You get what you annotate:
a pedagogically annotated corpus of coursebooks
for Swedish as a Second Language

Elena Volodina, Ildikó Pilán, Stian Rødven Eide, Hannes Heidarsson

University of Gothenburg, Sweden
COCTAILL time

• COCTAILL - Corpus of CEFR Textbooks as Input for Learner Level modeling
• January 2013-October 2014
• CLT-funded (University of Gothenburg)
Brewing COCTAILL

- corpus design
- interview with L2 teachers
- paper copies
- scan + OCR
- text extraction
- digitalized text
- annotators
- content annotation
- NLP tools
- linguistic annotation
COCTAILLL content “ingredients”

ELEMENTS
Coursebook → ATTRAIBUTES
  Extras → title, author etc.
  Lesson → CEFR level, (title)
  Text → topic (e.g. arts), (title)
  Genre → genre family (e.g. facts)
  Subheadings → macrogenre (e.g. dialogue)
  Activity instruction → skills/competences, activity type/format
  Task → skills/competences, linguistic unit
  List → Language Example
  Common attributes: → id, ref
COCTAILL quantitative explorations: target skills across levels
COCTAILL quantitative explorations: topics across levels

![Graph showing the percentage of topics across different CEFR levels (A1 to C1) for topics like daily life, family and relatives, culture and traditions, and politics and power.](graph_image)
COCTAILLL impact

• A possibility to study coursebooks in their entirety, correlating proficiency levels, L2 input, pedagogical and textual variables

• SLA-aware ICALL-relevant studies, such as CEFR profiling, vocabulary and grammar profiling, studies on sentence and text readability, automatic genre id, etc.

• A crystallized form of what should be taught, at which level and in which format -> for better material generation
Taste our COCTAILL
at the poster stand:

e.g. browse the corpus,
test the annotation tool,
zoom into inter-annotator agreement results, etc.