UniDive core group meeting 9

29 September 2023

See also the CG master document and the UniDive master document

Attendants

Atul, ~~Beata~~, Bruno, Dan, ~~Gülşen~~, ~~Joakim~~, Marie, ~~Nurit~~, Olesea, Verginica, Agata

Agenda/minutes

- [WG leaders: 10-20 min.] - Short feedback from the ongoing tasks
    - [AGATA for JOAKIM] WG3 - feedback from the Istanbul meeting
        - 2 ongoing tasks:
            - Documenting documenting multilingual language technology
            - Multilingual evaluation campaign promoting universality and diversity in language technology
                - Presentation of 4 proposals submitted to a call for ST ideas
                - A meeting of ST proposers is upcoming
    - [MARIE] WG4:
        - 2 meetings online; objective: develop a measure of diversity for a shared task; looking at past STs to estimate how well various languages are covered
    - [BRUNO] WG1: four tasks (with 4 or 5 co-leaders in each)
        - T1.1: Linguistic Typology and multilingual corpus annotation
        - T1.2: Extension and updates to MWE annotation guidelines / convergence with UD
        - T1.3: Extensions and updates to morphosyntactic annotation guidelines
        - T1.4: Sharing tools, formats, and infrastructure
        - Next meeting **25 October 11:00 CEST**
    - [VERGINICA] WG2 - monthly meetings:
        - T2.1: ongoing survey "What is a word in UD?", based on Haspelmath's paper; 2 calls so far, deadline: 15th Oct; by 5th Sept: 23 responses - 26 langs

- - - T2.2: extending the coverage of a semantically disambiguated parallel 10-lang corpus with new langs (3 new represented); adding UD syntactic parsing, MWEs and NEs; challenge: making the repository of sense openly available (for some languages)
    - T2.3: survey on   - phase: creating the pool of lexicons to be investigated
    - Next meeting **5th October**, 13:00 CEST
  - Interactions
    - WG3/WG4: Interfacing WG3 survey on multilingual tools with WG4 survey on inter-linguistic diversity - how to survey the results of multilingual systems on particular languages, and how to establish the diversity of a systems' results on this basis?
    - WG1/WG2-T2.2/: guidelines for MWEs identification (and classification)
    - WG1 and Task2.3: how MWE are encoded
    - WG1 1.2 and Task 2.1: words vs. corpus annotation
- [ALINA, AGATA: 15 min.] **COLING/LREC paper** submission about UniDive (to be a reference paper for the Action), category V (International and national activities, projects and initiatives), 4 pages, submission deadline **13 October**. Draft plan:
  - State of the art
  - Objectives of the action (with the logo) + instruments (meetings, STSMs, training schools, workshops)
  - Statistics of the action: countries, languages, dialects
    - [DAN, BRUNO] Aggregate the past survey:
      - [WG1 survey](#) (column E) - 73 answers
      - [Communication survey](#) (by Olesea)
      - [UD list of languages](#) (released or ongoing)
      - [post-meeting thought] Maybe also the AUTOGRAMM languages?
      - In future work we could also use [Glottolog](#) (for a more precise identification of languages and dialects)
  - [WG CO-LEADERS] Description of WGs
    - Ongoing tasks
    - Sketch the future plans for the WG → to be aggregated in the last section
  - Plans for the future for the whole Action
  - List of authors: UniDive Extended Core Group + Carlos
  - Deadlines:
    - 29 Sept: Overleaf and distribution of tasks
    - 9 Oct: first complete draft
    - 12 Oct: final version (aggregated and reviewed)
    - 13 Oct morning: submission
- [ALINA: 5 min.] Feedback from the logo contest
  - 25 beautiful proposals
  - Pre-selection - rejections of all proposals not fulfilling formal requirements
  - Vote - very balanced results; 2 winners; post selection by the ECG
  - https://unidive.lisn.upsaclay.fr/doku.php?id=logo
- [OLESEA, AGATA: 15 min.] Communication - current issues

- ○ Need for customizing the website graphical design with the new logo (maybe extra administration permissions are needed) - https://www.inmotionhosting.com/support/edu/dokuwiki/change-dokuwiki-logo/
- ○ Scientific communication /dissemination team
  - ■ Training about how to adapt the communication to the public we target (e.g LinkIn vs . Facebook),
- ○ We have UniDive accounts (to complete by Olesea with people in charge)
  - ■ Facebook - Lucie - https://www.facebook.com/profile.php?id=100095236907425
  - ■ Youtube Channel - https://www.youtube.com/@UniDiveCA21167
  - ■ TikTok
  - ■ LinkedIn - https://www.linkedin.com/company/unidive/about/
  - ■ Twitter
  - ■ Instagram - https://www.instagram.com/unidive__ca21167/?igshid=NGVhN2U2NjQ0Yg%3D%3D
- ○ Creating a LinkedIn profile (proposal from Anabela) - LinkedIN account already created
- ○ Would we like a video (about ideas) presenting the ideas of the action?
  - ■ [OLESEA] Other actions ask professionals to do it
  - ■ [AGATA] Maybe in budget year 3?
- ● [VERGINICA: 5 min.] Feedback from the COST workshops on EC research policy:
  - ○ Science for Policy WS: Engaging with policy makers (at local, national, European level) early in the duration of the action and keeping them co-interested
    - ■ Inviting Maria (ELE) to the Istanbul meeting was a good example
    - ■ Could we think of other such people whom we could involve? Maybe at the national level (talk to the MC).
  - ○ Science Diplomacy training: diplomacy in science: in international collaboration, with reasons such as access (knowledge and technology elsewhere), promotion (of own achievements abroad) or influence (other country's public opinion and leaders); SD uses strategic (policies, strategic docs), operational (cooperations, boards, networks, S&T events) and supporting (trainings, awareness building events with policymakers, personal networking) tools
    - ■ Concrete steps to take by UniDive not quite clear at this stage
- ● [AGATA: 5 min.] Confirming the dates for the 2nd General Meeting in Naples:
  - ○ the WG1 Meeting on **Wednesday 7 Feb**
  - ○ the 2nd General Meeting on **Thursday-Friday 8-9 Feb**
- ● [AGATA: 5 min.] Choosing the dates for the 1st Training School in Jerusalem:
  - ○ The poll showed preference for 29/07-2/08 but this slot is impossible for Chaya after all
  - ○ Remaining slots:
    - ■ July 8-12 2024
    - ■ July 15-19 2024
  - ○ No issues about both slots from the CG

- - [MARIE] Should we first find trainers and then decide about the slot?
- [BEATA: 5 min.] Current state of year 1 budget
  - The cost of the Istanbul meeting: 30 900 EUR (including LOS 2000 EUR)
  - In the budget for meetings we still have 9700 EUR
  - STSMs, we still have 4555 EUR
  - Conferences 530 EUR
  - For dissemination we spend 7500 EUR (1000 logo, 500 Olesea, 6000 website) instead of 3000
  - Savings - dissemination = **10 000 EUR** (we still have that amount to spend)
  - STSMs still possible until **15 Oct**
  - Olesea might be interested in submitting a proposal
- [AGATA: 10 min.] Year 2 **Work and Budget Plan**:
  - Global budget: 218,000 EUR (year 1: 165,000 EUR)
  - Events
    - General Meeting 1, February 2024, Naples, Italy
    - UD-MWE workshop, May 2024, Turin, Italy (if accepted)
    - [ATUL] Maybe also Eurasia (EURALI) workshop also to submit to in Turin
      - Joint invited speaker?
      - Joint paper session or joint panel?
      - Provided that both workshops are co-located with LREC-COLING
    - Training School 1, July 2024, Jerusalem, Israel
  - What should be our grant period goals (based on the ongoing surveys)?
    - [BRUNO] WG1
      - Increase the convergence of PARSEME and UD in treebanking
    - [VERGINICA] WG2
      - the notion of "syntactic word" across languages: taking stock and developing recommendations for future annotation projects
      - designing a lexicon-corpus interface for interlinking MWE lexicon entries with their occurrences in corpora
    - [JOAKIM] WG3
      - Set up a framework and workflow for documenting multilingual language technology tools within the scope of UniDive.
      - Launch at least one multilingual evaluation campaign promoting universality and diversity in language technology.
    - [MARIE] WG4
      Devise a "diversity" measure to be used in the WG3 shared task
    - [OLESEA] Communication - any specific goals?
      - [AGATA] Enhance the visibility of the Action on social media, integrate the its visual identity (logo) to the communication channels
    - [NURIT] STSMs
      - STSMs, Year 1: spent 33,270 EURO; 15 STSMs [of them 7 last-minute, 12,000 EURO]
      - ITC conference grants, Year 1: spent 970 EURO; 2 ITCs

- - How many STSMs for year 2?
- [5 min.] Forthcoming deliverable(s) - content, WGs involvement, deadlines → to discuss in the next meeting
- Next meeting: **Oct 17, 2023 11:00-12:30**.

TODO

- [AGATA] Publish the minutes from this meeting
- COLING-LREC paper
  - ~~[AGATA: **29 Sept**] Overleaf for the UniDive paper, initial draft, distribution of tasks. Email to all co-authors.~~
  - [ALL ECG: **9 Oct**] First complete draft
  - [AGATA: **11 Oct**]: Final version (aggregated)
  - [ALL: **12 Oct**]: Review of the final version
  - [ALL] Choice of the track - primary and secondary - among:
    - LC17   Multilinguality, Machine Translation, and Translation Aids (including Speech-to-Speech Translation)
    - LC14   Less-Resourced/Endangered/Less-studied Languages
    - LC03   Corpora and Annotation (including Tools, Systems, Treebanks)
    - LC08   Evaluation and Validation Methodologies
    - LC15   Lexicon and Semantics
    - LC22   Parsing, Tagging, Chunking, Grammar, Syntax, Morphosyntax, Morphology
    - Only the first, no secondary track
  - Topics:
    - Collaborative Resource Construction & Crowdsourcing
    - Corpus (Creation, Annotation, etc.)
    - Evaluation Methodologies
    - Less-Resourced/Endangered Languages
    - Lexicon, Lexical Database
    - Linked Data
    - LR National/International Projects, Infrastructural/Policy issues
    - Multilinguality
    - MultiWord Expressions & Collocations
    - Morphology
    - Parsing, Grammar, Syntax, Treebank
    - Part-of-Speech Tagging
  - [AGATA: 13 Oct morning] Submission
- WBP
  - ~~[ALL CG MEMBERS: **OCTOBER 6**] Any more Grant Period goals for year 2?~~
  - [AGATA] Announce the dates of the 2024 events
  - [AGATA] Talk to Chaya about the slot for the TS (keep the choice open)
  - [AGATA] Draft the WBP for year 2 → submit to the ECG
- Communication
  - ~~[OLESEA, AGATA] Integrate the UniDive logo on the website~~

- - - [OLESEA] Send the links to the Action profiles on social media and the list of people who administer them
    - [AGATA] Respond to Anabela (LinkedIN already handled)
    - [ALL] Think about people to involve in UniDive who would bring us closer to science policy making (also at the national level)
    - [AGATA] Add this to the MC agenda
- [AGATA] Next CG meeting agenda
  - Organization of the Naples meeting (posters? Invited talk? WG meetings in parallel?)
  - Ideas of courses and trainers for Jerusalem
  - Forthcoming deliverable(s) - content, WGs involvement, deadlines
- [AGATA] Organize the new MC meeting
  - Poll
  - Announcement, Zoom link
  - [BEATA] Invitation
  - Agenda
    - Involving policy makers in UniDive
    - Year 1 report
    - Presentation of the WBP

UD languages (released treebanks: Dan)

Abaza, Afrikaans, Akkadian, Akuntsu, Albanian, Amharic, Ancient Greek, Ancient Hebrew, Apurina, Arabic, Armenian, Assyrian, Bambara, Basque, Beja, Belarusian, Bengali, Bhojpuri, Bororo, Breton, Bulgarian, Buryat, Cantonese, Catalan, Cebuano, Chinese, Chukchi, Classical Chinese, Coptic, Croatian, Czech, Danish, Dutch, English, Erzya, Estonian, Faroese, Finnish, French, Frisian Dutch, Galician, German, Gheg, Gothic, Greek, Guajajara, Guarani, Hebrew, Hindi, Hittite, Hungarian, Icelandic, Indonesian, Irish, Italian, Japanese, Javanese, Kaapor, Kangri, Karelian, Karo, Kazakh, Khunsari, Kiche, Komi Permyak, Komi Zyrian, Korean, Kurmanji, Kyrgyz, Latin, Latvian, Ligurian, Lithuanian, Livvi, Low Saxon, Madi, Maghrebi Arabic French, Makurap, Malayalam, Maltese, Manx, Marathi, Mbya Guarani, Moksha, Munduruku, Naija, Nayini, Neapolitan, Nheengatu, North Sami, Norwegian, Old Church Slavonic, Old East Slavic, Old French, Old Irish, Old Turkish, Persian, Polish, Pomak, Portuguese, Romanian, Russian, Sanskrit, Scottish Gaelic, Serbian, Sinhala, Skolt Sami, Slovak, Slovenian, Soi, South Levantine Arabic, Spanish, Swedish, Swedish Sign Language, Swiss German, Tagalog, Tamil, Tatar, Teko, Telugu, Thai, Tupinamba, Turkish, Turkish German, Ukrainian, Umbrian, Upper Sorbian, Urdu, Uyghur, Vietnamese, Warlpiri, Welsh, Western Armenian, Western Sierra Puebla Nahuatl, Wolof, Xavante, Xibe, Yakut, Yoruba, Yupik, Zaar

Languages from the WG1 survey (Bruno)

- 73 people answered
- 67 languages (with the number of times they were mentioned):

44 English
11 French
 8 Serbian
 7 German
 5 Spanish
 5 Russian
 5 Latin
 5 Italian
 5 Albanian
 4 Slovenian/Slovene
 4 Romanian
 4 Hindi
 4 Chinese
 3 Turkish
 3 Swedish
 3 Slovak
 3 Portuguese
 3 Polish
 3 Macedonian
 3 Lithuanian
 3 Hebrew
 3 Dutch
 3 Croatian
 2 Sanskrit
 2 Naija
 2 Korean
 2 Hungarian
 2 Greek
 2 Georgian
 2 Czech
 2 Beja
 2 Armenian
 2 Arabic
 1 Yoruba
 1 Wolof
 1 Warlpiri
 1 Vietnamese
 1 Urdu
 1 Upper Sorbian
 1 Ukrainian
 1 Tamil
 1 Tagalog
 1 Surzhyk
 1 Pomak
 1 Persian
 1 Old Italian
 1 Neo-Aramaic
 1 Modern Greek
 1 Mingrelian
 1 Maltese
 1 Malayalam

1 Magahi
1 Latvian
1 Latgalian
1 Kurdish
1 Haitian Creole
1 Frisian
1 Finnish
1 Farsi
1 Estonian
1 Catalan
1 Bulgarian
1 Bhojpuri
1 Azeri Turkic
1 Awadhi
1 Ancient Greek
1 Afrikaans