on the Same Page in Representing Multiword Expressions) put forward a roadmap for the UD-PARSEME unification, with short-, mid- and long-term objectives.

DRAFT

Deliverables

This section covers only deliverables that were foreseen for the Action, not additional outputs that were generated during the Action (these additional outputs will be added in the following section). Please select and comment on the progress with achieving each deliverable.

For deliverables that are:

- Delivered, please provide proof to enable the Action Rapporteur to confirm the delivery
- Not delivered but delivery is foreseen within 2 years please explain how the delivery will be achieved
- Not foreseen to be delivered please explain why not

Deliverable	A scientific publication describing measures of inter- and intra-linguistic diversity in language resources and tools.		
Progress with achieving deliverable	Not delivered, but expected before end of Action	**Month deliverable due**	12
Explanation	Pending delivery: WG4 is surveying the NLP field and assessing how diversity is measured and represented there. Estimated result: a survey paper. Two provisional abstracts have been submitted to the Computational Linguistics journal. The abstract was rejected, we head for another venue. Delivered: Estève et al. (2024) Vector Spaces for Quantifying Disparity of Multiword Expressions in Annotated Text Ploeger et al. (2024) A Principled Framework for Evaluating on Typologically Diverse Languages Ploeger et al. (2024) What is "Typological Diversity" in NLP?		

Deliverable	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.		
Progress with achieving deliverable	Not delivered, but expected before end of Action	**Month deliverable due**	24
Explanation	Ongoing delivery: Task 1.2 in WG1 has drafted the guidelines for non-verbal (nominal, adjectival, and functional) MWEs. Pending delivery: WG1 is reorganizing the UD guidelines, adopting a new architecture based on syntactic constructions. This revised guidelines version will facilitate more consistent and efficient application across multiple languages Extra delivery: WG1 is improving the guidelines for cross-lingually harmonized morphosyntactic annotation of speech-specific phenomena (in progress). Delivered: Omer Goldman, Leonie Weissweiler, Reut Tsarfty (2024) New annotation guidelines for morphosyntax https://github.com/omagolda/msap-docs (suggestions were provided by other WG3 members; guildelines are compatible with the UD framework).		

Deliverable	Centralized documentation of the nationally funded software infrastructures coordinated in WG1 and WG4, to support universality and diversity in language resources.		
Progress with achieving deliverable	Not delivered, but expected before end of Action	**Month deliverable due**	24
Explanation	Pending delivery: - A beginner-friendly introduction to UD guidelines is created. - A survey on manual annotation tools was launched to gather an inventory of these tools and their functionalities. Delivered: PARSEME Wiki (https://gitlab.com/parseme/corpora/-/wikis/home) – centralised documentation of PARSEME software and resources. The new and updated pages concern: updating morpho-syntactic anotations to synchronise them with UD annotations, integrating the UD validation script with the PARSEME validator, continuous development and continuous integration for language corpora, the FLAT centralized annotation platform – it was recently migrated to a new server, controlled by UniDive, and upgraded to the most recent version.		

DRAFT

Deliverable	Centralized documentation of unified file formats and conventions for corpora and lexica.		
Progress with achieving deliverable	Not delivered, but expected before end of Action	**Month deliverable due**	18
Explanation	Pending delivery: A community discussion https://github.com/UniDive/WG1/discussions/2 on the pro and cons of the currently used formats is being conducted to propose an optimized and universal format. Delivered: Manual annotation tools comparison https://docs.google.com/spreadsheets/d/1FZo6sSdIkxXCm9p9FcV8PzVKCeXot6apnnA-zXYhRwk – a survey of formats used in existing tools.		

Deliverable	Centralized documentation of (new or enhanced) annotated corpora for at least 100 languages.		
Progress with achieving deliverable	Not delivered, but expected before end of Action	**Month deliverable due**	36
Explanation	Pending delivery: Maintaining and expanding the two large multilingual structured data infrastructures: a. Universal Dependencies https://universaldependencies.org – 22 more languages and 38 new treebanks were released in November 2023 and May 2024. Each new language is documented in the UD infrastructure. Currently 161 languages are covered. b. PARSEME https://gitlab.com/parseme/corpora/-/wikis/home#languages – GitLab repositories for 4 new languages were added. Currently 28 languages are covered.		

DRAFT

Deliverable	Documentation of the prototypes of NLP-applicable lexica of MWEs and idiosyncratic constructions.		
Progress with achieving deliverable	Not delivered, but expected before end of Action	**Month deliverable due**	42
Explanation	Pending delivery: - Standardizing Lexica of MWEs https://unidive.lisn.upsaclay.fr/lib/exe/fetch.php?media=meetings:general_meetings:2nd_unidive_general_meeting:task_2_3_christian_slides.pdf – introduction by Ch. Chiarcos. -A canonical form for flexible multiword expressions https://drive.google.com/file/d/1-vm13_MhKmdC01hsqn6CBBGubBu7XXlI/view – the Dutch draft by J. Odijk		

Deliverable	Centralized documentation of multilingual and cross-lingual NLP tools coordinated in WP3: syntactic and semantic parsers, MWE discovery tools, MWE identifiers, prototypes of identifiers of idiosyncratic constructions.		
Progress with achieving deliverable	Not delivered, but expected before end of Action	**Month deliverable due**	45
Explanation	Pending delivery: WG3 has so far explored, evaluated and visualized existing documentation repositories, such as CLARIN and ELG. In the future, we will also document multilingual and cross-lingual tools that are developed for the evaluation campaigns organized by WG3.		

Deliverable	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.

Progress with achieving deliverable	Not delivered, but expected before end of Action	**Month deliverable due**	45
Explanation	Pending delivery: 1st evaluation campaign: WG3 is in the process of organizing the first evaluation campaign on morphosyntactic parsing, and has recently started the process for organizing a second campaign (with a different theme). 2nd evaluation campaign: In collaboration with WG3, WG4 is organizing a call to the NLP community to gather tasks that focus on different aspects of linguistic diversity, in order to create a diversity benchmark.		

Deliverable	Website: describing the Action's organisation, and gathering the links to its events and outcomes.		
Progress with achieving deliverable	Delivered	**Month deliverable due**	3
Proof of progress with achieving the deliverable	https://unidive.lisn.upsaclay.fr		

Deliverable	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).		
Progress with achieving deliverable	Not delivered, but expected before end of Action	**Month deliverable due**	48
Explanation	Pending delivery: Dissemination materials have been made available on the UniDive website (see the 9th deliverable): - A list of STSM topics and grant recipients is published under the "Grants" section, with reports accessible only within the e-COST system to mitigate the risk of misappropriating research topics. - The "Outcomes" section provides materials from the 1st UniDive training school, the UniDive webinar, proceedings from two workshops held during General Meetings, a list of publications related to UniDive, and other dissemination outcomes.		

DRAFT

Additional outputs / achievements

Co-authored Action publications

Please enter below ONLY publications (including publications that are submitted but not yet accepted):

- that are on the topic of the Action, and
- that are co-authored by at least two Action participants from two countries participating in the Action, and
- for which the Action networking was necessary.

Please pay special attention to the COST Excellence and Inclusiveness policy and ensure the inclusion of publications with authors from COST Inclusiveness Target Countries (ITCs), from the underrepresented gender in the Action and from Early Career Investigators/Young researchers.

DRAFT

	Bibliographic data	Countries participating in the Action among authors	Open Access	COST cited?	COST funds?	Relevance to H2020 Societal challenge	Peer Reviewed?
1	doi:10.3384/nejlt.2000-1533.2023.4453	FR, RO, SE	Y	N	N	Europe in a changing world, inclusive innovative and reflective societies	Y
2	doi:10.18653/v1/2023.mwe-1.6		Other	Y	Y		Y
3	doi:10.18653/v1/2024.americasnlp-1.5		Other	Y	Y		Y
4	Verginica Barbu Mititelu, Voula Giouli, Kilian Evang, Daniel Zeman, Petya Osenova, Carole Tiberius, Simon Krek, Stella Markantonatou, Ivelina Stoyanova, Ranka Stanković, Christian Chiarcos (1024) Multiword Expressions between the Corpus and the Lexicon: Universality, Idiosyncrasy, and the Lexicon-Corpus Interface, in Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD) @ LREC-COLING 2024, pages 207–219, Torino, Italia. ELRA and ICCL.	BG, CZ, DE, EL, NL, RO, RS, SI	Y	Y	N	Europe in a changing world, inclusive innovative and reflective societies	Y

5	Archna Bhatia, Gosse Bouma, A. Seza Doğruöz, Kilian Evang, Marcos Garcia, Voula Giouli, Lifeng Han, Joakim Nivre, and Alexandre Rademaker (2024) Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD) @ LREC-COLING 2024, ELRA and ICCL, Torino, Italia	BE, DE, EL, NL, SE, UK	Y	Y	Y	Europe in a changing world, inclusive innovative and reflective societies	Y
6	Roberto Díaz Hernández, Marco Passarotti (2024) Developing the Egyptian-UJaen Treebank, to appead in the Proceedings of the 22nd Workshop on Treebanks and Linguistic Theories (TLT 2024), Hamburg, Germany, December.	IT, ES	Y	Y	Y	Europe in a changing world, inclusive innovative and reflective societies	Y
7	doi:10.5281/zenodo.10949960		Other	Y	Y		Y
8	Svetlozara Leseva, Verginica Barbu Mititelu, Ivelina Stoyanova, Mihaela Cristescu (2024) A uniform multilingual approach to the description of multiword expressions] (2024), in Giouli, Voula & Barbu Mititelu, Verginica (eds.). 2024. [[https://langsci-press.org/catalog/book/440|Multiword expressions in lexical resources: Linguistic, lexicographic, and computational perspectives. (Phraseology and Multiword Expressions 6). Berlin: Language Science Press. DOI: 10.5281/zenodo.10949960	BG, RO	Y	N	N	Europe in a changing world, inclusive innovative and reflective societies	Y
9	Esther Ploeger, Wessel Poelman, Andreas Holck Høeg-Petersen, Anders Schlichtkrull, Miryam de Lhoneux, Johannes Bjerva (2024) A Principled Framework for Evaluating on Typologically Diverse Languages, arXiv:2407.05022	BE, DK	Y	N	N	Europe in a changing world, inclusive innovative and reflective societies	Y
10	Esther Ploeger, Wessel Poelman, Miryam de Lhoneux, Johannes Bjerva (2024)	BE, DK	Y	N	N	Europe in a changing world, inclusive innovative and reflective	Y

DRAFT

	What is "Typological Diversity" in NLP?, in EMNLP 2024.					societies	
11	Agata Savary, Daniel Zeman, Verginica Barbu Mititelu, Anabela Barreiro, Olesea Caftanatov, Marie-Catherine de Marneffe, Kaja Dobrovoljc, Gülşen Eryiğit, Voula Giouli, Bruno Guillaume, Stella Markantonatou, Nurit Melnik, Joakim Nivre, Atul Kr. Ojha, Carlos Ramisch, Abigail Walsh, Beata Wójtowicz, Alina Wróblewska (2024) UniDive: A COST Action on Universality, Diversity and Idiosyncrasy in Language Technology, in the Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024, pages 372–382, Torino, Italia. ELRA and ICCL.	BE, CZ, FR, EL, IE, IL, MD, PL, PT, RO, SI, SE, TR	Y	Y	Y	Europe in a changing world, inclusive innovative and reflective societies	Y
12	Jonathan Washington, Çağrı Çöltekin, Furkan Akkurt, Bermet Chontaeva, Soudabeh Eslami, Gulnura Jumalieva, Aida Kasieva, Aslı Kuzgun, Büşra Marşan, Chihiro Taguchi (2024) Strategies for the Annotation of Pronominalised Locatives in Turkic Universal Dependency Treebanks, in Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD) @ LREC-COLING 2024, pages 207–219, Torino, Italia. ELRA and ICCL.	DE, TR	Y	Y	Y	Europe in a changing world, inclusive innovative and reflective societies	Y
13	Ana Ostroški Anić, Kristina Strkalj Despot, Luka Terčon (2023) Detecting patterns of implicit offensive language in multilingual data, 1st UniDive Workshop, Université-Paris-Saclay, 16 March 2023.	HR, SI	Y	N	Y	Europe in a changing world, inclusive innovative and reflective societies	Y
14	Simon Krek, Carole Tiberius, Kaja Dobrovoljc, Polona Gantar, Jelena Kallas, Kristina Koppel, Svetla Peneva Koeva, Veronika Lipp, László Simon (2023) The ELEXIS parallel sense-annotated corpus, 1st UniDive Workshop, March 16-17 March 2023, Université Prais-Saclay, France.	BG, EE, HU, NL, SI	Y	N	Y	Europe in a changing world, inclusive innovative and reflective societies	Y

DRAFT

15	Christian Chiarcos, Anas Fahad Khan, Maxim Ionov, Elena Simona Apostol, Besim Kabashi, Ciprian-Octavian Truica, Katerina Gkirtzou (2023) OntoLex-FrAC: Standardizing the Corpus-Lexicon Interface, 1st UniDive Workshop, March 16-17 March 2023, Université Prais-Saclay, France.	DK, DE, IT, RO	Y	Y	Y	Europe in a changing world, inclusive innovative and reflective societies	Y
16	Gülşen Eryiğit, Ali Şentaş, Johanna Monti (2023) Dodiom: a Gamified Bot supporting Diversity and Multilinguality for Idiom Corpora Construction, 1st UniDive Workshop, Université-Paris-Saclay, 16 March 2023, France	IT, TR	Y	N	Y	Europe in a changing world, inclusive innovative and reflective societies	Y
17	Yuval Pinter, Miryam de Lhoneux (2023) Transitions all the Way Down: From Characters to Full Document Annotation in One System, 1st UniDive Workshop, Université-Paris-Saclay, 16 March 2023, France	BE, IL	Y	N	Y	Europe in a changing world, inclusive innovative and reflective societies	Y
18	Carole Tiberius, Jaka Čibej, Jelena Kallas, Kadri Muischnek, Simon Krek (2024) UD syntax for the ELEXIS parallel sense-annotated corpus: a pilot study, 2nd UniDive Workshop, University of Naples L'Orientale, 8 February 2024, Italy	EE, NL, SI	Y	N	Y	Europe in a changing world, inclusive innovative and reflective societies	Y
19	Ranka Stanković, Christian Chiarcos, Milica Ikonić Nešić (2024) Leveraging Linked Data, NIF, and CONLL-U for Enhanced Annotation in Sentence Aligned Parallel Corpora, 2nd UniDive Workshop, University of Naples L'Orientale, 8 February 2024, Italy	DE, RS	Y	Y	Y	Europe in a changing world, inclusive innovative and reflective societies	Y
20		CZ, DE	Y	N	Y	Europe in a changing	Y

DRAFT

	Kilian Evang, Daniel Zeman (2024) Word Segmentation in Universal Dependencies, 2nd UniDive Workshop, University of Naples L'Orientale, 8 February 2024, Italy					world, inclusive innovative and reflective societies	
21	Cagri Coltekin, Furkan Akkurt, Olcay Taner Yildiz, Sardana Ivanova, Tunga Gungor, Suzan Uskudarli, Mehmet Oguz Derin, Aida Kasieva, Gulnura Dzhumalieva, Bermet Chontaeva, Jonathan Washington, Soudabeh Eslami, Chihiro Taguchi, Aslı Kuzgun, Büşra Marşan (2024) Unifying the Annotations in Turkic Universal Dependencies Treebanks, 2nd UniDive Workshop, University of Naples L'Orientale, 8 February 2024, Italy	DE, TR	Y	Y	Y	Europe in a changing world, inclusive innovative and reflective societies	Y
22	Christian Chiarcos, Maxim Ionov, Andrius Utka, Sigita Rackeviciene (2024) Enhancing Interoperability for Under-Resourced Language Datasets: A Case Study on Linking Lithuanian-English Cybersecurity Data, 2nd UniDive Workshop, University of Naples L'Orientale, 8 February 2024, Italy	DE, LT	Y	Y	Y	Europe in a changing world, inclusive innovative and reflective societies	Y
23	Thomas Pickard, Aline Villavicencio, B. Madalina Zgreabăn (2024) MultiNCI - A Multilingual Noun Compound Idiomaticity Dataset, 2nd UniDive Workshop, University of Naples L'Orientale, 8 February 2024, Italy	NL, UK	Y	N	Y	Europe in a changing world, inclusive innovative and reflective societies	Y
24	Louis Estève, Kaja Dobrovoljc (2024) A new pipeline for measuring diversity across various linguistic levels. 3rd UniDive Workshop. Budapest, Hungary. (under review)	FR, SI	Y	Y	Y	Europe in a changing world, inclusive innovative and reflective societies	Y

DRAFT

Projects resulting from Action activities

Please enter below all the projects on the topic of the Action resulting from Action activities, involving at least one Action participant, and for which the Action networking was necessary.

The Action reported 3 project(s) and 0 proposal(s) resulting from the Action networking.

Key details of the projects are shown below.

#	Title	Countries participating in the Action among proposers	Main proposer name	Funder	Amount	Call identifier	Relevance to H2020 Societal Challenge
1	TESLA	RS	Ranka Stanković	National	212452 €	PRIZMA-AI	Europe in a changing world, inclusive innovative and reflective societies
2	Model adaptation for language varieties	CH	Tanja Samardzic	National	13300 €	Scientific Exchanges 2023	Europe in a changing world, inclusive innovative and reflective societies
3	Multilingual Coreference Resolution	TR	Gülşen Eryiğit	National	15000 €	Tubitak 2515 European Cooperation in Science And Technology Program	Europe in a changing world, inclusive innovative and reflective societies

DRAFT

Other outputs / achievements

Please enter below any additional outputs/ achievements on the topic of the Action that contribute to the COST mission: “COST enables break-through scientific developments leading to new concepts and products and thereby contributes to strengthen Europe’s research and innovation capacities", and for which the Action networking was necessary (e.g. a patent, standards, white paper).

Output / achievement description	Dependence of achievement on the Action networking

Software: diversutils - library for measuring diversity of linguistic resources, as well as system predictions (France	High
DUST - a library for measuring syntactic diversity in treebanks (France, Slovenia)	High
LangDive – library for measuring the level of linguistic diversity in multilingual NLP datasets (Serbia, Switzerland)	High
Updates to the Universal Dependencies tools	Low
Updates to Grew-match - corpus browser form (S)UD treebanks, PARSEME corpora and many others	Medium
Updates to Arborator Grew - a collaborative annotation tool for treebank developpement	Medium
Updates to PARSEME utilities - for language leaders and corpus release experts	Medium
Migration and upgrade of the FLAT annotation platform - used for manual annotation of multiword expressions in PARSEME	High

DRAFT

Impacts

Please describe the impacts (the short- to long-term scientific, technological, and / or socioeconomic changes produced by a COST Action, directly or indirectly, intended or unintended) that have resulted, or might result, from the Action in the following table (one impact per line).

Description of the impact, i.e. what will change, and for whom, as a result of what the Action achieved	Type of impact	Timing of impact
Specific impact submitted by Roberto **Díaz Hernández**: UniDive has given Díaz Hernández the opportunity to meet scholars specialised in computational linguistics and to introduce concepts and methods from this discipline into Egyptian philology. On the one hand, Díaz Hernández has started to develop the UD-EUJA treebank, the first morphosyntactic treebank for pre-Coptic Egyptian. It currently contains 1,573 sentences and 14,650 words and will become a useful linguistic tool for understanding the synchronic and diachronic use of pre-Coptic Egyptian. On the other hand, Díaz Hernández has opened a new line of research in Egyptian philology dedicated to the identification and analysis of multiword expressions in Egyptian. His aim is now to create a database of Egyptian multiword expressions in PARSEME, which will help him to carry out a study on the evolution of Egyptian multiword expressions. As a result, two new digital resources for Egyptian will be available at the end of this COST action—the UD-EUJA treebank and a database of Egyptian multiword expressions in PARSEME. These two tools will contribute to the development of deep learning methods and applications for Egyptian, such as the automatic morphosyntactic analysis of Egyptian texts for philologists and the automatic translation of Egyptian inscriptions for archaeologists and visitors to Egypt. Díaz Hernández's contribution to UniDive will help him to become a professor after completing the "Beatriz Galindo" junior professorship programme at the University of Jaén (2023–2027).	• Scientific / Technological	Achieved
Specific impact soubmitted by **Amal Haddad** (**YRE**, Spain) : I am Amal Haddad Haddad from the University of Granada(https://www.ugr.es/personal/amal-haddad-haddad), and I am under 40 (36years old). I am also a member in the LexiCon research Group(http://lexicon.ugr.es/).I have just come back in September from an STSM in Bulgaria, where Ijoined the Bulgarian Academy of Science. At short and long terms, I believe this visit was so useful for me tolearn new annotation methods and to learn about the work of the team ofthe Bulgarian Academy of Science, headed by Prof Svetla Koeva. Hopefully, there will be future collaboration between the BulgarianAcademy of Science and the LexiCon research group.	• Scientific / Technological	Achieved
Specific impact submitted by **V. Mititelu** and **I. Stoyanova**: Verginica Barbu Mititelu (Romania) and Ivelina Stoyanova (Bulgaria) presented jointly a tutorial on the annotation of multiword expressions as part of the International Conference Computational Linguistics in Bulgaria 2024 (https://dcl.bas.bg/clib/tutorials/) drawing on their joint work within UniDive (as well as previous COST Action PARSEME). Joint activities facilitate the long-term collaboration between the individual researchers as well as between their institutions, both from inclusiveness target countries. Moreover, the participants at the tutorial were from Bulgaria, Albania, Romania, Croatia, Latvia, Spain,	• Scientific / Technological	Achieved

DRAFT

most of them ITCs, and working on low-resourced languages, sharing their experience and efforts in areas closely related to the UniDive activities.		
Specific impact submitted by **Abigail Walsh** (**YRE**, Ireland): A Short-Term Scientific Mission (STSM) funded visit enabled collaboration between researchers at universities in Ireland and England and resulted in the development of specialised Irish language resources targeting noun compounds. These resources have been integrated into the development of testing datasets for assessing MT engine and LLM quality, as well as furthering linguistic understanding of these constructions for Irish. This research has been extended into a PhD project exploring other rare and diverse phenomena that remain a challenge with current NLP models. Furthermore, the development of these resources has inspired further collaborative work between researchers at universities in Ireland and France, to perform evaluation experiments on LLMs using these testing datasets. These ongoing projects represent a significant part of the research work being undertaken by a new postdoctoral researcher at Dublin City University, and have allowed them to connect with many successful senior researchers in this field. These connections are likely to lead to even more collaborations in the future, to perform more cross-lingual analyses of these idiosyncratic constructions.	• Scientific / Technological	Achieved
Specific impact submitted by **Stella Markantonatou** (a Research Director at Athena Research Center, Greece): Three computational linguists from Greece, two Greek postdoctoral students and a mature scientist, have taken advantage of the tools offered by UniDive as follows: **Stavros Bompolas** (**postdoc**). The topic of Stavros' research is in the domain dialectometry. He is trying to apply AI tools, such as models and LLMs, to dialectometry. He works with Greek dialects. UD resource development is crucial for his postoc research. Stavros attended the UniDive 2024 Chisinau Training School. He considers it "instrumental in shaping his research trajectory. The training provided essential skills and insights, particularly in mastering the Universal Dependencies framework and the development of treebanks. This knowledge has been vital in his efforts to build dialectal treebanks for Greek varieties. It has also enabled him to examine how the concept of diversity is applied in NLP, leading to more accurate and comprehensive analyses." He is currently planning to join WP4. **Vivian Stamou** (**postdoc**). Vivian has been active in two domains (a) processing of under-resourced language varieties (b) text classification. She is trying to bridge the gap between under-resourced language varieties and reasonably resourced ones by exploiting the power of LLMs. Vivian is an active member of Athena Research Center in the domains of text classification and language modelling. The STSM has added to both her knowledge and her profile as a specialist in language modelling. Vivian was awarded an STSM (Sept 2024). Vivian visited the team of Valerio Basile and Cristina Bosco at the University of Turin. The team has an expertise in offensive speech detection and in UDs. Vivian explored the aspect of offensive VMWEs in offensive speech detection (most work on offensive speech detection has not considered the VMWE aspect). **Stella Markantonatou** is a **Research Director** at Athena Research	• Scientific / Technological • Societal	Achieved

DRAFT

Center. Through a UniDive STSM she visited the team of Aline Villavicencio at the University of Sheffield and worked on nominal MWEs. Stella develops MWE resources for Modern Greek for NLP purposes. She has used the STSM experience to involve students of MSc courses in NLP in the production and evaluation of such resources with LLMs. This is a contribution to the students' careers and to the issue of providing resources to Modern Greek that has no substantive lexica and only medium general purpose corpora.		
Impact on the **University of Moldova**: The University of Chișinău hosted the 1st UniDive Training School in July 2024. It was one of the first international events organized by this institution. Its impact on the University is many fold: • increasing its international visibility, • showing its scientific and administrative preparedness for international events, • supporting European integration of Moldova in science, • promoting NLP locally, • integrating local students into the group of the trainees	• Scientific / Technological • Societal	Achieved
Specific impact submitted by **Dawit Jembere** (**YRI**, Sweden): As a student of language technology with a purely linguistics background, my involvement in UniDive has opened up opportunities for me to work on dependency parsing in low-resource languages, specifically Amharic and Gedeo. It has also expanded my professional network and shifted my career focus from pure linguistics to a more hybrid field that integrates both linguistics and computer science. Additionally, I have drafted a master's project titled: Parsing for Amharic and Gedeo Using Universal Dependencies (pls see the attached short description). This project is a direct result of my involvement in UniDive and reflects how the Action has influenced both my research direction and professional growth.	• Scientific / Technological	Achieved
Impact submitted by **Agata Savary** (the Action chair): Coordinating COST Actions, both UniDive and, previously, IC1207 PARSEME (2013-2017) has crucial importance for my career. Between the two Actions, I obtained a full professor position at a major French university, in probably the biggest NLP research group in France, and my COST experience was a major factor in this success. Currently UniDive is one of important international visibility factors in my institution (the LISN lab at Paris-Saclay University). In November 2024, a 6-year evaluation of our university (by the national HCERES certification body) is taking place. UniDive was selected as one of 2 activities to be highlighted during oral presentations of my research group in this evaluation. I'm also in the process of setting up an ERC (advanced grant) proposal, to be submitted in 2025. It directly builds upon UniDive assets. In case of success, the impact on my career and on stakeholders of the project will be invaluable.	• Scientific / Technological • Societal	Achieved
Specific impact submitted by **Liudmila Mockienė** (Director of the Institute of Humanities, Faculty of Human and Social Sciences, Mykolas Romeris University, Lithuania):	• Scientific / Technological	Foreseen within two years of the end of the Action

DRAFT

Participation in this COST action has contributed greatly to involvement of younger (early-career) and more experienced researchers into the scientific activities of the department, which will also enhance internationalization of the research at our university, which is in line with our institutional strategic goals. The researchers have also expanded their network of cooperation and will continue doing so by taking part in working groups and tasks which are carried on by researchers in the same field thus providing a forum for effective professional communication and potential involvement of our researchers in other scientific projects, events, etc. They also have (and will in the future) developed their research skills by attending trainings and workshops. This new knowledge and experience will allow them to apply new research methods and share them with colleagues to create added value by producing more efficient research outputs which is vital for accreditation of the study programmes we deliver and comparative research evaluation reports (which we have to provide every five years).		
Impact for the researchers, technology developers and speakers of low-resourced languages: Since the beginning of UniDive, the Universal Dependencies collection of treebanks was extended by 32 new languages (22 created by UniDive members), all of them low resourced: Azerbaijani, Bavarian, Bororo, Cappadocian, Egyptian, Georgian, Gujarati, Haitian Creole, Hausa, Highland Puebla Nahuatl, Latgalian, Macedonian, Malayalam, Middle French, Old Irish, Ottoman Turkish, Paumari, Sinhala, Western Sierra Puebla Nahuatl, Xavante, Zaar, Abaza, Abkhaz, Classical Armenian, Gheg, Kyrgyz, Luxembourgish, Maghrebi Arabic-French, Nheengatu, Telugu-English, Tswana, and Veps. Also the PARSEME collection of treebanks annotated for multiword expressions was extended by Arabic, Serbian, Dutch, Ancient Greek, Ancient Egyptian, all low-resourced. For most of these 30+ languages, the corresponding treebank is a are (if not the only) the only structured language resource, which paves the way towards the documentation of the language (linguistic objective) and its integration into NLP tools (technological objective). A possible expected impact is both a better preservation of the language and its coverage by language technology, to the benefit of the speakers.	• Scientific / Technological • Economic • Societal	Foreseen within two years of the end of the Action
Impact on the researchers and speakers of Turkic languages: UniDive organised the UD Turkic Workshop 2023 (https://ud-turkic.github.io/udtw23/), co-located with the WG3 meeting in Istanbul on 8 September 2023. The objective was to bring together people working on Universal Dependencies (UD) treebanks for Turkic languages. The focus was consistent annotations across different treebanks and linguistic phenomena in Turkic languages that are not easy to annotate with the current UD guidelines. The impact comes from a better coordination and consistency of morphosyntactic modelling of Turkic languages, to the benefit of	• Scientific / Technological • Economic • Societal	Foreseen within two years of the end of the Action

DRAFT

researchers, technology developers and speakers of these languages.		
Impact on Young Researchers and Investigators: YRIs are rather well integrated in the Action, many of them take important roles as Extended Core Group members and task leaders. Many of them benefited from STSMs and of the Training School, which increased their visibility in the Action and allowed them to open new collaborations, including those leading to joint publications.	• Scientific / Technological	Foreseen by the end of the Action
Specific impact submitted by **Kilian Evang** (**YRI**, University of Düsseldorf): Serving as a Task leader in this Action is being one of the most significant networking opportunities of my career so far, not least due to the large number of actively engaged participants (21 people at the last Task meeting). If successful, the Task will lead to at least one highly collaborative publication with high visibility due to the large number of authors, and to a set of guidelines that will be used by treebank creators for years to come.	• Scientific / Technological	Foreseen by the end of the Action

Please describe how the Action is advancing the careers, skills and network of researchers, including ECIs (for example: joint supervision of graduate and PhD students, research exchanges not funded by the Action, collaborations, Training Schools with ECTS accreditation, joint projects and jobs prospects).

All researchers, and especially YRCs, who are active in the action gain good international visibility, which will open their careers to international collaboration and potential recruitment (in which international aspects are highly valued). At least one new joint PhD supervision (UK/France) is planned for 2025. Several STSMs created collaboration which did not exist before and opened perspectives for joint publications. 3 spin-off projects helped career advancement of senior researchers. Two other spin-offs are planned for 2025, including one bilateral. At least two new research lines were opened by senior researchers in connection with the Action.

DRAFT

The career benefits are mainly to researchers with the following amount of experience after their PhD: ≤ 8 years.

Which of the stakeholders described in the "Plan for involving the most relevant stakeholders" in the Action's MoU have been engaged and how? What additional stakeholders have been, or will be, engaged and how?
Experts in theoretical linguistics, typology, NLP and Young Researchers and Innovators are directly integrated in the Action. Researchers working on low-resourced and endangered languages are directly integrated in the Action. They authored 22 out of 32 new UD treebanks and 2 new PARSEME corpora. They benefit from centralized language repositories and discussion platforms (on Github for UD, on Gitlab for PARSEME), a centralized annotation platform (FLAT), shared data validation and release infrastructure (UD tools, PARSEME utilities), centralized documentation (UD guidelines, PARSEME guidelines, FLAT user guide). Representatives of the speakers of low-resourced and endangered languages are directly integrated since the UD and PARSEME treebanks are mostly annotated by native speakers. Experts in empowering low-resourced languages are integrated via strong links with the European Language Grid project (https://live.european-language-grid.eu/) and the European Language Equality (https://european-language-equality.eu/) - see task 1.3 in WG3. At least one professional in language technology and one in language teaching are involved via Linguse (https://linguse.com/) - a reading app for language learners, whose beta version features PARSEME-based identification of idioms. Other professional in language technology will become involved via the shared tasks organized by WG3 in 2025 and 2026. At least one educator is active in several UniDive tasks.

Dissemination and exploitation of Action results (other than co-authored Action publications listed previously)

Please describe the Action's dissemination and exploitation approach as well as all activities undertaken to ensure dissemination and exploitation of the Action results and the effectiveness of these activities.

Dissemination and exploitation approach of the Action

The general dissemination approach in UniDive is to publish all major outcomes of the Action (publications, language resources, software) under open licenses, so as to facilitate their access, in particular by experts of under-resourced languages. Convergences are also sought with related initiatives such as special Interest Groups at the Association of Computational linguistics (SIGLEX-MWE, SIGUL, SIGSlav). Events are co-organized with major international and national NLP venues: LREC, COLING, CLIB, etc. The inclusive UniDive membership policy allows us to integrate members from NLP-related domains, who help disseminate the Action's outcomes.

Dissemination

Dissemination meetings funded by the Action (possible only until 31st October 2021)

Title of Dissemination meeting	Meeting date	Meeting country	Action participant	Event name and hyperlink to the website	Title of presentation	Description of added value to the Action
N/A						

DRAFT

Other dissemination activities

E.g. participation to non-Action meetings, e.g. EU Parliament, meetings with policy makers, experts in the field, regional authorities.

Item/activity	Target audience	Outcome	Hyperlink
SIGUL 2024 (3rd Annual Meeting of the Special Interest Group on Under-resourced Languages)	Academic and industry researchers on under-resourced languages worldwide, members of of the SIGUL special interest group.	- UniDive reference paper presented as poster - UniDive represented at the panel discussion "In a post-ChatGPT world, what are the Challenges and Opportunities for Under-resourced Languages?"	https://sigul-2024.ilc.cnr.it/workshop/programme/
COST training “Science for Policy” and “Science Diplomacy” on 27-28 June 2023	The audience of the training were COST Action members but the ultimate audience of the resulting dissemination activities will be European policy makers.	The training was taken by a UniDive WG leader, and reported to the Core Group.	https://www.cost.eu/cost-events/science-policy-diplomacy-workshops-2023/

Amorós, L. (2024). Emilio, Sofía and language technologies. EDITUM DOI: 10.6018/editum.3072 (in Spanish)	Young scholarly public in Spain, to be educated for language diversity	Book dedicated to young people, published in 2024, introducing the notions of diversity measurement and low resourced languages, mentioning UniDive and PARSEME (pp. 125-146).	https://publicaciones.um.es/publicaciones/public/obras/ficha.seam?numero=3072&edicion=1&cid=4588
Panorama global de las tendencias educativas en 2023. Libro de resúmenes del III Congreso Internacional de Innovación y Tendencias Educativas. Sevilla: Egregius Ediciones.	Participants of the III CONGRESO INTERNACIONAL Innovación y Tendencias Educativas and readers of the proceedings	Three presentations about UniDive by L. Amoros: - COST ACTION-UNIDIVE. Language technology as a response to inclusion - Some natural language processing tools in teaching-learning contexts - Quantifying Linguistic Diversity since NLP	https://innted.org/ponencia/cost-action-unidive-language-technology-as-a-response-to-inclusion/
4.0 World Forum Industrias Creativas Desafios 2023, 19 May 2023, Buenos Aires, Argentina (not sure why 13 words were not enough)	Attendants of the World Forum and readers of the proceedings, including top politicians from dozens of countries (https://worldforum40.org/2023/oradores-ediciones-anteriores/)	A tak by Lucia Amoros Poveda: "Universalidad, diversidad e idiosincrasia en las Tecnologías del Lenguaje para una educación inclusiva" + recording (https://worldforum40.org/2023/world-forum-4-0-dia-1/) minutes 1:51:20 to 2:01:54	https://worldforum40.org/2023/agenda/
Dissemination paper in the 10th International Conference on Modern Greek Dialects and Linguistic Theory, by Konstantinos Sampanis	Attendants of the conference, who are mostly researchers from outside UniDive but interested in the application of theoretical and applied tools to Modern Greek, thus potentially concerned by UniDive outcomes.	Abstract and poster by Konstantinos Sampanis, University of Vienna; Prokopis Prokopidis, Institute for Language and Speech Processing - ATHENA RC; Furkan Akkurt, Boğaziçi University The “Asia Minor Greek in Contact” (AMGiC) Universal Dependencies Treebank	https://e-services.cost.eu/activity/grants/9c054b3e-71fd-4588-9130-4fd1af306c57/download/128974
Till Überrück-Fries: "Multiword Expressions: The Spice of Language Learning" - a presentation at the Polyglot Gathering 2023	The Polyglot Gathering (https://www.polyglotgathering.com) is the world’s biggest international event for polyglots and language lovers, organized every year at the end of May. It gathers language learners and teachers, and other enthusiasts of langue knowledge and practice.	A presentation was given about the Linguse (https://linguse.com/) - a reading app for language learners, whose beta version features PARSEME-based identification of idioms	https://www.youtube.com/watch?v=q016Sn3aljk
User experiments with the Linguse language learning tool, to assess contribution of idiom identification in language learners	A class of students studying the French language at the B1 level at the Institute of Romance Studies at University of Warsaw	User evaluation of the Linguse (https://linguse.com/) tool beta version, and especially of its idiom identification and learning module	https://nlp4call2024.sciencesconf.org/data/pages/Programme_NLP4CALL_2029.pdf

DRAFT

Dagstuhl Seminar 23191 on Universals of Linguistic Idiosyncrasy in Multilingual Computational Linguistics , co-organized by UniDive members on May 07 – May 12, 2023.	37 experts in linguistics, NLP and typology, from 5 continents and dozens of coutries.	Dagstuhl Seminar are know for their inspirational effect on research. The direct outcomes include slides, recordings, minutes (https://gitlab.com/unlid-dagstuhl-seminar/unlid-2023/-/wikis/home) and a volume of abstracts pubslished in the Dagstuhl Reports journal (https://drops.dagstuhl.de/entities/document/10.4230/DagRep.13.5.22). Since this Dagstuhl Seminar, groups of typology and UniDive experts collaborate and published innovative research publications related to universality and idiosyncrasy.	https://www.dagstuhl.de/en/seminars/seminar-calendar/seminar-details/23191
Publication of all UD and PARSEME corpora on Grew-match corpus browser	Researchers concerned about morphology, syntax and multiword expressions in a large number of languages, both well- and low-resourced	All UD and PARSEME corpora readily available for browsing, with a unified query language, and a number of aggregation operators	https://universal.grew.fr/

DRAFT

Exploitation activities

Please describe below any activities undertaken to ensure exploitation (use, in particular in a commercial context) of the Action's achievements.

Item/activity	Target audience	Outcome
Systematic releases, under open licenses and via the European CLARIN/LINDAY infrastructure, of the Universal Dependencies treebanks and of the PARSEME corpora, for a growing number of languages	NLP researchers worldwide, linguists working under-resourced languages, NLP tool constructors adapting their tools to under-resourced languages	The UD treebanks are one of the most used datasets in NLP. For instance at the LREC-COLING 2024 conference several dozens of articles (many of which authored by people not from UniDive) contained "Universal Dependencies in their title.
Proof of concept for implementing future UniDive share tasks on CodaBench.	Language technology experts both in academia and in language technology industry.	CodaBench (https://www.codabench.org/) is a platform for easy development of machine learning competitions. It is an upgrade of CodaLab (https://codalab.lisn.upsaclay.fr/). CodaLab has 50,000 registered users, and has been used to run over 1000 competitions (over 400 in the last year), and receives about 600 submissions per day. According to an independent evaluation by ML Contests in 2023 (https://mlcontests.com/state-of-

		competitive-machine-learning-2023/), CodaLab scored as the first (out of 18) machine learning competition platforms, in terms of hosting the largest number of competitions which fit sufficient inclusion criteria. As its successor since 2023, CodaBench already has 2,500 registered users and has been used to run 60 public competitions, some of which dedicated to NLP. Both CodaLab and CodaBench are flagship platforms developed and maintained by the LISN lab at the University of Paris Saclay in Frence. A master student internship by Achille Desreumaux in 2024, supervised by Agata Savary (the UniDive chair) and Anne-Catherine Letournel (the head of the research engineer team in charge of the platform), addressed porting a past PARSEME shared task to CodaBench. In this way, we achieved a proof of concept for future UniDive shared tasks possibly to run on CodaBench.

DRAFT

Other matters

This section is confidential to the Management Committee, the Action Rapporteur and the COST Association, and is not included in the version of the report that is published on the COST website.

Difficulties in implementing the Action

If any difficulties are experienced in the implementation of the Action (e.g. imbalances of participation across the Working Groups, inactive country representatives) please describe these below. Please also describe the efforts made by the MC to address these.

During the first two years, the Action's communication has been ensured mainly by mailing lists created for the MC, the Core Group, each WG, and the whole action. These lists were hosted by the Paris-Saclay University (the MC Chair's institution). In August 2024, Paris-Saclay University was a victim of a major cyberattack, which encrypted all its main servers and backup servers, including the mailing list server.

DRAFT

Endangerment Measures

Taking into account the issues identified throughout this report, please summarise the measures the Action will implement in the coming two years to overcome any issues identified as potentially endangering the achievement of the objectives of the Action.

The mailing lists have been newly implemented at the University of Chisinau, Moldova, by the institution of the Science Communication Coordinator. The usage of this lists is still unstable, likely due to aggressive antispam policies from some universities against content stemming from Moldova. Some configuration issues of these new lists still need to be solved.

Suggestions for improvements to COST framework/ procedures

The mandate of the Scientific Committee includes providing advice to the COST Committee of Senior Officials on possible improvements to the COST framework. Please describe below any improvements that you believe should be made to the COST framework.

We were instructed that whenever two or more people are to be hosted at the same time in the same place for a short visit, a small WG meeting is preferred over multiple STSMs. We found that this preference is not sufficiently documented in the COST guidelines. We would also like to suggest that the 2nd Progress Report and the 3rd Work and Budget Plan are not requested at the same period. This creates a huge work overload for the Core Group and especially for the action chair.

Sustaining the network beyond the Action

Are there any plans to sustain the network beyond the end of the Action?	YES

Please describe how the network will be sustained beyond the end of the Action.

The goals of UniDive are long-standing. The two initiatives underlying the action, Universal Dependencies, and

PARSEME, have been operating for many years even without funding (although with lesser momentum). They were supported by some local funding and especially by tenure position researchers in various countries believing in common objectives. We expect this effect to continue, in a more converging spirit, after the UniDive funding is over.

We also believe that new spin-off projects will provide further support to the aims set up by UniDive. Some young researchers who are now active in the action, will likely continue working towards the same objectives in their future careers.

Emerging topics/ developments in the field of the Action

Please describe any emerging topics or potentially important future developments identified during the Action and that could potentially be addressed by future COST activities such as Actions S&T Conferences or Exploratory Workshops.

We have nothing of this sort to report so far.

DRAFT

Annex 1: Types of objectives

1 - Coordination of scientific and technological activities at a European level

1.a - Development of a common understanding/definition of the subject matter

1.b - Coordination of information seeking, identification, collection and/or data curation

1.c - Coordination of experimentation or testing

1.d - Comparison and/or performance assessment of a theory, model, methodology, technology or technique

1.e - Development of knowledge needing international coordination, pertaining to a new or improved theory, model, methodology, technology or technique

1.f - Achievement of a specific tangible output that cannot be achieved without international coordination (e.g. due to practical issues such as database availability, language barriers, availability of infrastructure or know-how, etc.)

1.g - Input to stakeholders (e.g. standardization body, policy-makers, regulators, users), excluding commercial applications

1.h - Input for future market applications (including cooperation with private enterprises)

1.i - Dissemination of research results to the general public

1.j - Dissemination of research results to stakeholders (excluding specific input in view of knowledge application)

2 - Community building

2.a - Building a community around a topic of scientific and/or socio-economic relevance, allowing for knowledge exchange and the development of a joint research agenda

2.b - Building a community around a new or emerging field of research

2.c - Bridging separate fields of science/disciplines to achieve breakthroughs that require an interdisciplinary approach

2.d - Acting as a stakeholder platform or trans-national practice community, pertaining to a certain area of socio-economical or societal application, or to a certain market sector

2.e - Building capacity in the demographic inclusiveness of networks in science and technology, including representation of newly established research groups, Early-Career Investigators, the under-represented gender and teams from countries/regions with less capacity in the field of the Action

DRAFT

Annex 2: Dimensions of successes

1 - Breakthroughs

- 1.a - Scientific breakthrough
- 1.b - Technological breakthrough
- 1.c - Breakthrough in socio-economic or societal applications

2 - Policy contribution

- 2.a - Contribution to regulatory policy
- 2.b - Contribution to environmental, infrastructural or agricultural policy
- 2.c - Contribution to economic or socio-economic policy
- 2.d - Contribution to social, cultural or legal policy

3 - Capacity building

- 3.a - Building capacity in an existing field of science and technology
- 3.b - Building capacity in bridging separate fields of science and technology
- 3.c - Building capacity in a new or emerging field of science and technology
- 3.d - Building capacity in valorising and implementing advances and applications in science and technology
- 3.e - Building capacity in the demographic inclusiveness of networks in science and technology, including representation of newly established research groups, Early-Career Investigators, the under-represented gender and teams from countries/regions with less capacity in the field of the Action

DRAFT