Encoding Language Learner Corpora

A case study focusing on the KCT and H2 corpora

Matteo Brivio & Pia Schwarz

Universität Tübingen
Seminar für Sprachwissenschaft

HS Linguistic Corpus Annotation
WS 2020/21

Goals

Encode the Karlsruhe Children's Text corpus and the H2 corpus into a computationally digestible format

- Agree on an encoding format
- Develop a corpus model (layers, annotations...)
- Automate as much as possible

Corpora: KCT and H2

- spontaneously written texts from school grades 1-8
- KCT (Karlsruhe Children's Text)
  - 1701 texts + 1-6 jpgs per text
- H2 corpus (H2, E2, ERK1 Children's Writing)
  - 2117 texts + 1 pdf per text

Achieved:

```
\\Kind6-begin
dann noch fernsehen Kucken b[. §]z[. §]w. Horror§film mit chips****{F} und so weiter halt.
\\Kind6-end
\\Kind7-begin
Dan nochmal mit ihn{G} Playstaison{N}_spielen
\\Kind7-end
```

Target:

```
\\Richtig5-begin
Dann heim§gehen und dann noch mit ihm{G} Playstation{N} spielen.
\\Richtig5-end
\\Richtig6-begin
Dann noch Fernseh gucken b[. §]z[. §]w. Horror§film mit Chips****{F} und so weiter halt.
\\Richtig6-end
```

Issues and Obstacles

- orthographic errors in target text: `wei (weiß)`, `Geschindigkeit (Geschwindigkeit)`
- erroneous annotation: `[§/]` instead of `[§ /]` ...
- inconsistency: `lieste{G}{3}` vs. `Schnüfeln{2}{G}`
- combination of annotations symbols
- non-resolvable annotation: `ABFOJA=Aller§Beste Freunde ohne jede ausnahme`

⇒ more issues than expected
⇒ we only covered the KCT corpus
⇒ **but:** with some changes our script allows to also cover H2

Corpus model: layers and annotations

tag und früte sie ch daser lesen Konte und daser imer

⊟ grid (error)

error				§		_				_	
tok	tag	und	früte	sie	ch	daser	lesen	Konte	und	daser	imer

Figure: achieved layer with error annotations

Tag und freute sich , dass er lesen konnte und dass er immer

⊟ grid (pos)

pos	NOUN	CCONJ	VERB	ADV	PUNCT	X	ADV	VERB	VERB
tok	Tag	und	freute	sich	,	dass	er	lesen	konnte

Figure: normalized layer with POS annotations

Dealing with PAULA XML

High modularity, high scalability and *lots* of XML files...

```
KCT
└── 0_0000
    ├── achieved.kct.0_0000.text.xml
    ├── achieved.kct.0_0000.tok.xml
    ├── ...
    └── pos.kct.0_0005.pos_mark.xml
```

Generating XML files

- `lxml` to create XML files
- `HunSpell` to spellcheck the normalized tokens
- `SpaCy` to generate POS tags
- `img2pdf` to convert JPG scans into PDF

Achieved and Target Texts

- Automatic generation of:
  - expanded achieved layer
  - normalized layer

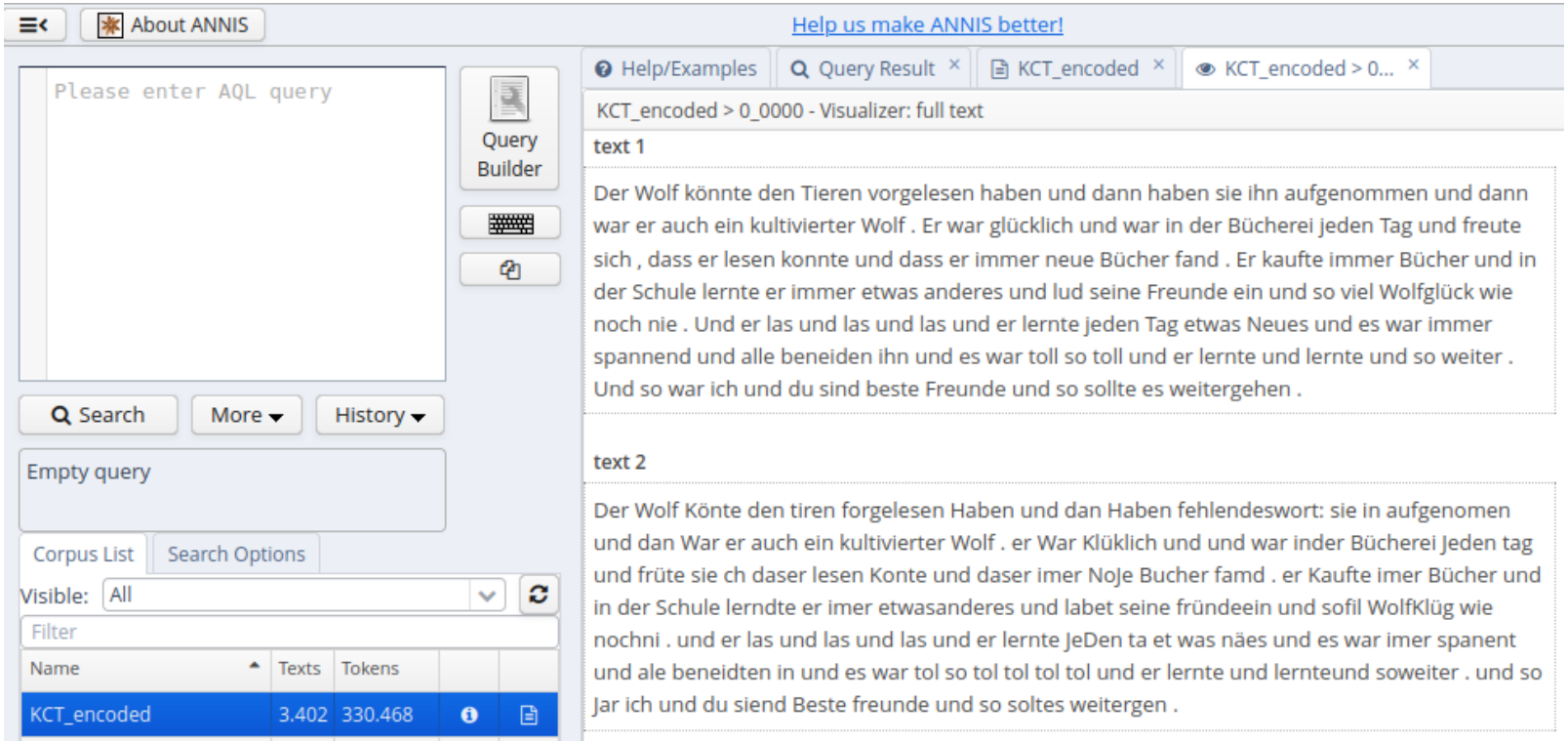
```
achieved:   dan get er in_den Park aba_mir, wan{G} [§ noch] --chlagweilich--e
resolved:   dan get er inden Park abamir, wan  fehlendeswort:noch chlangweiliche
errors:                  _          _      {G}  [§ noch]            -- --

target:     Dann geht er in_den Park aber_mir ist{G} [§ noch] --langweiliger--,
normalized: Dann geht er in den Park aber mir ist noch langweiliger,
```

- Lots of special cases that need extra handling (inflating the python code)
- Normalized and expanded achieved output make up the base layers of the corpus

Converted Corpus

- KCT is now encoded in the PAULA XML format
- Using pepper it is easily convertible (e.g. into Annis)
- Short demo: `error=/\{\d\}/ error=/\[§ \w+\]/`



Have we reached our goal and expectations?

- Yes:
  - automating the process of conversion was successful (despite many special cases)
  - the corpus in PAULA XML format is easily extensible to add further annotation
- No:
  - H2 is not converted
  - There is no *single* base layer to which all annotations are attached, instead we have two (achieved and normalized):

```
target:     Dann geht er in_den Park aber_mir ist{G}   [§ noch]  --langweiliger--,
normalized: Dann geht er in den Park aber mir ist noch INSERTION langweiliger,
```

Where to next?

- Improve POS tagging
- Run the spellchecker on the whole corpus
- Include further annotations
- Polish the code
- Investigate possible ways to align achieved and normalized layers
- Encode H2
- Refine error layer (real errors {2} vs non-real ones {F})

References

Karlsruhe Children's Text:
https://catalog.ldc.upenn.edu/LDC2015T22

H2, E2, ERK1 Children's Writing:
https://catalog.ldc.upenn.edu/LDC2018T05

Lavalley, R., Berkling, K., and Stüker, S. (2015). *Preparing Children's Writing Database for Automated Processing*. In Workshop on L1 Teaching, Learning and Technology (L1TLT), Leipzig, Germany.

Berkling, K. (2018). *A 2nd longitudinal corpus for children's writing with enhanced output for specific spelling patterns*. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).