JOINT EVENT CORLI, LIFT, TAL

# WORKSHOP

## ANNOTATION COLLABORATIVE DE CORPUS

## LES 15 & 16 MAI 2023
## UNIVERSITÉ PARIS CITÉ

En parallèle, des sessions de formation CORLI auront lieu les 16 & 17 mai

### Programme

**Lundi 15 mai - 14h à 17h30**
Présentation de CORLI, du GDR TAL et du GDR LIFT
Conférence d'Amir ZELDES suivie d'une table ronde
**Où ?** Salle 720, bât. Olympe de Gouges

**Mardi 16 mai - 10h à 12h30**
Réunion du projet CORLI-GUM
**Où ?** Salle 531, bât. Olympe de Gouges

### Se rendre à l'événement

**Université Paris Cité - Campus Grands Moulins**
8 Place Paul Ricoeur, 75013 Paris
Bâtiment Olympe de Gouges

JOINT EVENT CORLI, LIFT, TAL

# WORKSHOP

## ANNOTATION COLLABORATIVE DE CORPUS

LES 15 & 16 MAI 2023
UNIVERSITÉ PARIS CITÉ

En parallèle, des sessions de formation CORLI auront lieu les 16 & 17 mai

**Conférence d'Amir ZELDES**

**Crafting Rich and Diverse Data for Computational Models of Discourse**

Thanks to pre-trained language models, computational tools are now better than ever at extracting entities, events, and other information from arbitrary text and speech. However, models are only successful and useful when they learn from the right kind of data, making the representations and materials we choose, and what we teach practitioners in the field, more important than ever. In this talk I will present the Georgetown University Multilayer (GUM, https://gucorpling.org/gum/) family of datasets, which are constructed as part of the Computational Linguistics curriculum at Georgetown, and target discourse level representations of spoken and written language across 20 distinct text types. After discussing tools, architectures and pedagogical challenges for developing data in an academic curriculum, I will present a series of studies illustrating the added value of richly annotated resources, the importance of genre diversity, limitations of current models and some new theoretical directions in computational discourse models. Our results show that carefully crafted datasets are crucial for a range of popular tasks, including salient entity recognition, Wikification, coreference resolution, automatic summarization, discourse parsing, and more.