Moldova Training School with UniDive

Andrew Dyer

Language Science and Technology
Universität des Saarländes
Germany

andrew.dyer@uni-saarland.de

cost EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

Date: 8-12 July 2024

# Towards universal annotation guidelines for coreference resolution and information status

## The goal

To develop cross-lingually consistent guidelines for Universal Dependencies-native annotation of coreference and information structure in the world's languages.

## Background

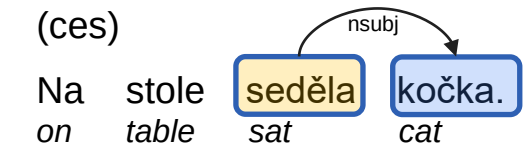
- **CorefUD** and **Universal Anaphora** provide a cross-lingually consistent *format* for layering coreference resolution datasets *on top of* Universal Dependencies.
- Several existing coreference datasets are included in this format, from various languages.
- *However*, these datasets are still not ideal for cross-lingual study.
  - *They were compiled with different ideas of what is a markable in mind.*
  - *Only some contain information status/structure as annotation features.*

## Our use case

We are **computational linguistic typologists** studying the effects of information structure on word order.

- *Topic/focus, givenness/newness*
- How do these impact word order rules and choices in the world's languages?

(ces)



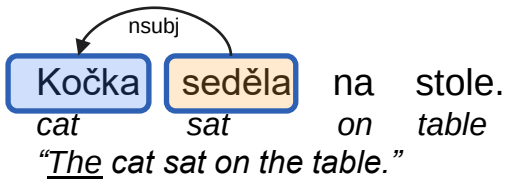




Fig 1: An example of the effect of information status in Czech. The surface forms of each token are identical, but the word order communicates givenness and definiteness. In this case, *a cat* (new) or *the cat* (given).

## Our work so far

- We are annotating coreference and information status on top of a *parallel, parsed multilingual* corpus.
- We have annotated data for *English, Portuguese, Greek, Ukrainian, Indonesian and Turkish*.
- We have tried to maintain common guidelines that apply parsimoniously to all languages.

## Challenges

Keeping guidelines consistent between languages is a challenge. Some linguistics issues:

- **Null anaphora/zero tokens:** To capture "pro-drop", we have included these when referents are *indexed* through morphology. However, this does not capture non-indexing pro-drop.
- **Inclusion of topic/focus:** We have restricted ourselves to given/new, as topic/focus is subjective and difficult to annotate. But should we include this?
- **"Accessible" information status:** Is it possible to define what is *accessible* universally?

## Over to you

- How does information status impact word order in your language?
- What do we need to consider when annotating coreference and information status for your language?

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 232722074 – SFB 1102

European Union

UNIVERSITÄT DES SAARLANDES