RIKSBANKENS JUBILEUMSFOND
STIFTELSEN FÖR HUMANISTISK OCH SAMHÄLLSVETENSKAPLIG FORSKNING

GÖTEBORGS UNIVERSITET

SPRÅKBANKENTEXT

SWE-CLARIN

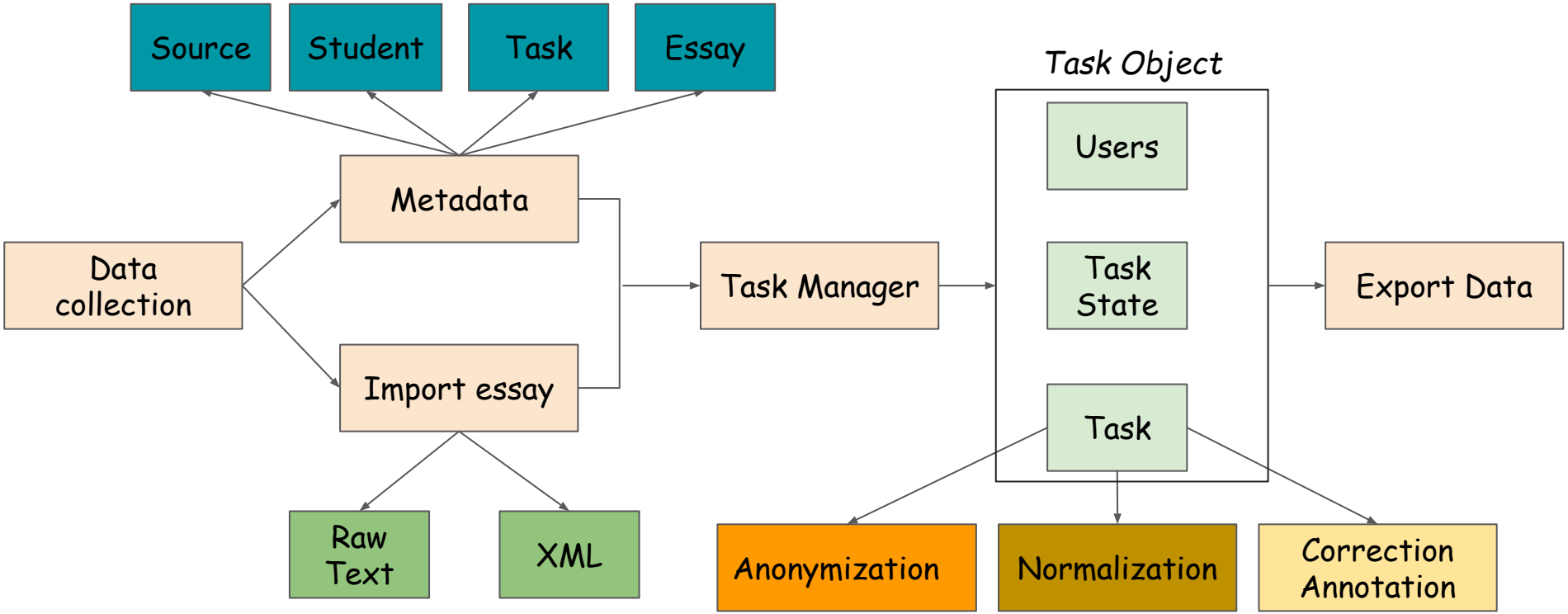
SweLL Portal

Yousuf Ali Mohammed (Samir)

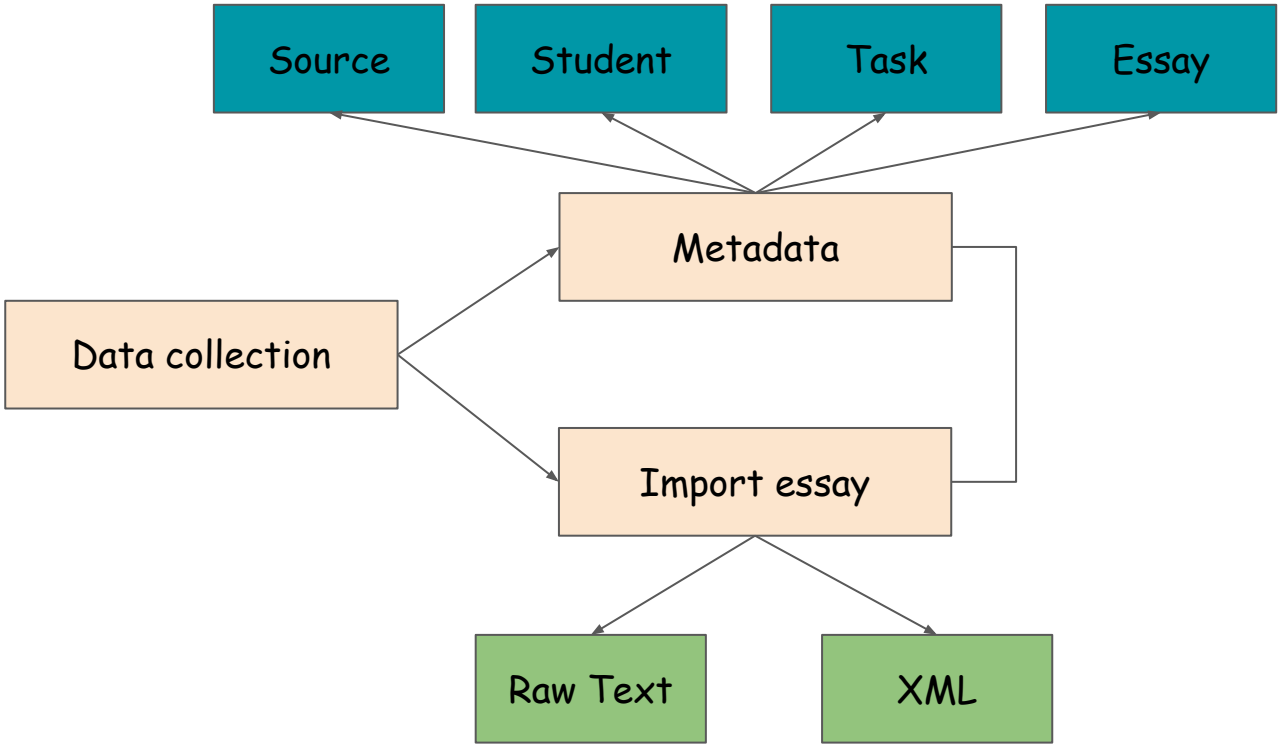
Data collection
and
Data management

Derivative
Datasets

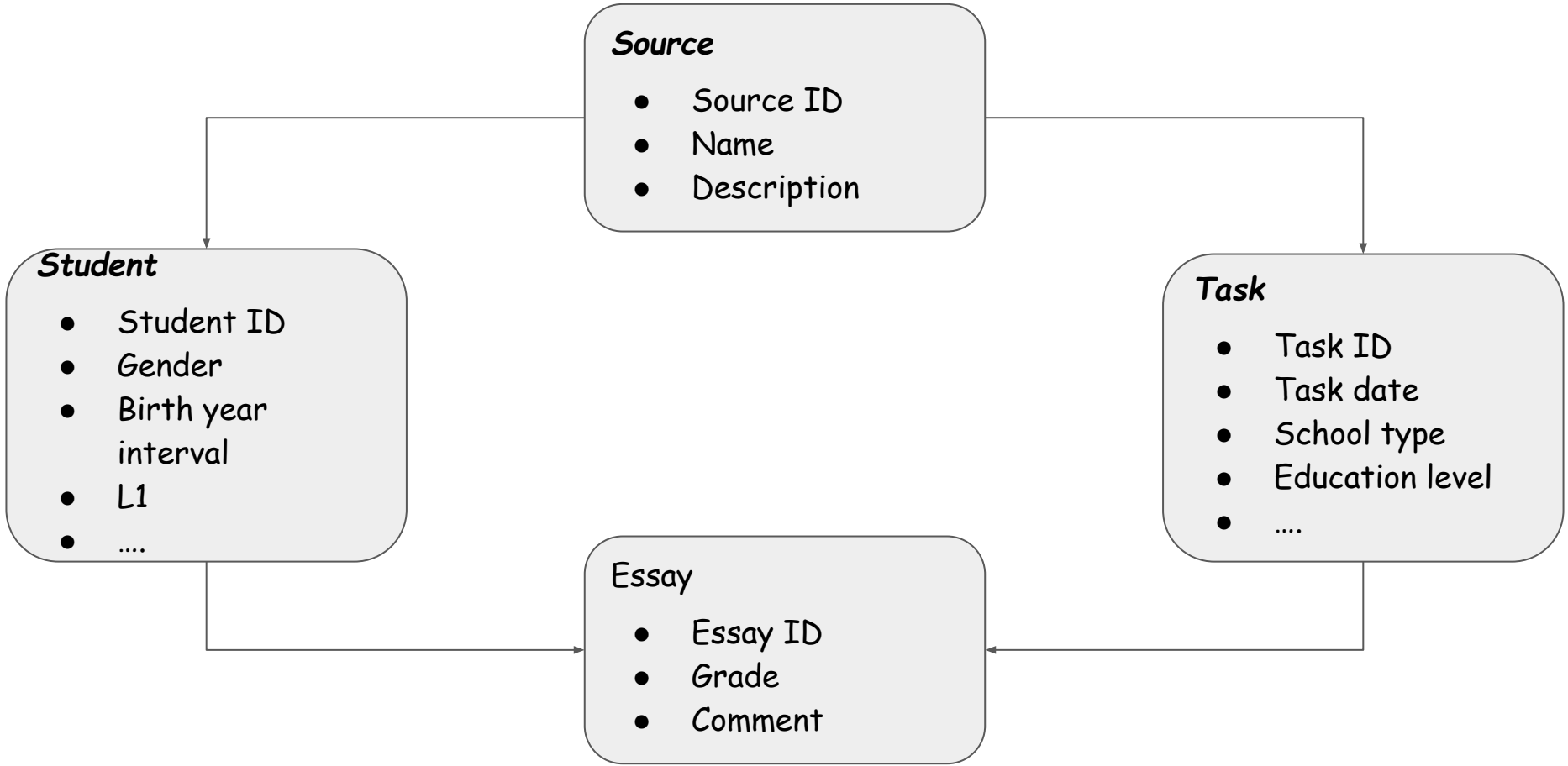
Data collection and Data management



Data collection



Metadata


Source
Source ID
Name
Description
Student
Student ID
Gender
Birth year interval
L1
….
Task
Task ID
Task date
School type
Education level
….
Essay
Essay ID
Grade
Comment

Import Essay

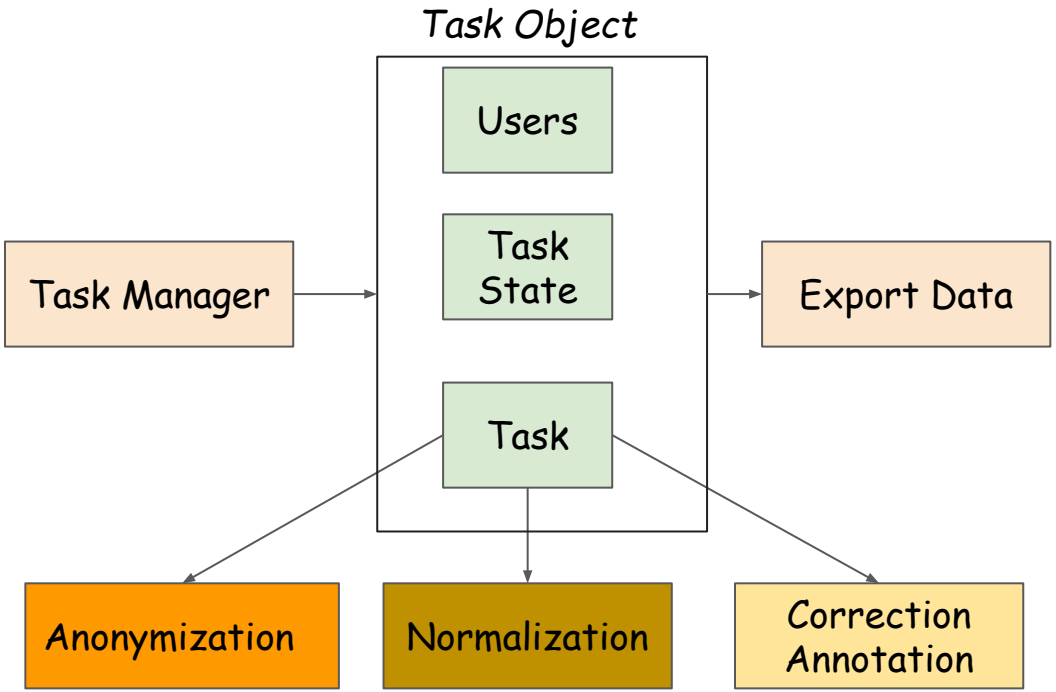
Raw text

Inför gratis körkortsutbildning ␤ Körkort är jätteviktigt för alla människor . Min åsikt är : Inför gratis körkortsutbildning så att alla människor kan lära sig att köra bil . ␤ ␤ För det första behöver man körkort när man söker jobb därför att många företag frågar " har du körkort ? " , och då kan man säga " JA , jag har körkort . " ␤ För det andra tycker jag att körkort gör det lättare att åka någonstans , t.ex. för att handla eller köpa något man behöver . Jag förstår att köra bil inte är bra för miljön , men man behöver inte köra bil varje dag . Man kör bara bil när man vill eller åker långt bort . ␤ För det tredje sparar man pengar därför att man betalar mycket pengar när man går och lär sig i körskolan . ␤ ␤ Mitt förslag är att kommunen ska göra något . Inför gratis körkortsutbildning nu , för alla som är intresserade av att lära sig att köra bil ! ␤ Viktigt för livet

XML

```
<?xml version="1.0" encoding="utf-8" ?>
<swell>
  <meta>
    <student id="A1">
      <language situation="Modersmål">Norska</language>
    </student>
    <task id="AT1" grading_scale="CEFR" />
  </meta>
  <essays>
    <essay student="A1" task="AT1" grade="B2">
      <content type="svala">
        {
          "source": [
            {"id": "s0", "text": "Hello "},
            {"id": "s1", "text": "world "}
          ],
          "target": [
            {"id": "t0", "text": "Hello "},
            {"id": "t1", "text": "world "}
          ],
          "edges": {
            "e-s0-t0": {
              "id": "e-s0-t0",
              "ids": ["s0", "t0"],
              "labels": [],
              "manual": false
            },
            "e-s1-t1": {
              "id": "e-s1-t1",
              "ids": ["s1", "t1"],
              "labels": [],
              "manual": false
            }
          }
        }
      </content>
    </essay>
  </essays>
</swell>
```

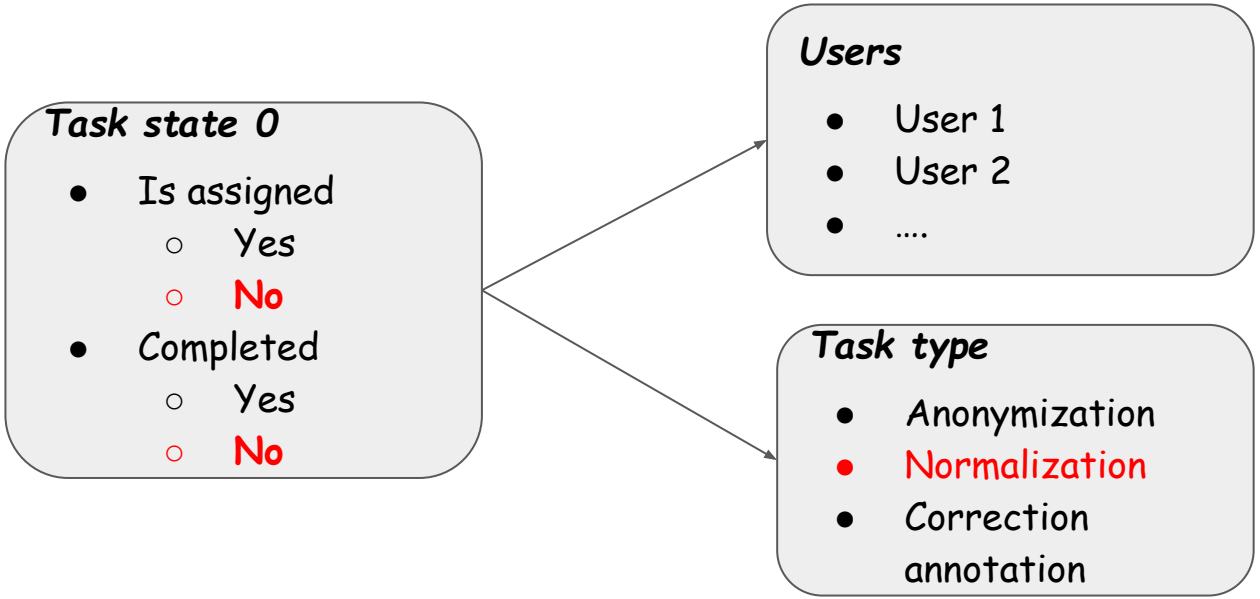
Data management



Task manager (step 0)

Essay_id

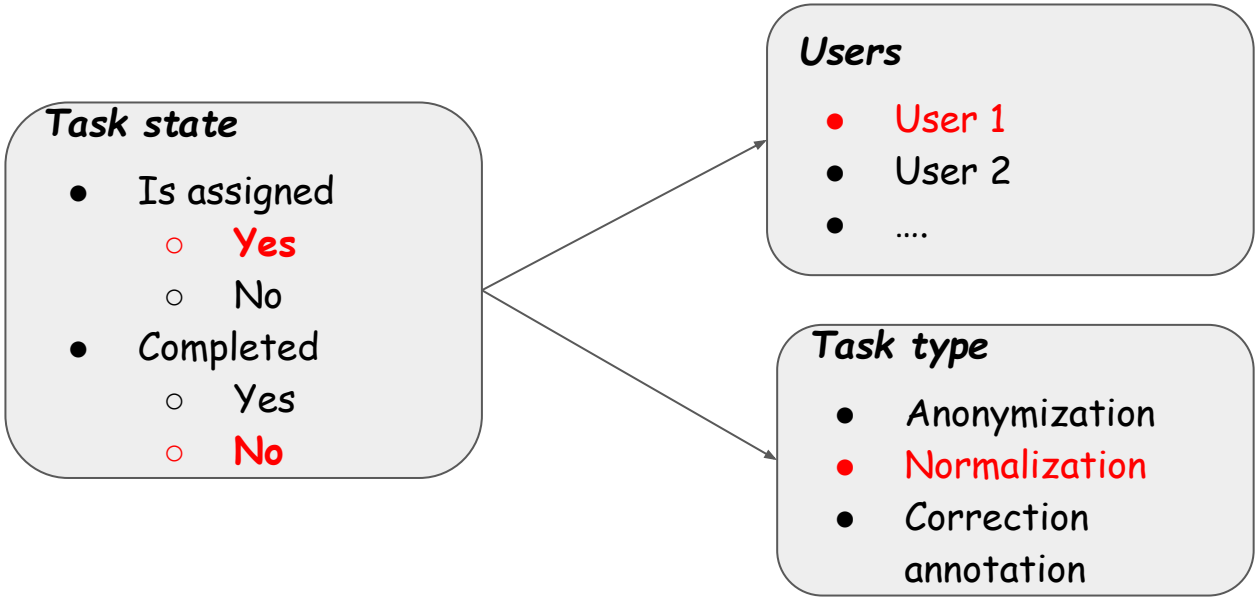
{id:X1XT1, text: raw}



Task manager (step 1)

Essay_id

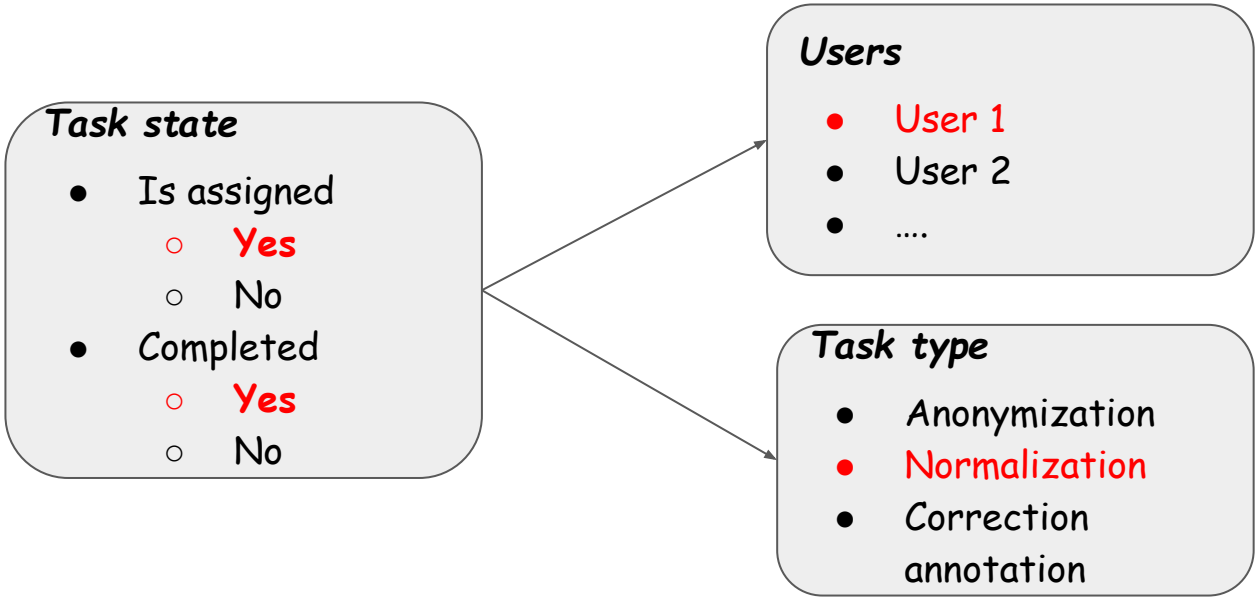
X1XT1



Task manager (step 2)

Essay_id

X1XT1



Ready for Correction annotation

Export dataset

Task type

- Anonymized
- Normalized
- Correction anno

School

- A
- B
- C

L1

- English
- Spanish
- Arabic

Data type

- Source
- Target
- Both

Status

- Complete
- All

File format

- XML
- JSON
- Text

Derivative Datasets

Da • L • A • J (v.1.0)

Dataset for • **L**inguistic • **A**cceptability • **J**udgments for Swedish

- … one error per sentence
- … is a dataset of 9 596 Swedish sentences, equally distributed between correct and incorrect sentences (4 798 each)
- … is based on L2 Swedish essays, manually corrected and labeled
- … correction labels - 4
- … dataset split 80-10-10 (training, validation, testing)

---

freely available from https://spraakbanken.gu.se/en/resources/superlim

Da • L • A • J (v.2.0)

Dataset for • **L**inguistic • **A**cceptability • **J**udgments for Swedish

- … one error per sentence
- … is a dataset of 48 500 Swedish sentences, equally distributed between correct and incorrect sentences (24 250 each)
- … is based on L2 Swedish essays, manually corrected and labeled
- … correction labels - 31
- … dataset split 80-10-10 (training, validation, testing)

original sentence	corrected sentence	error indices (original)	corrected indices	error-corr pair	error label	L1	Approximate level
Förr i tiden kunde vi byta en sak till en annan .	Förr i tiden kunde vi byta en sak mot en annan .	34-37	34-36	till--mot	L-W	Polska	B:Fortsättning
Jag kan ta några exempel av betydelsen av pengar .	Jag kan ta några exempel på betydelsen av pengar .	25-26	25-26	av--på	L-W	Somaliska	B:Fortsättning
Och där efter köper vi nästan allt vi behöver .	Och därefter köper vi nästan allt vi behöver .	4-12	4-11	där efter--därefter	O-Comp	Somaliska	B:Fortsättning
För det andra är det nyckeln av livet .	För det andra är det nyckeln till livet .	29-30	29-32	av--till	L-W	Somaliska	B:Fortsättning
Det är svårt ibland men det kommer inte på engång .	Det är svårt ibland men det kommer inte på en gång .	43-48	43-49	engång--en gång	O-Comp	Somaliska	B:Fortsättning
Bostäder i D-hemland är ett lite hett topik att diskutera .	Bostäder i D-hemland är ett lite hett ämne att diskutera .	38-42	38-41	topik--ämne	L-FL	Ungerska	A:Nybörjare
Bostäder i D-hemland är ett lite hett ämne att diskussera .	Bostäder i D-hemland är ett lite hett ämne att diskutera .	47-56	47-55	diskussera--diskutera	L-FL	Ungerska	A:Nybörjare
Många lägenheter och hus är gamla och har dålig energi-niveå .	Många lägenheter och hus är gamla och har dålig energinivå .	48-59	48-57	energi-niveå--energinivå	O-Comp	Ungerska	A:Nybörjare
Man betalar mycket i vintern , vilket är likadant som i Sverige .	Man betalar mycket på vintern , vilket är likadant som i Sverige .	19-19	19-20	i--på	L-W	Ungerska	A:Nybörjare
Man betalar mycket på vintern , vilket är det samma som i Sverige .	Man betalar mycket på vintern , vilket är likadant som i Sverige .	42-50	42-49	det samma--likadant	L-W	Ungerska	A:Nybörjare

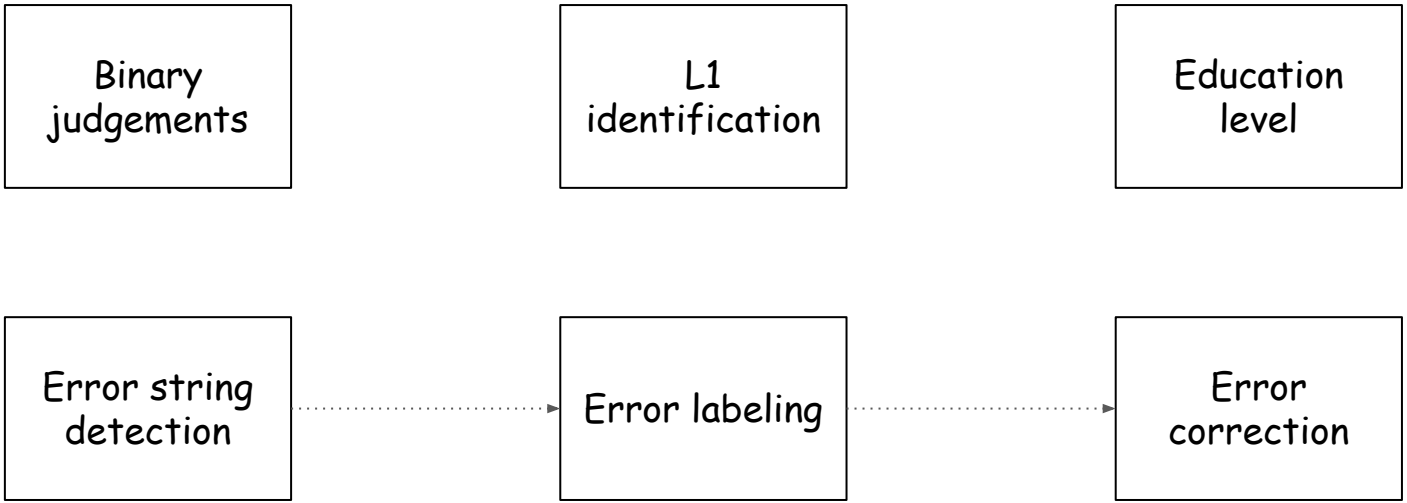
1. original sentences
2. corrected sentences = copies of (1) with all errors corrected

3&4. character indices for error (1) & its correction (2)

5. error-correction pair
6. correction label
7. mother tongue(s)
8. approximate level (A, B or C)

(+ a column marking sentence pairs for training, validation, test as used by us for getting the first baseline)

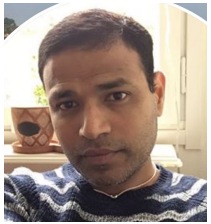
NLP tasks for DaLAJ datasets



Other derivative datasets

- Pseudonymized dataset
  - One pseudonymised error per sentence
- Multiple error per sentence dataset
  - Either one or more error per sentence
  - Around 6600 sentence pair

SweLL Portal - Svala Editor team



Samir
Ali Mohammed



Arild
Matsson



Dan
Rosén



David
Alfter



Ildikó
Pilán



Carl-Johan
Schenström

RIKSBANKENS JUBILEUMSFOND
STIFTELSEN FÖR HUMANISTISK OCH SAMHÄLLSVETENSKAPLIG FORSKNING

GÖTEBORGS UNIVERSITET

SPRÅKBANKENTEXT

SWE-CLARIN

Thank you!

Comments? Questions?