

Courses at the 1st UniDive training school

Back to the [UniDive 1st training school page](#)

Dependency syntax, Surface-Syntactic UD, and UD

- **Trainers**

- [Sylvain Kahane](#) (Université Paris Nanterre and Institut Universitaire de France)

- **Objectives:**

- Recall the principles of dependency syntax (already done in Naples but necessary to start)
 - Connectedness: words that form a unit must be connected
 - Headedness: the distribution of a unit is generally controlled by one word
 - Categories: words must be assigned the same category if and only if they can occupy the same positions
 - Relations: similar constructions with similar properties must be labeled with same syntactic relation
- SUD annotation scheme
- Conversion between SUD and UD
- A word on mSUD, SUD annotation at the morph level for people starting with Interlinear Glossed Texts

- **Exercises:**

- understanding the [SUD](#) (and [UD](#)) annotation scheme by exploring some treebanks with [Grew-match](#) (SUD_English, converted from UD; [SUD_Naija](#), a native SUD treebank of a pidgin creole of English; [mSUD_Beja](#), a native morph-based SUD treebank glossed in English) (joint session with Bruno Guillaume?)
- example of a SUD annotation from scratch based on data from the participants which are glossed and translated in English
 - creation of a project on ArboratorGrew
 - annotation on ArboratorGrew
 - automatic completion of the annotation with Grew

- **Pre-requisites:**

- being concerned by syntactic annotation
- ideally, having some data you want to annotate (please take contact before the summer school for the preparation of the data)

- **Preparatory work** (offered in a parallel course by Bruno Guillaume):

- looking at treebanks on Grew-Match
- comparing UD and SUD annotation

- **Further readings:**

- Lucien Tesnière (2015), [Elements of structural syntax](#), Benjamins. ebook in open access.
- Igor Mel'cuk (1988), *Dependency syntax: theory and practice*. SUNY press.
- Timothy Osborne (2019), *A Dependency Grammar of English*. Benjamins.
- Sylvain Kahane, 2003, [The Meaning-Text Theory](#), in *Dependency and Valency, Handbooks*

of Linguistics and Communication Sciences, 25 : 1-2, Berlin/NY: De Gruyter, 32 p.

- De Marneffe, M. C., Manning, C. D., Nivre, J., & Zeman, D. (2021). [Universal dependencies](#). Computational linguistics, 47(2), 255-308.
- Gerdes K., Guillaume B., Kahane S., Perrier G. (2018) [SUD or Surface-Syntactic Universal Dependencies: An annotation scheme near-isomorphic to UD](#), Proceedings of the Universal Dependencies Workshop (UDW), EMNLP.
- Gerdes K., Guillaume B., Kahane S., Perrier G. (2021) [Starting a new treebank? Go SUD!](#), Proceedings of 6th international conference on Dependency Linguistics (DepLing), SyntaxFest, ACL.

Annotation of multiword expressions for newcomers

• Trainers

- [Verginica Mititelu](#) (Romanian Academy, Bucarest, Romania)
- [Voula Giouli](#) (ATHENA Research Centre, Athens and Aristotle University of Thessaloniki, Greece)

• Objectives: Upon completion of the course, the trainees will be able to

- define the notion of multiword expressions and manually identify them in corpora;
- classify the multiword expressions identified in corpora according to the types defined for them in PARSEME and UniDive;
- briefly describe this phenomenon in their language

• Form of instruction

- lectures
- practical exercises - annotating the corpus prepared by the trainees
- the last session will give the trainees the opportunity to present their observations from the practical exercise

• Contents

- The course will start from the definition and characteristics of multiword expressions as agreed upon in PARSEME and UniDive, telling them apart from other phenomena involving word combinations. Results already obtained in PARSEME and beyond will also be presented, highlighting the importance of corpora annotation for tasks of automatic identification of the phenomenon in corpora and for lexicography.
- The decision trees for establishing the status of multiword expressions of several types (verbal, nominal, modifier) will be presented, with examples from languages (preferably from different language families and types).
- Having access to this knowledge, the trainees will then be involved in a practical session, in which they will annotate a corpus of their own language with the types of multiword expressions presented in the theoretical part of the course.
- Insights from the annotation will be presented by the trainees and will be recorded so as to be further taken into account as feedback for the annotation guidelines.

• Pre-requisites

- theoretical linguistics knowledge (parts of speech, inflection, syntactic structures)
- Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. In Nitin Indurkha and Fred J. Damerau, editors, Handbook of Natural Language Processing, 2 edition, pages 267-292. CRC Press, Taylor and Francis Group, Boca Raton, FL, USA.

- **Preparatory work:** To be done by the trainees before the training school:
 - prepare a parallel corpus or a monolingual one; it would preferably contain a new language, a new dialect, or a new genre; by “new” we mean “not already covered in the PARSEME 1.3 corpus”.

Corpus annotation infrastructure

- **Trainers**
 - [Bruno Guillaume](#) (INRIA, LORIA, France)
 - [Daniel Zeman](#) (Charles University, Czech Republic)
 - [Agata Savary](#) (Université Paris-Saclay, CNRS, LISN, France)
- **Objectives:**
 - Understand and efficiently use the technical infrastructure supporting UD and PARSEME corpus annotation and query
- **Form of instruction**
 - mostly practical exercises in corpus querying and processing
- **Contents (not necessarily in chronological order)**
 - Session 1 (by Bruno Guillaume), joined with Sylvain's course in dependency syntax
 - Storage formats of data: ConNLL-U, CUPT
 - Basic usage of Grew-match of morpho-syntactic treebanks
 - Hands-on: observe main difference between UD and SUD
 - ArboratorGrew basic usage: users roles, graphical edition, conllu edition, metadata
 - Sessions 2-3 (by Bruno Guillaume)
 - Advanced usages of Grew-match
 - On PARSEME data
 - Usage of clustering / tables for corpus maintenance, error mining and checking annotation consistency
 - Advanced usage of ArboratorGrew
 - usage of rewriting rules for corpus pre-annotation / maintenance
 - usage of Parser for pre-annotation
 - usage of Github synchronisation
 - Session 4 (by Agata Savary)
 - Git for beginners:
 - a repository, a clone, a commit
 - Git operations: clone, pull, add, commit, push
 - branches
 - Gitlab vs. Github
 - PARSEME Git infrastructure
 - PARSEME project on Git and its repositories
 - Managing language repositories
 - PARSEME utilities
 - PARSEME/UD consistency
 - Sessions 5-6 (by Daniel Zeman)
 - UD GitHub repositories
 - Branches, push access, pull requests
 - How to upload: Use git diff before committing and pushing

- TortoiseGit
- Prescribed structure of the dev branch
 - Do not pull history from the master branch
 - The docs repository, language-specific documentation
 - Working with personal UD repositories
- Validator
 - On-line report after uploading data
 - How to run locally (there are two scripts!)
 - How to locate and fix the error
 - Demonstrate some common errors, validation levels
 - How to register language-specific features, relation subtypes, auxiliaries
 - How to fix documentation errors (demonstrate)
- Fixing the errors
 - Annotation tool (cf. Grew)
 - Text editor (do not use Word!)
 - Udapi
- UD Github issues: asking for linguistic help in docs, reporting bugs in treebank-specific repos
 - Referring to particular commits, files and lines in the repo.

From:
<https://unidive.lisn.upsaclay.fr/> - **Universality, diversity and idiosyncrasy
in language technology
CA21167 COST Action**

Permanent link:
https://unidive.lisn.upsaclay.fr/doku.php?id=meetings:other-events:1st_unidive_training_school:courses&rev=1714391250

Last update: **2024/04/29 13:47**

