UniDive

Universality, diversity and idiosyncrasy in language technology

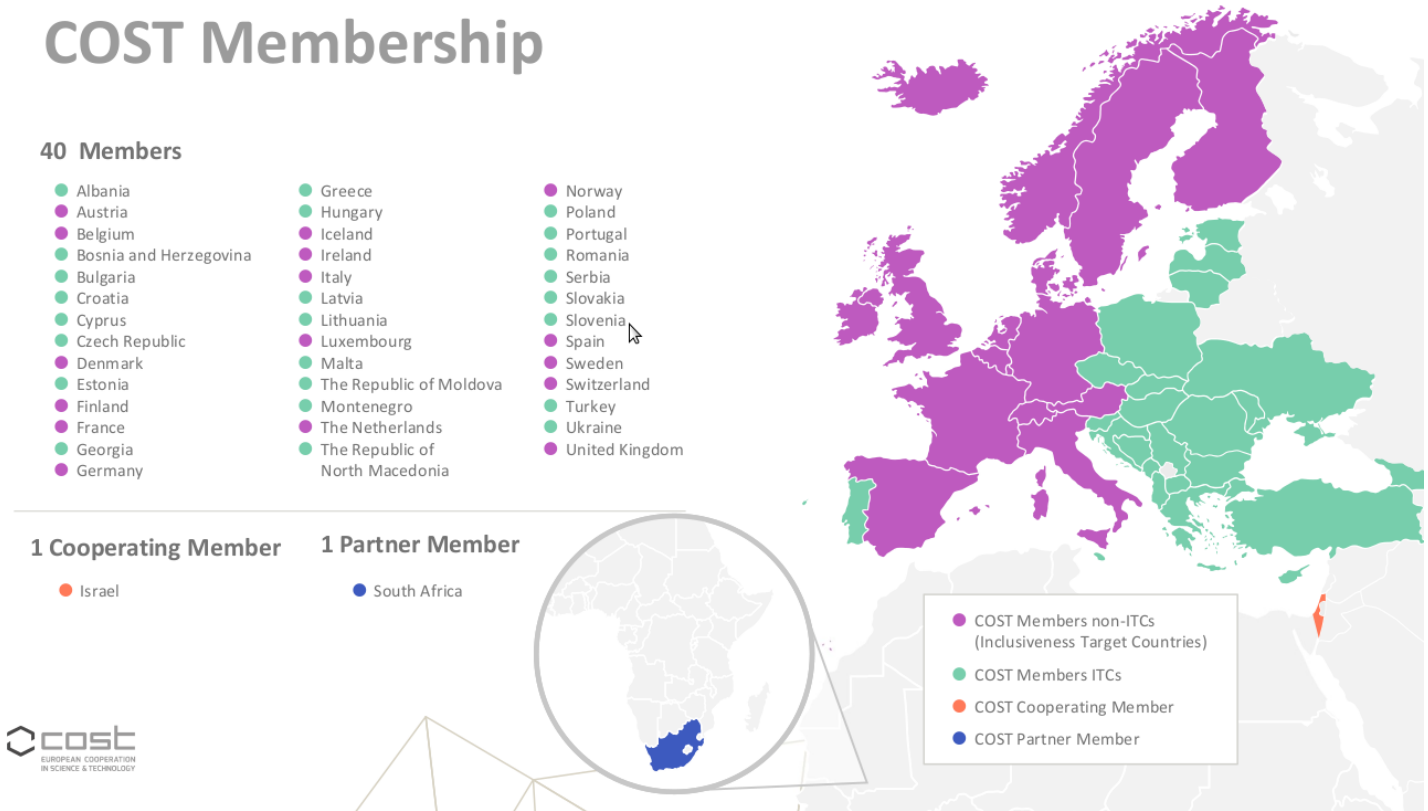
COST Action CA 21167

1st general meeting, Paris-Saclay, opening session

16 March 2023

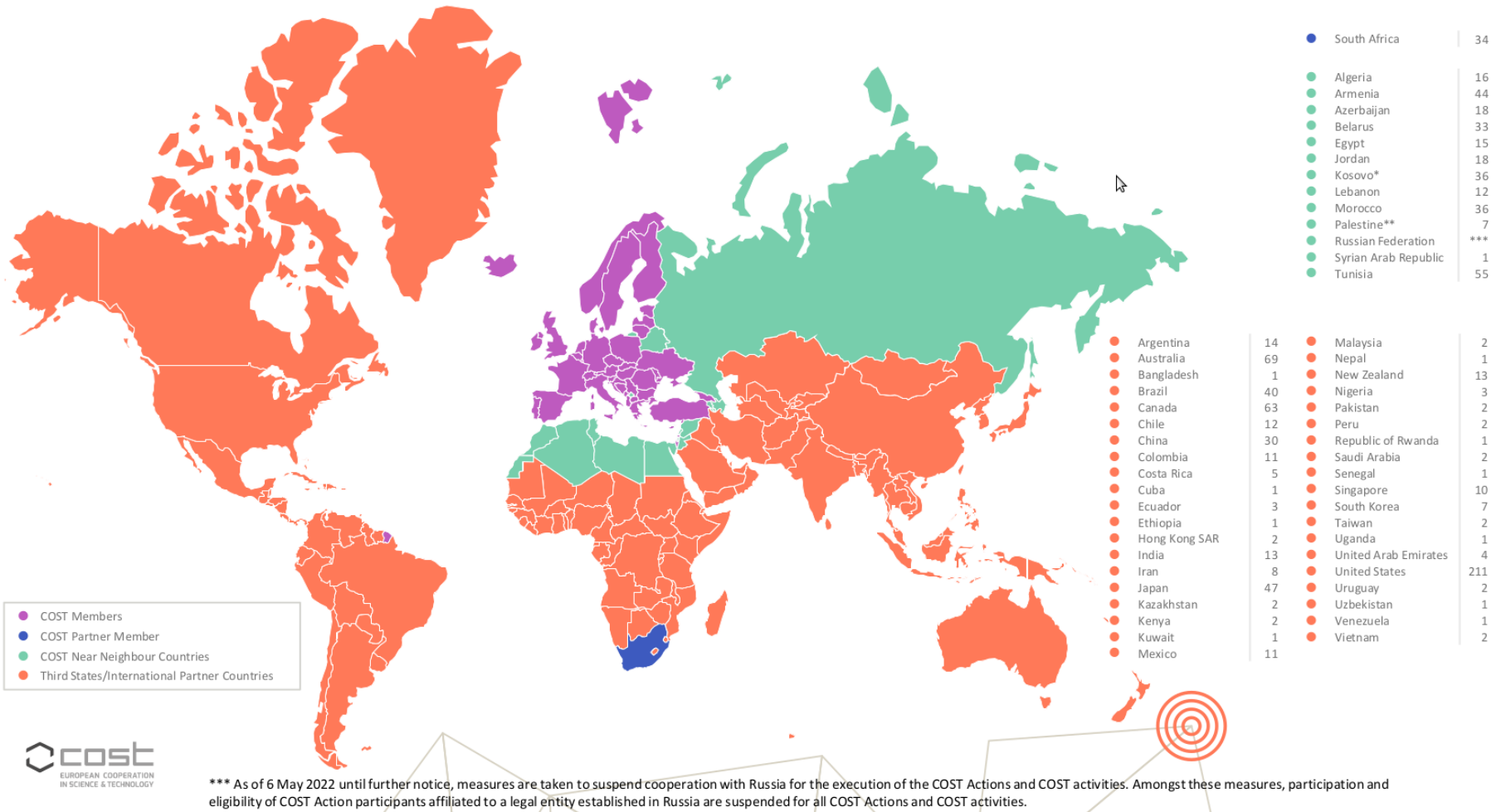
cost EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY
Funded by the European Union
LISN LABORATOIRE INTERDISCIPLINAIRE DES SCIENCES DU NUMÉRIQUE
université PARIS-SACLAY GRADUATE SCHOOL Computer Science

COST



- inter-governmental framework (founded in 1971),
- coordination of nationally-funded European research,
- funded by Horizon Europe.

COST global networking

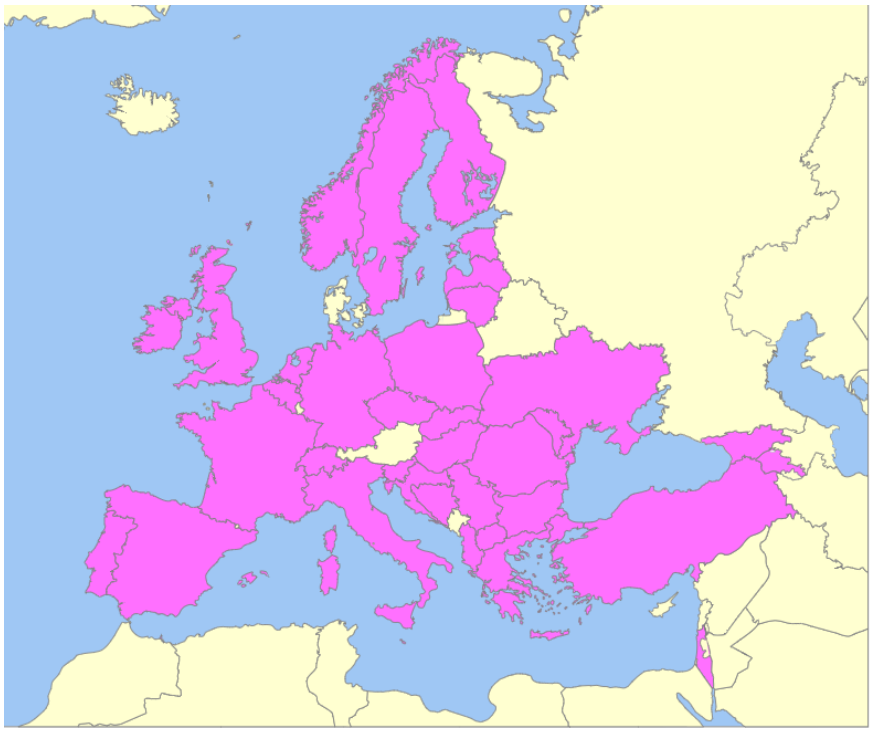


*** As of 6 May 2022 until further notice, measures are taken to suspend cooperation with Russia for the execution of the COST Actions and COST activities. Amongst these measures, participation and eligibility of COST Action participants affiliated to a legal entity established in Russia are suspended for all COST Actions and COST activities.

What Is a COST Action?

- **network** of **individual** researchers and innovators interested in the same objective
- **bottom-up approach**: the scientific challenges are defined by the researchers,
- **instruments**: meetings, workshops, short-term missions, training schools, conferences
- **no** direct **research funding**,
- promoting **inclusiveness** and **balance** (gender, age, geography)
- budget: **120,000–180,000 euros** per year for all partners,
- proposal **success rate** (in 2022): 17.4%.
- precursor role for **other European programmes**,

CA21167 COST Action: **UniDive**



* scientific network
* 36 COST countries
* 23 inclusiveness-target countries
* 4 Working Groups
* kick-off: 23 Sep 2022 Brussels

Duration

4 years: 23 Sept 2022 – 22 Sept 2026

Landscape

Language and technology

- Language diversity – vital heritage to be preserved
- Language technology (NLP) – flourishing data science branch

Inter-language diversity in NLP

- Overwhelming dominance of **English** (Bender 2011)
- Booming **multilingual** NLP
- Data **scarceness** (low-resourced and endangered languages)

Intra-language diversity in NLP

- Most linguistic phenomena have a **Zipfian** distribution
- Data **sparseness** - most phenomena are individually infrequent
- NLP favors **few frequent** phenomena and under-performs in the **many rare** ones
- **Rare** phenomena are **interesting** (unbounded dependencies, idiosnycrasies in multiword expressions)

Landscape

Fragmentation

- Linguistic phenomena appear "on the surface" in **language-specific** ways
- An expert can only reason in terms of **a few languages** she is familiar with
- This leads to **divergent** theories, terminologies and methods

Universality

- Major linguistic debate:
  - Long-standing tradition of **language universals** (Greenberg, 1966; Chomsky, 1976; Tallerman, 2009)
  - vs. no property is universal, **diversity** is the very nature of language (Evans & Levinson, 2009)
- **Universality** in NLP
  - **Cross-linguistically** consistent and **applicable** language descriptions
  - Focus on "**statistical universals**" while leaving room for true **peculiarities**
  - Universal Dependencies (UD, Nivre et al. 2020), PARSEME (Savary et al. 2018, Ramisch et al., 2020) and UniMorph (Kirov et al. 2018)

Universality vs. diversity

Universality (convergence of different points of view on similar phenomena) might be perceived as opposed to diversity but ...

Contributions of universality-driven initiatives to diversity

- Unification makes true **peculiarities more visible**
- Promoting **inclusiveness**:
  - centralized infrastructures
  - easy integration of new experts, grassroot contribution
  - open licenses
  - UD: 35 **endangered/vulnerable** languages, 12 **extinct**, 20 **underrepresented** in NLP although demographically strong
- **multilingual** tools - universal algorithms, language-specific input data
- **cross-lingual** tools - tools for a language benefit from resources in other languages
  - annotation transfer (Yarowsky et al., 2001; Hwa et al., 2005)
  - model transfer (Zeman and Resnik, 2008; McDonald et al., 2011; Agić et al., 2014; Ponti et al., 2018)
  - corpus sampling for variety (de Lhoneux, 2017)
- **multilingual** language **models** (Devlin et al. 2019), expected to encode statistical universals, **fine-tuned** for languages with few or no data

Objectives

General aim

To reconcile language diversity with rapid progress in language technology.

Research coordination objectives

- **quantifying** inter- and intra-linguistic diversity
- **common understanding of language universals** (across 70 languages)
- **coordinating** unified language **resources** (i.a. in UD and PARSEME)
- better coverage of inter-/intra-linguistic **diversity** in NLP **tools**
- **raise awareness** about **diversity preservation** in NLP

Capacity-building objectives

- **network** of experts (morphology, syntax, semantics) in many languages
- promotion of **young** researchers and **inclusiveness** countries
- coordinating **universality**-driven initiatives (also outside Europe)
- **roadmap** for the joint efforts of the universality-driven NLP community

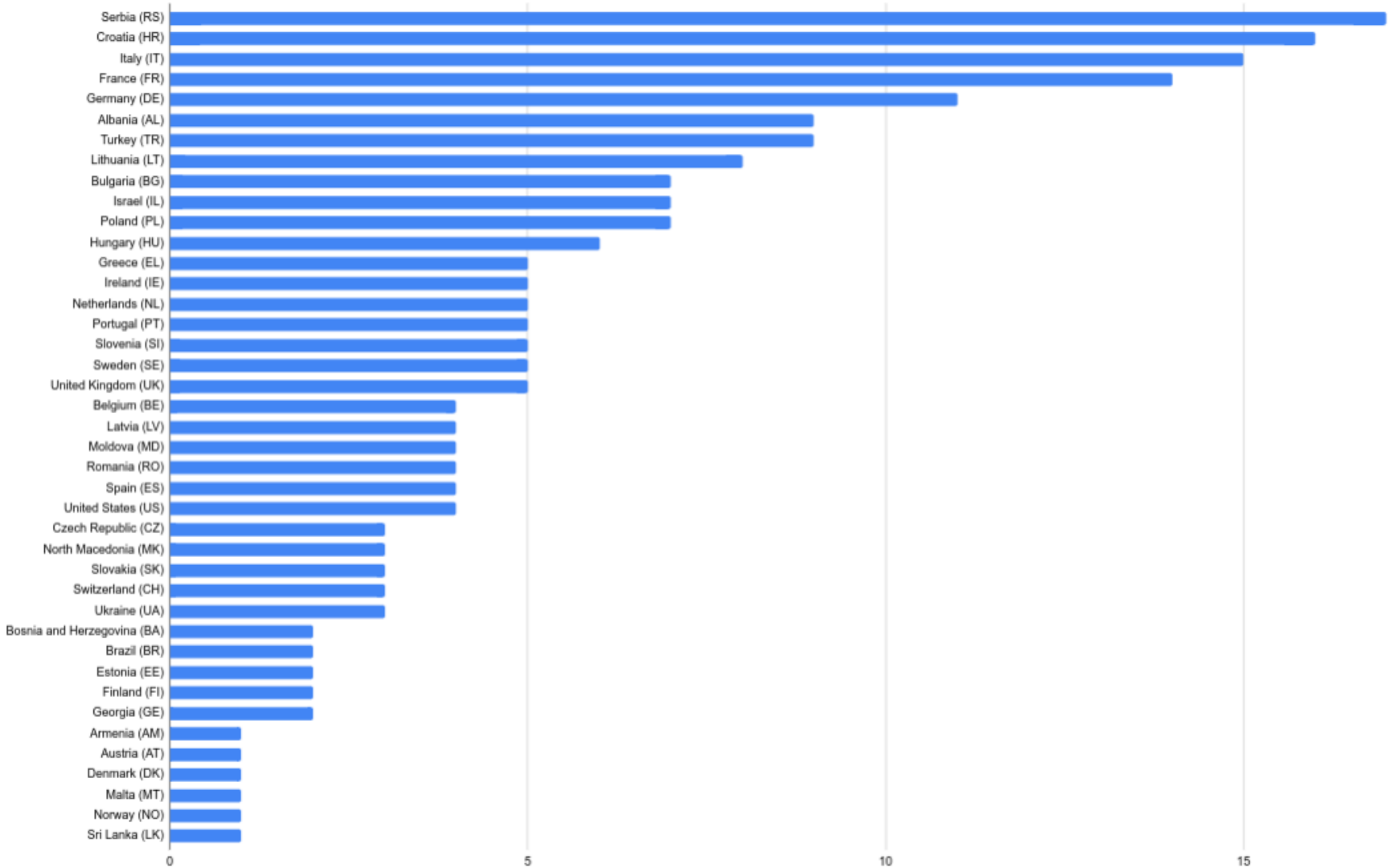
Working Groups

- WG1: Corpus annotation
- WG2: Lexicon-corpus interface
- WG3: Multilingual and cross-lingual language technology
- WG4: Quantifying and promoting diversity

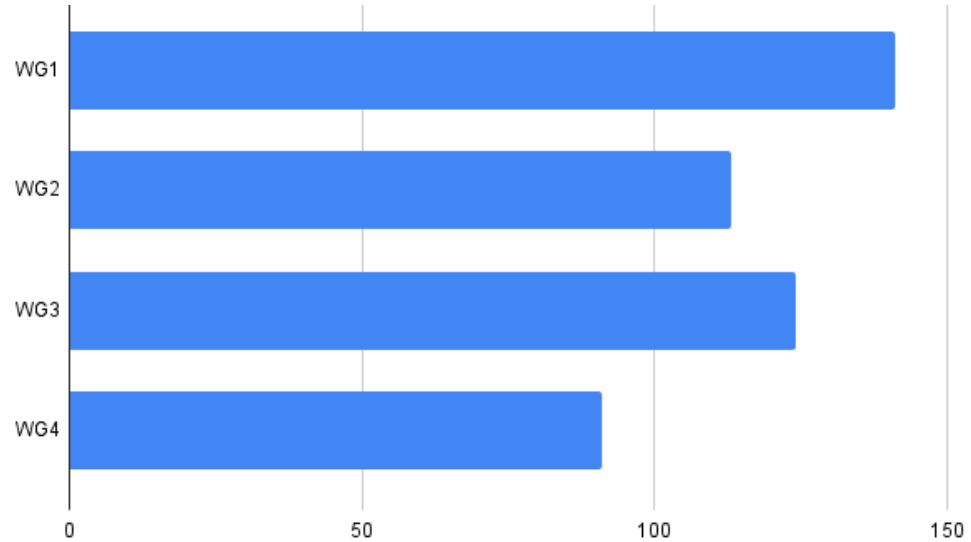
Membership

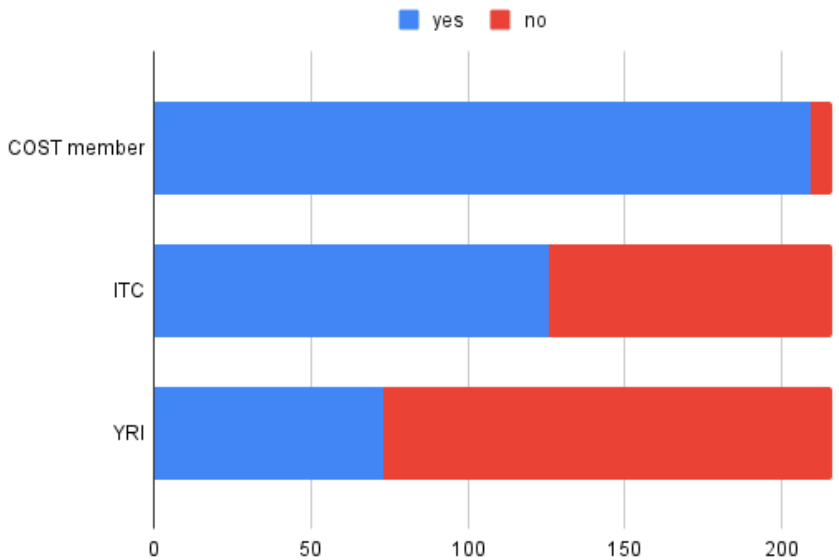
- Open to all relevant researchers and innovators for all duration of the Action
- Young Researchers and Innovators particularly welcome
- See **How to join us** on the Wiki website

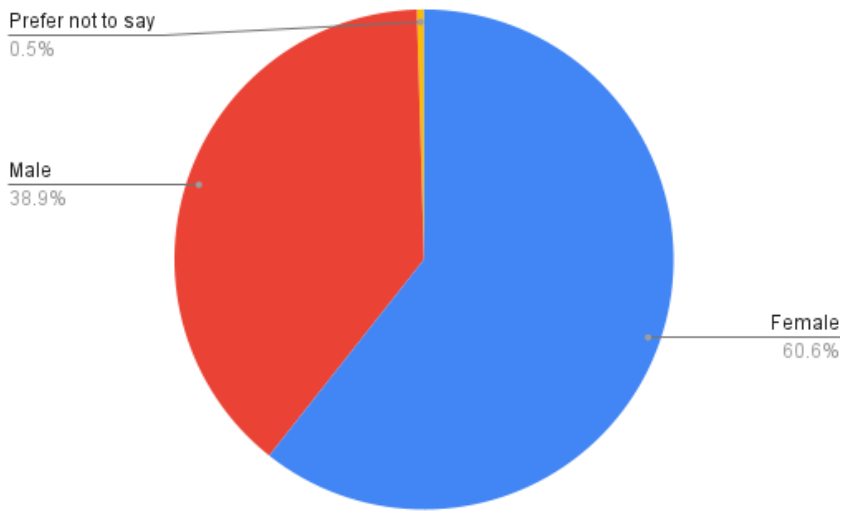
WG members (234 WG applications, 216 approved)



WG members


WG1
WG2
WG3
WG4
0
50
100
150


yes
no
COST member
ITC
YRI
0
50
100
150
200


Prefer not to say
0.5%
Male
38.9%
Female
60.6%

Action Management

Management Committee

- Up to 2 official representatives of all COST countries
- Nominated by their COST National Coordinators (CNCs)
- Validated by an MC vote

MC kick-off: Brussels, 23 Sep 2023

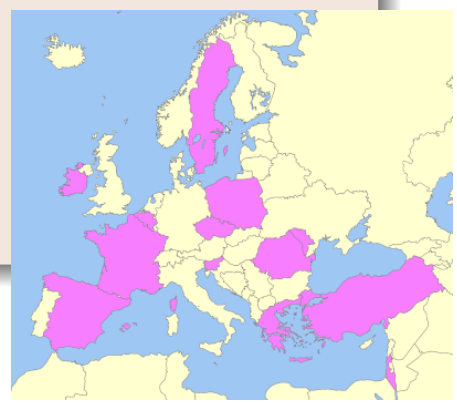
- Introduction to COST
- Breakout rooms
- Delegation of powers to the Core Group

Core Group meetings

- Every-day organization of UniDive
- **Monthly** online meetings
- **Minutes** available on the wiki website

Extended Core Group

- MC Chair: **Agata Savary** (France) (YRIs in green)
- MC Vice-Chair: **Daniel Zeman** (Czech Republic)
- WG1 (Corpus annotation):
  - Leader: **Carlos Ramisch** (France)
  - Co-leader: **Kaja Dobrovoljc** (Slovenia)
- WG2 (Lexicon-corpus interface):
  - Leader: **Verginica Barbu Mititelu** (Romania)
  - Co-leader: **Voula Giouli** (Greece)
- WG3 (Multilingual and cross-lingual language technology):
  - Leader: **Joakim Nivre** (Sweden)
  - Co-leader: **Gülşen Eryiğit** (Turkey)
- WG4 (Quantifying and promoting diversity):
  - Leader: **Marie-Catherine de Marneffe** (Belgium)
  - Co-leader: **Abigail Walsh** (Ireland)



Extended Core Group

- Grant Holder
  - Grant Holder Scientific Representative: **Alina Wróblewska** (Poland)
  - Grant Holder Manager: **Beata Wójtowicz** (Poland)
- Grant Awarding:
  - Coordinator: **Nurit Melnik** (Israel)
  - Vice-Coordinator: **Stella Markantonatou** (Greece)
- Science Communication
  - Coordinator: **Olesea Caftanatov** (Moldova)
  - Vice-Coordinator: **Anabela Barreiro** (Spain)
- Young Researcher and Innovator Representative: **Atul Kumar Ojha** (Ireland)
- **Open positions**: Task leaders (whenever tasks are defined in or across WGs)

Work and budget plan for year 1

Year 1 goals

- Establishing internal and external **communication means**
- Establishing **working relationships** and structuring the community around the **Working Groups**
- **Planning** the activity of each Working Group
- Understanding the **state of the art** for the work program of each WG
- Extending the Action to **new** potential **countries, languages and dialects**
- Facilitating and **coordinating** the development of language resources and tools for new **under-resourced languages**

Budget

Budget year 1

1 Nov 2022 – 31 Oct 2023

Allocated budget

- Initial budget: **125,000€**
- Ongoing amendment: **40,000€**
- Total: **165,000€**

Communication

Websites

- Official COST website: `https://www.cost.eu/actions/CA21167/`
- Wiki website: `https://unidive.lisn.upsaclay.fr/`

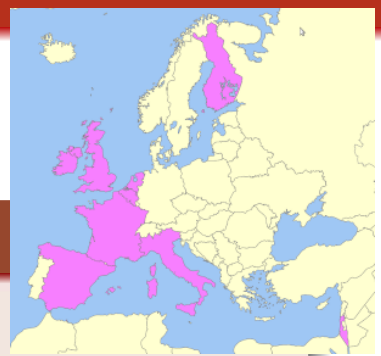
Mailing lists **@lisn.upsaclay.fr**

- unidive-**all** (WG and MC members), unidive-**mc**, unidive-**core**, unidive-**ext-core**, unidive-**wg1**, unidive-**wg2**, unidive-**wg3**, unidive-**wg4**

Upcoming

- Survey on additional communication and document sharing **channels**
- Survey on **expectations** from the UniDive members (**volunteers welcome!**)
- **Logo** contest

Grants



Short-Term Scientific Missions

- Dec 2023: call for STSM proposals
  - **12** submissions, **7** selected, 5 on stand-by, 2 rejected
- permanent call for hosting institutions - **2** submissions
- more **submissions from ITCs** encouraged

Conference Grants

- Small budget for ITC or Dissemination Conference Grants
- Call to appear

Future events

Online MC meeting

22 March 2023

Online webinar

June 2023

WG meeting

Istanbul, Turkey

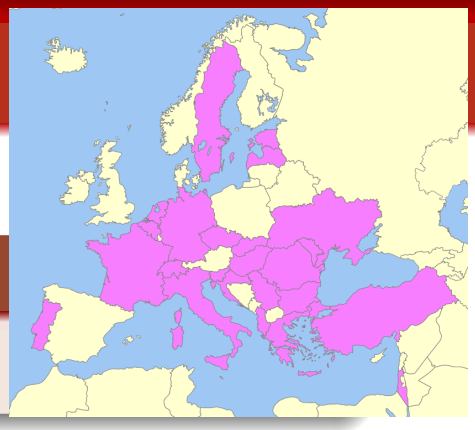
8 September 2023

Istanbul Technical University

WG3; maybe parallel WG1 Turkic session

Local Organizer: **Gülşen Cebiroğlu Eryiğit**

This meeting



Plenary talks

Origins of UniDive.

Posters

- 66 submissions, 4 withdrawn, 38 selected, acceptance 62,5%
- 24 author affiliation coutries, 69 explicitly mentioned languages, 14 multi-/cross-lingual methods

WG meetings

- Getting to know each other, brainstorming of workplans
- Plenary wrap-up session

Ackowledgements

Sponsors

- Laboratoire Interdisciplinaire des Sciences du Numérique (LISN), Paris-Saclay University
- Graduate School in Computer Science, Paris-Saclay University
- Solid States Physics Laboratory (LPS), Paris-Saclay University

Organizers

- LISN: Sophie Rosset (LISN Director), Bénédicte Daly, Till Überrück-Fries
- Program Committee (36 members)
- PC Chairs: Atul, Joakim, Carlos
- WG Leaders: Abigail, Carlos, Gülşen, Joakim, Marie-Catherine, Kaja, Verginica, Voula

Providers

- CESFO (lunches, coffee breaks)
- Solen Boivin and Caroline Widmer (bird watching)
- Campanile Hotel

Questions?