# Annotated Corpora and UCCA-App

*Dotan Dvir*
*COLING 2020 UCCA Tutorial*

https://github.com/UniversalConceptualCognitiveAnnotation/tutorial

# Annotated Data

| | | Tokens |
|---|---|---|
| English | Wikipedia<br>The English Web Treebank (online reviews)<br>20 Thousand Leagues under the Sea<br>The Little Prince<br>Wall Street Journal (financial news) | 159,000<br>56,000<br>13,000<br>100 sentences<br>100 sentences |
| German | 20 Thousand Leagues under the Sea<br>The Little Prince | 145,000 (full version)<br>19,000 (full version) |
| French | 20 Thousand Leagues under the Sea | 13,000 |
| Hebrew | The Little Prince<br>Wikipedia | 22,000 (full version)<br>annotation underway |
| Russian | The Little Prince | annotation underway |

Parallel Corpus

# The Annotation Process

- Annotator recruitment: a linguistic background is not necessarily a requirement

- Training

  – ~ 30-50 hours

  – first stage: self-learning

  – second stage: feedback and agreement tests

- Current workforce: 2 annotators + project manager

3

# The Annotation Process

- Annotation process

  - 2 annotators per passage: annotator and reviewer

  - corrections by the project manager (only clear-cut errors are corrected at this stage)

  - currently exploring: parsing+ manual corrections

- Means of communication with annotators:

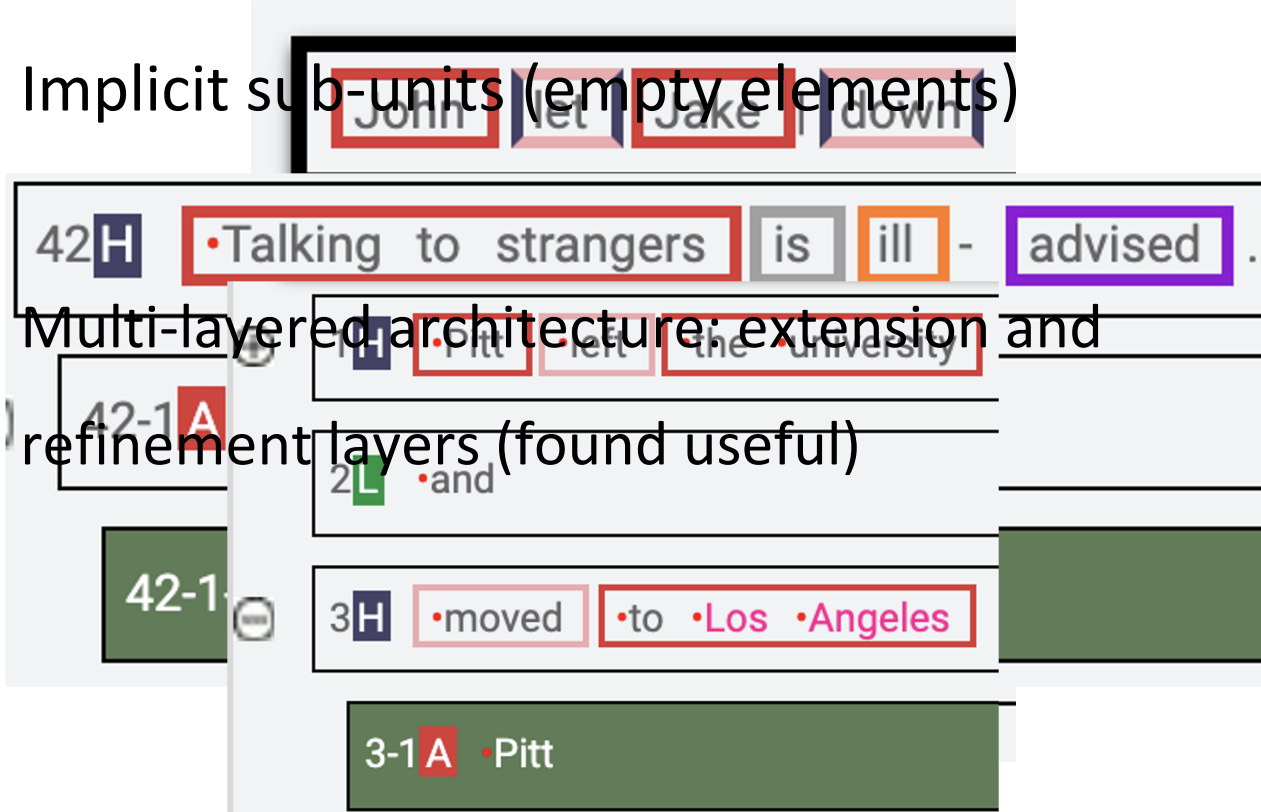  - comments in the task itself

  - meetings

# UCCA-App

- Open source, flexible web application

- UCCA annotation is done on the UCCA-App

- Supports syntactic and semantic phrase-based annotation in general

- Simple and intuitive UI, supports annotation by non-experts

# UCCA-App

Supports various properties such as:

- Discontiguous phrases

- Implicit sub-units (empty elements)

- Multi-layered architecture: extension and refinement layers (found useful)

# UCCA-App

# UCCA Tutorial

1) Bird's Eye View of UCCA - Omri Abend

2) Annotation of English - Nathan Schneider

**3) Annotated Corpora and UCCA-App - Dotan Dvir**

4) Extension Layers and Comparison to Other Formalisms – Jakob Prange

5) Parsing, Evaluation and Applications – Daniel Hershcovich

6) Crosslinguistic Studies – Omri Abend

## Thanks to my co-presenters for feedback!

https://github.com/UniversalConceptualCognitiveAnnotation/tutorial