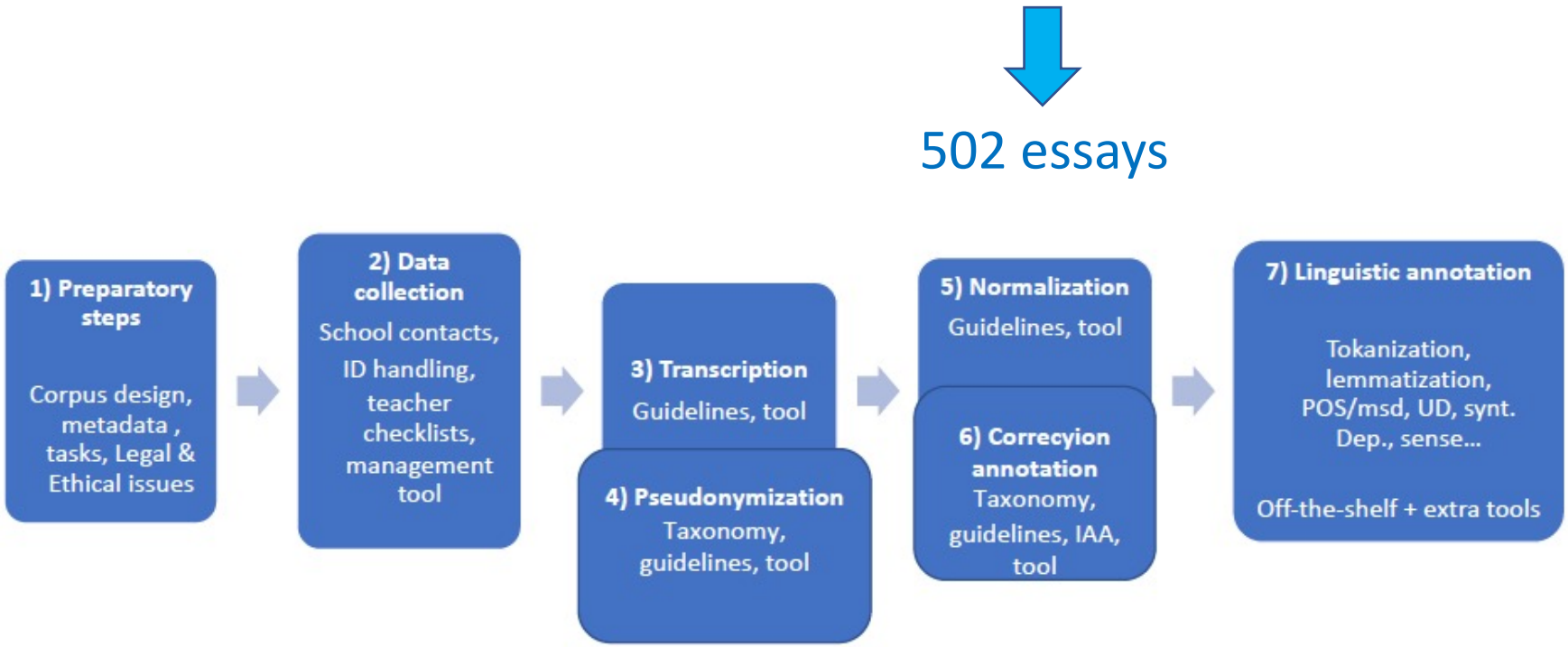
Normalization

https://spraakbanken.github.io/swell-project/Normalization_guidelines

SweLL release August 26-27, 2021

Lisa Rudebeck & Gunlög Sundberg

Dept. of Swedish Language and Multilingualism, Stockholm University



Normalization

Editing of the original learner text in such a way that the normalized version of the text adheres to standard Swedish text norms.

Number of tokens gold corpus
Original version: 147 842
Normalized version: 151 851



Normalization: tool SVALA

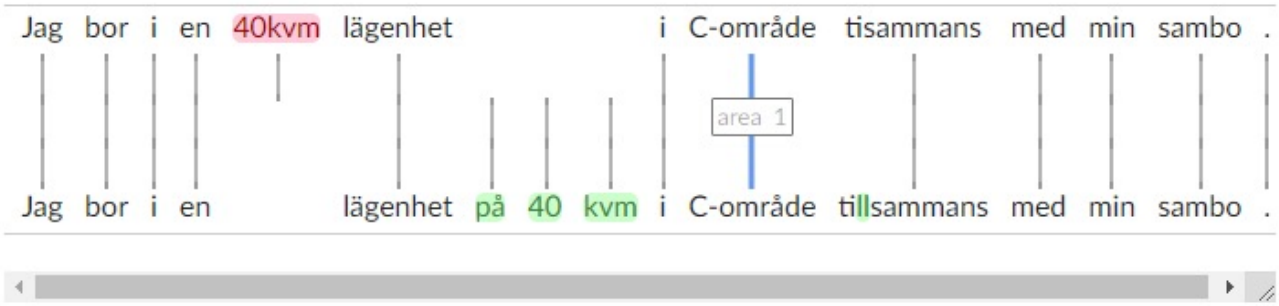
Svala normalization – essay A4AT1

Source text:

Jag bor i en 40kvm lägenhet i C-område tisammans med min sambo . Vi har bott tisammans för två eller tre år nu . Vår legenhet ligger på första våning , so många folk passera och gå upp trapperna . Det finns många affärer och resturanger nära , och B-plats-def B-transport station finns bara 200m promonera . Från mitt fönster jag kan ser C-plats-def-gen busshållplats . Det är bra så man kan springer till nästa bussen . Många bilar körar upp och ner gatan värje kväll , det är inte så bra .

Target text:

Jag bor i en lägenhet på 40 kvm i C-område tillsammans med min sambo . Vi har bott tillsammans i två eller tre år nu . Vår lägenhet ligger på första våningen , så mycket folk passerar och går upp för trapporna . Det finns många affärer och resturanger nära , och B-plats-def B-transport station finns bara 200 m bort att promenera till . Från mitt fönster jag kan se C-plats-def-gen busshållplats . Det är bra så man kan springa till nästa buss . Många bilar kör upp och ner för gatan varje kväll , det är inte så bra .



Normalization: purpose

- Makes the whole text meaningful in its context.
- Makes it possible to use the same normalized text version with different tools, annotation paradigms or categorization taxonomies.
- A necessary methodological step and basis for high accuracy and reliability of automatically generated annotation, or manual annotations, especially for learner corpora.
- Makes the corpus easily searchable with many possible applications, e.g POS.

Normalization : a multiple text level approach

- **Orthography**
  - På **Måndag** är det den 3 **Mars**. → **måndag** 3 **mars**.
  - **24 åringar** → **24-åringar**
- **Morphology**
  - När kommer **den buss** → **bussen**
  - Min mormor har **kommat att bli** jättesjuk → har **blivit**
- **Punctuation**
  - **5.5** procent → **5,5** procent
- **Syntax**
  - I texten **sägs det av henne** att hon efterlyser ….. → I texten **säger hon** att hon efterlyser ….
  - **litteraturverket** → **litterära verket**
- **Lexikon**
  - Internet **uppmanar** vår förståelse → **utmanar**
  - Traditioner ger människorna tid att **stå still** och fundera över livet → … att **stanna upp**
  - Men **har det hänt något problem med mig** → Men **jag har fått ett problem**

Normalization: a target hypothesis (TH)

Hirschmann et al. (2007), Rehbein et al. (2012)

- **Original learner text**:
- *Och det finns olika former av svarta tjänster*
  ***exempelvis av detta är** svartjobb , svart hyra och svartmarknad*

- **TH/Normalization 1**:
- *Och det finns olika former av svarta tjänster,*
  ***exempelvis** svartjobb, svart hyra och svartmarknad*

- **TH/Normalization 2**:
- *Och det finns olika former av svarta tjänster.*
  ***Exempel på detta är** svartjobb, svart hyra och svartmarknad*

Normalization: fundamental values

balancing the following:

- 1) adherence to standard Swedish norms
- 2) fidelity to the original text
  - similarity to the original text string
  - effective communication of the content intended by the writer (according to the normalizer's interpretation)

- *Firsta gång såg jag henne i **gymnastik sckola***

Number of sentences Gold corpus
Original version: 7 807
Normalized version: 8 137

Adherence to standard Swedish norms

- Accepted forms
- *sen/sedan*, *mejl/mail*, *nån/någon*, *dom/de*(*m*) etc.

- Accepted inflectional patterns for specific words
- *partner* (null plural)/*partners*, *kolleger/kollegor, dåligare/sämre*, *givit/gett*, *lyste/lös* etc.

- Acceptance for *satsradning* is fairly high.

Normalization: methodological practices

- Few annotators
- Expertise in linguistic structure, Swedish language norms, and Swedish as a second language research and education.
- Pilot work and discussions with the project group.
- A second opinion from a co-normalizer for particularly difficult decisions.
- Documentation of problematic instances and classifications.

- Context sensitivity
- Understand the text in relation to the task.
- An overview and reading of the whole text gives clues for the interpretation of the text and the use of vocabulary.

X = the text is unintelligable

- *Jag gick till dem och frågade henne om hennes namn och inte berätta något **han** blyg av mig så jag spårade tills jag kände hennes hem **och bad henne och pjäsen idag dig varje morgon jag går att se det och låta dig gå till skolan***

- 137 instances of X, mostly at the Beginner's level

Normalization: final reflections



- *When?*
- Time consuming, but makes the next steps faster and gives high accuracy
- The first normalization step could be a combination of automatic and manual work on orthography.
- 15 % of correction labels are Orthography (spelling mistakes, capitalisation errors etc.)
- *By whom?*
- Second language experts and linguistic norm experts for an up-to-date view on Swedish text norms.
- *How?*
- Basic principals, a systematic approach & tool **Svala**