

Språk-
BANKEN



RIKSBANKENS
JUBILEUMSFOND

STIFTELSEN FÖR HUMANISTISK OCH
SAMHÄLLSVETENSKAPLIG FORSKNING



SWE-CLARIN

SweLL in a nutshell

August 31 2017



GÖTEBORGS
UNIVERSITET



UMEÅ UNIVERSITET



UPPSALA
UNIVERSITET



Stockholms
universitet



RIKSBANKENS
JUBILEUMSFOND

STIFTELSEN FÖR HUMANISTISK OCH
SAMHÄLLSVETENSKAPLIG FORSKNING

SweLL -

Research infrastructure for Swedish as a Second Language

Elena Volodina, Beata Megyesi, Mats Wirén,

Lena Granstedt, Julia Prentice, Monica Reichenberg,
Gunlög Sundberg

Key terminology

SweLL **Swedish Learner Language**

L2 **Second (and foreign) language**

What is infrastructure?



"'Infrastructure'? — You mean like rocks and sticks?"

An electronic research infrastructure

- (free accessible) data in electronic format
- technical platform for exploring data, including tools and algorithms for data analysis, and visualization
- a set of tools and technical solutions for new data collection and preparation, including data processing and annotation
- a network of experts in the relevant disciplines, incl. legal and ethical questions



How can we help?

- Collect and annotate data (L2 essays, error logs, ...)
- Develop tools for analyzing and visualizing L2 data (e.g essays, in Strix / Korp)
- Gain expert knowledge
 - to promote research on L2 Swedish
 - to support course book writers, L2 teachers, L2 assessors, L2 students
 - to support instruction of future L2 teachers
 - ... and to promote app development for
 - L2 Swedish learning



As you know...



© Enache Dumitru Bogdan • www.free-cartoon-clipart.blogspot.com

- Riksbankens Jubileumsfond, infrastructure project IN16-0464:1

7 mln SEK



shutterstock

- 2017-2019

Partners

- University of Gothenburg: NLP, L2, assessment
- Stockholm university: NLP, L2
- Uppsala university: NLP
- Umeå university: L2/assessment





Presentations

SweLL participants

✓ Göteborgs universitet (svenska, pedagogen)



Julia
Prentice



Monica
Reichenberg



Elena
Volodina



Presentations

SweLL participants

✓ Stockholms universitet (svenska, lingvistik)



Gunlög
Sundberg



Mats
Wirén



UMEÅ UNIVERSITET

Presentations

SweLL participants



Umeå universitet (språkstudier)



Lena
Granstedt



UPPSALA
UNIVERSITET

Presentations

SweLL participants



Uppsala universitet (lingvistik)



Beáta
Megyesi



Presentations

developers: Sparv, Mink, Lärka



Markus
Forsberg



David
Alfter



Ildikó
Pilán



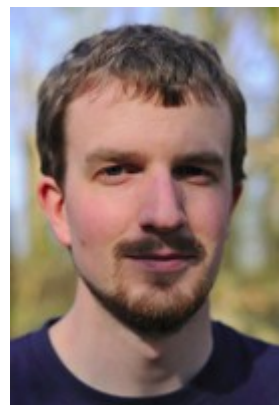
Carl-Johan
Schenström



Dan
Rosén



Anne
Schumacher Hammarstedt



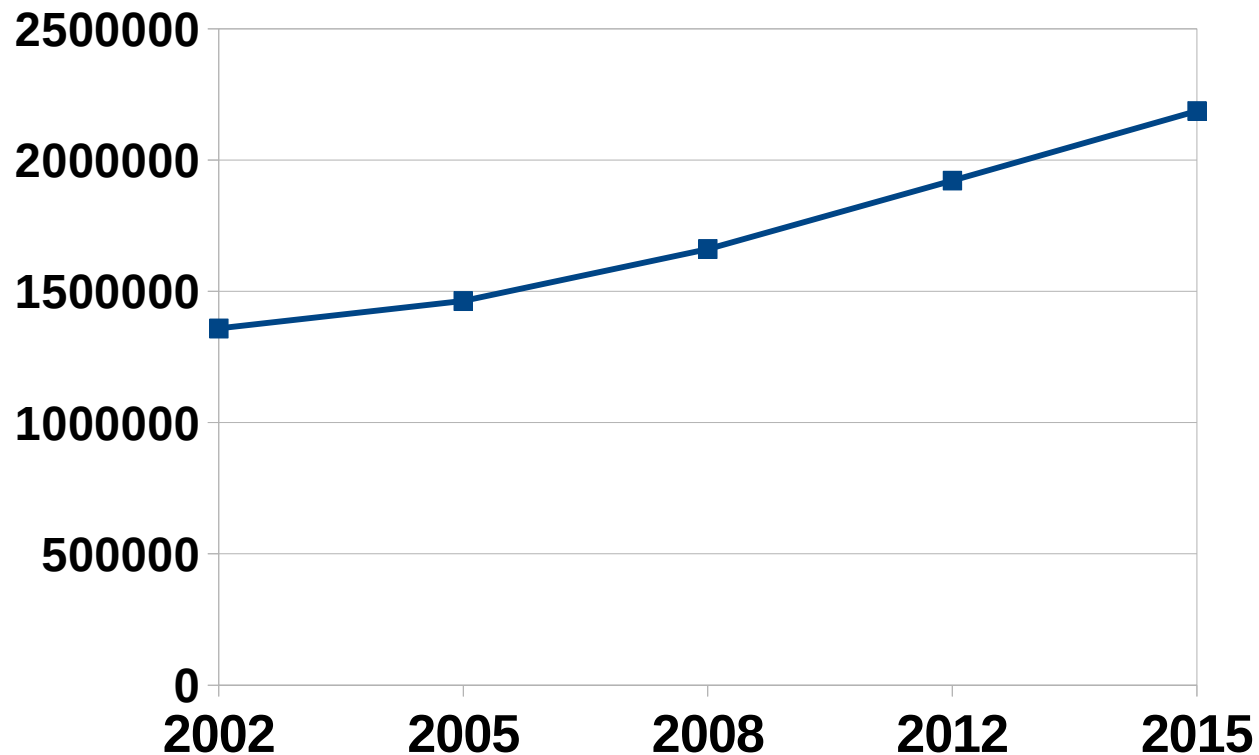
Martin



Jonas
Lindh

Societal need

Citizens with foreign background, 2002-2015



2015: out of **9,9 mln** citizens, **2,2 mln** have foreign background. i.e. **22,2 %**
(Statistiska centralbyrån)

Focus on literacy

- Dutch study:

→ Average reading comprehension ~B1 level

Velleman, E., van der Geest, T.: Online test tool to determine the CEFR reading comprehension level of text. Procedia Computer Science 27 (2014)



Literacy: Sweden

- PIAAC study with focus on literacy (2013)
- Sweden among 5 “**best**” out of 23 countries (on average)
- **Biggest** discrepancy between citizens with native and non-native origins
 - low employability
 - bigger risk for bad health

OECD. 2013. OECD Skills Outlook 2013. First Results from the Survey of Adult Skills.

PIAAC. 2013. Survey of Adult Skills (PIAAC).

SCB. 2013. Tema utbildning, rapport 2013:2, Den internationella undersökningen av vuxnas färdigheter. Statistiska centralbyrån.

Our focus is on...

- MAIN: L2 essays (writing)
- SIDE: exercise logs (reading and listening comprehension, vocabulary and grammar training)
- NO speech data – yet
- target group: adult learners

Problem 1: lack of L2 data

- Electronic L2 production is very difficult to collect
 - NOT available online
 - Need learner permits
 - Need learner variables (gender, age, native language, etc)
 - Sensitive in nature
 - Need to be carefully agreed with the existing laws
- We need an infrastructure/environment for storing and collecting L2 data

Problem 2: lack of coordination

- There is a national need to coordinate various (individual and bigger-scale) efforts aimed at collecting L2 production (e.g. which permits, learner variables, formats etc so that the data could be comparable and usable between projects, and REUSABLE between projects)
- There is a need to digitize and process hand-written L2 language samples (e.g. National tests in Swedish and L2 Swedish) in an organized nation-wide effort

Problem 3: lack of L2 tools and models

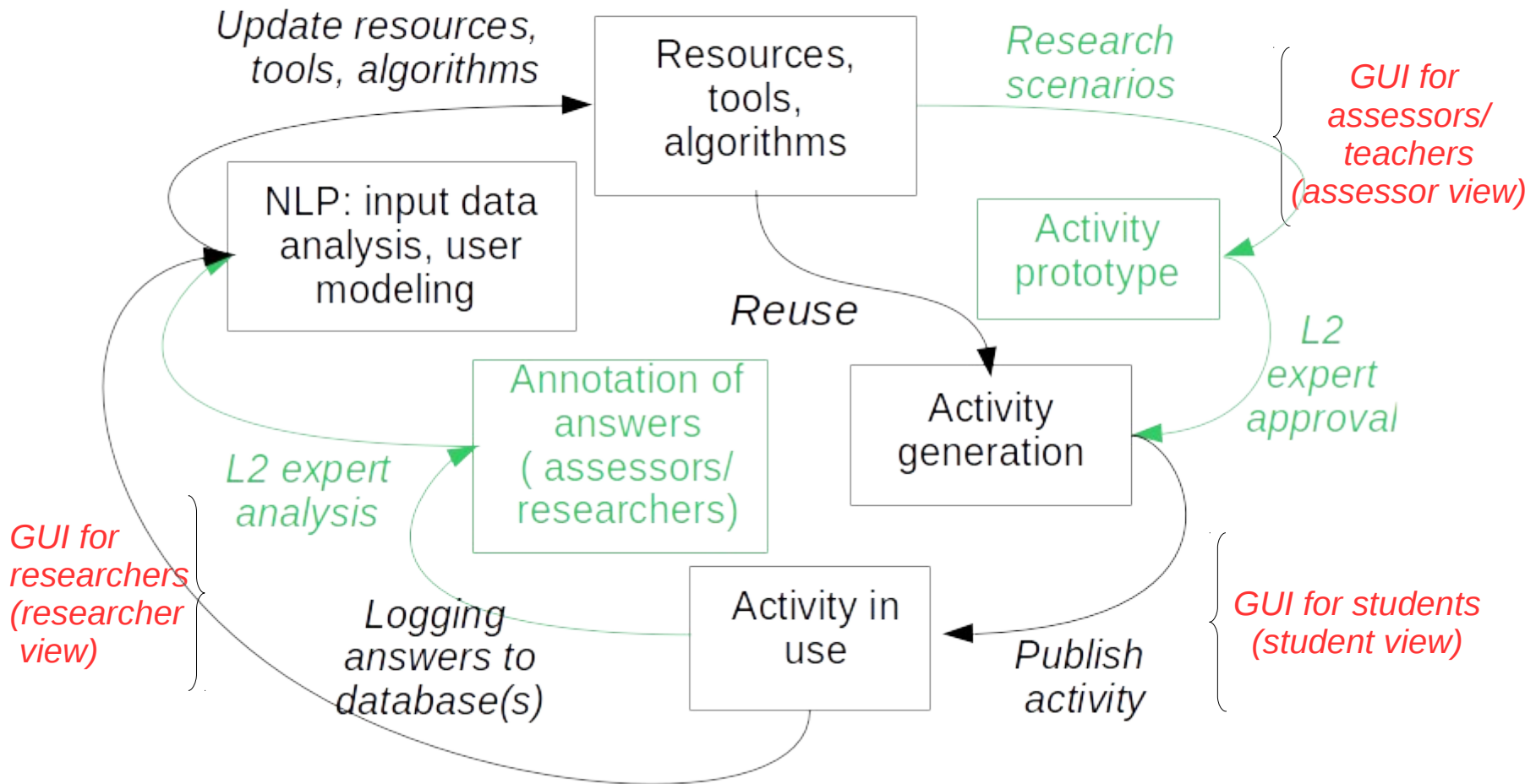
- Existing NLP tools are not capable to analyze L2 learner language due to numerous infelicities (normative language analysis versus error analysis)
 - Adaptation of existing NLP tools required
 - Adaptation of tools targeting "deviating" forms of language: historical texts or social media
- Development of new tools require specific, often hand-annotated data
 - Error-tagged corpora
 - Learner profiles (grammar, vocabulary, etc. per level of proficiency)
- ...

Initial steps and pilot studies

- **Data** collection and digitiation
 - SweLL corpus (core corpus)
 - The Uppsala Corpus of Student Writings (monitor corpus)
 - “Alternative” L2 data
- **Resource creation** (e.g. SweLLex – L2 productive vocabulary)
- **Algorithm development:** L2 error detection & normalization
- **User-oriented tools:**
 - L2 annotation tools (SweGRAM)
 - L2 essay analysis (Lärka-based online tool)

The ultimate goal

L2 infrastructure activity development cycle





Data collection



Relevant laws & regulations

the core ones

- Personal Privacy Act, PPA (*Personuppgiftslagen*)
- Ethical Review Act (*Etikprövningslagen*)
- *Tryckfrihetsförordningen*
- Copyright (*Upphovsrättslagen*)

Different priority, and pretty conflicting

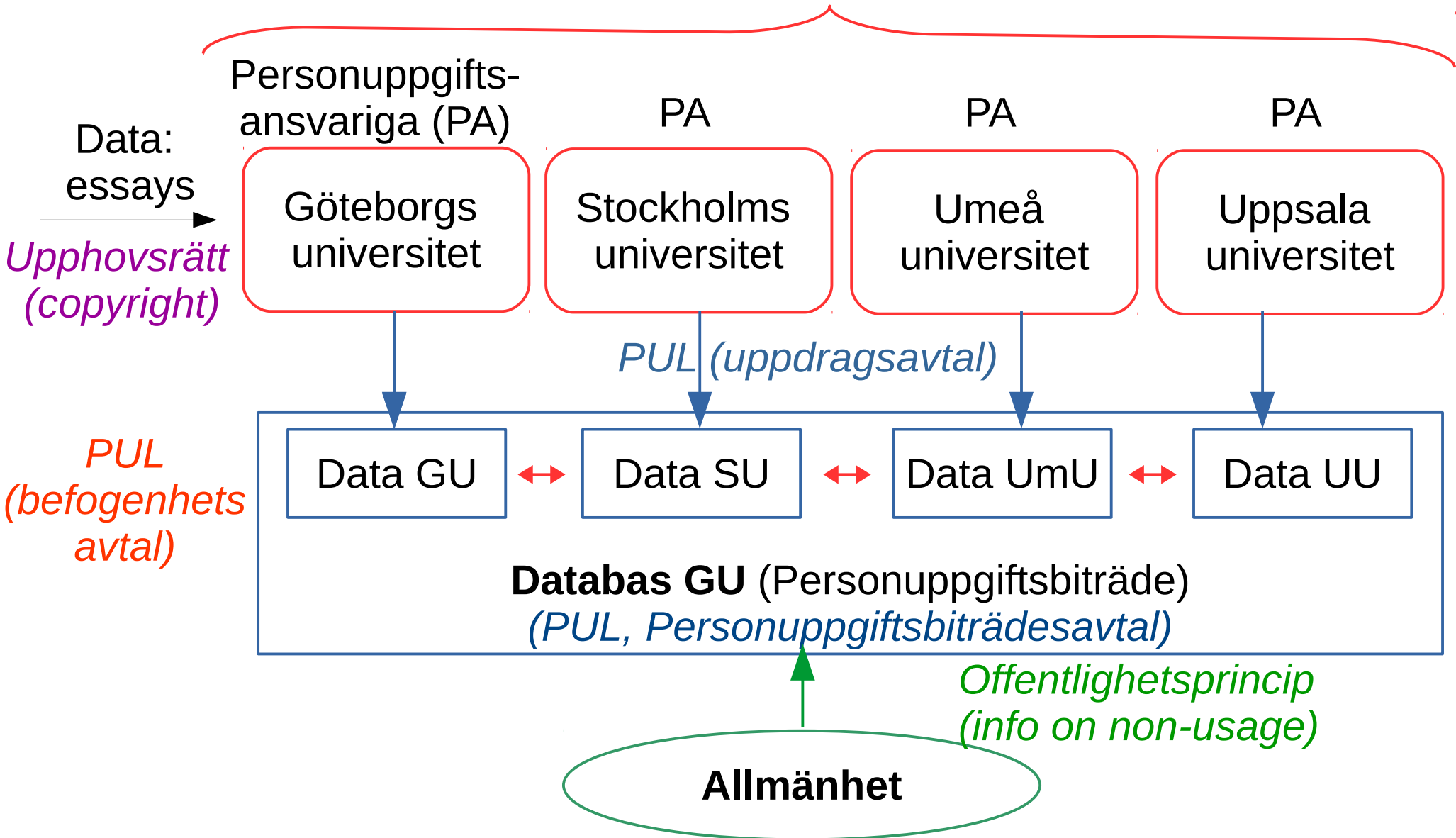
What is “personal information”?

- **Personal information:** Any kind of information that directly or indirectly can be linked to a physical person (alive one).
 - Agreements (signed)
 - Hand-written versions *versus* ip-addresses
 - Demographic metadata: age, gender, L1, residence time in Sweden (month/year of arrival), education level, other languages
 - Essay metadata: date, school, class, teacher, topic, genre, grade, level, additional information/handouts
- If possible to identify a person
 - applications to PUL, Ethical Review Act / Board
 - sketch all scenarios where the data can be used

More “personal information” terms

- **Personuppgiftsansvarig:** Den som ensam eller tillsammans med andra bestämmer ändamålen med och medlen för behandlingen
- **Personuppgiftsbiträde:** Den som behandlar personuppgifter för den personuppgiftsansvariges räkning.
- **Tredje man:** Någon annan än den registrerade, den personuppgiftsansvarige, personuppgiftsombudet, personuppgiftsbiträdet och sådana personer som under den personuppgiftsansvariges eller personuppgiftsbiträdets direkta ansvar har befogenhet att behandla personuppgifter.

PUL, Ethical Review Act (collaboration agreement)



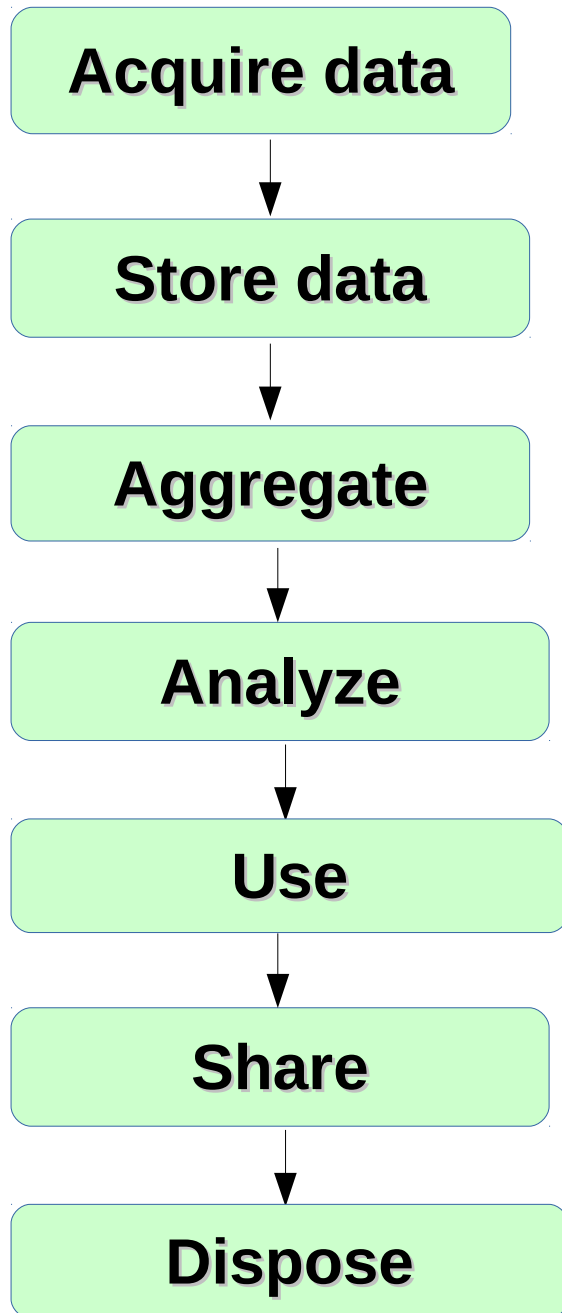
Documents we need

the ones I know

- *Samtycke* (permit)
- *PUL application* (to GU lawyer)
- *Application to Rthical Review Board*
- *Samverkansavtal* (between project partners)
- *Personuppgiftsbirädesavtal* (all partners with GU)
- *Uppdragsavtal* (all partners with GU)
- *Befogenhetsavtal* (between all partners)
- *Information to the public* (in case some researchers would want to get access to the data)

non-PUL approach

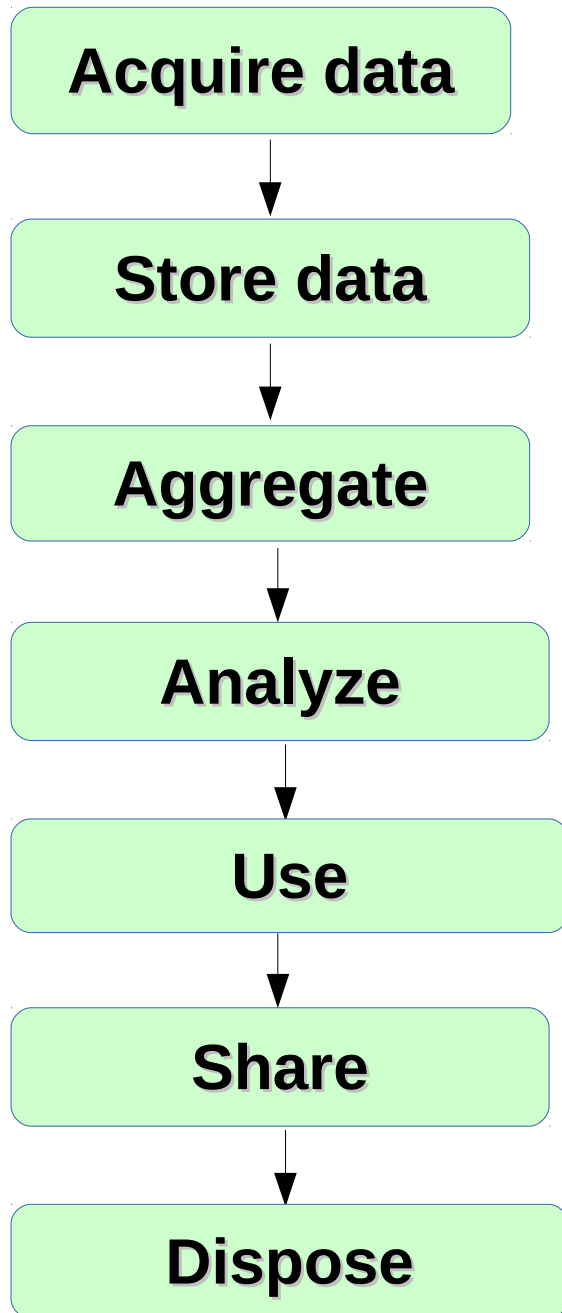
<https://goo.gl/7LFoTC>



Model from *Building digital trust: The role of data ethics in the digital age*

non-PUL approach

<https://goo.gl/7LFoTC>



Implications

- for kind of applications (PUL, Ethical Review)
- for portal development
- for the overall data handling flow
- who can have access to data
- for data usage scenarios

SweLL corpus (2013-2016)

core data

SweLL workflow



Learner variables
Collected through permits*

- Student variables:**
- Age/birthyear
 - Gender
 - Mother tongue(s)
 - Residence time in Sweden
 - Education level

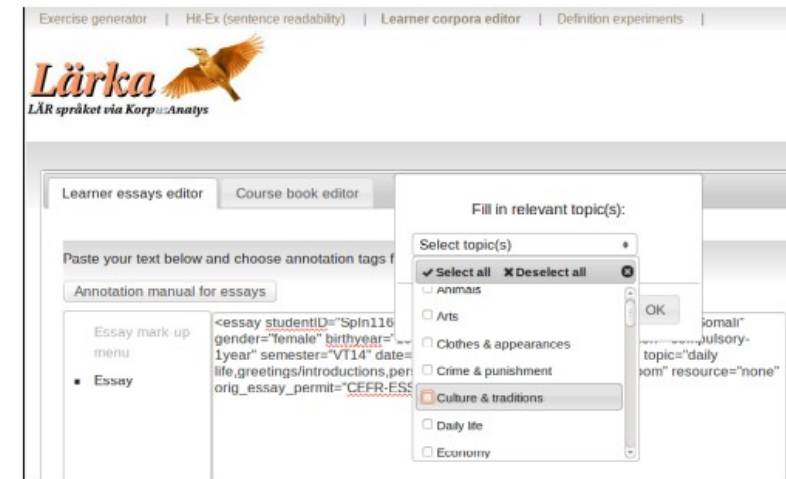
- Essay variables**
- Assigned CEFR level
 - Essay setting (exam/home)
 - Use of extra materials
 - Academic term and date
 - (Title, topic, genre, grade)

Assessors
Minimum of two trained assessors

- Inter-annotator agreement**
- A degree to which several annotators agree about assigning attributes
 - Reported for SW1203 subcorpus
 - Krippendorff's alpha for pairwise agreement = 0.80
 - **0.80 = good annotation quality** (Artstein & Poesio 2008)

SweLL digitization principles

- 1. Do not reveal author identity**
 - * revealing names → replace with *NN*
 - * addresses → replace with *NN-street*
- 2. Do not correct errors**
 - * if several interpretations possible → make *positive assumption*, i.e. that the learner made no mistake
- 3. Preserve illegible handwriting**
 - * each illegible letter → replace with @
 - * stricken text → leave out



* http://spraakbanken.gu.se/sites/spraakbanken.gu.se/files/tilstand_eng-24042013_v03.pdf

SweLL corpus

core data

Sub- corpus	A1	A2	B1	B2	C1	Un- known	Total
Tisus	-	-	-	27	78	-	105
Sw1203	-	-	33	45	11	1	90
SpIn	16	83	42	2	-	1	144
Total	16	83	75	74	89	2	339

The Uppsala Corpus of Student Writings

reference corpus

Level	Age	School level and curriculum	Number of essays	Number of tokens	Tokens per essay
C-3	9	Compulsory, Lpf94 + Lgy11	91	8,644	95
C-5	11	Compulsory, Lpf94	66	13,121	199
C-6	12	Compulsory, Lgr11	47	17,741	377
C-9	15	Compulsory, Lgr94 + Lgr11	249	137,689	553
US-1	16	Upper Secondary, Lgy11	131	76,521	584
US-3	18	Upper Secondary, Lgy11	410	347,836	848
GY-3	18	Upper Secondary, Lpf94	1,506	1,055,468	701
Total			2,500	1,657,020	663

Table 1: Distribution of the subset of texts by school year, given as number of texts, sentences and tokens, and average number of tokens per essay used in the pilot study.

Handwritten essays	Printed essays
Transcription	Scanning-conversion-editing
Coding	Coding
Proofreading and final editing	Proofreading and final editing

Table 2: Preparation of the essays.

SweLL digitization process: flow, guidelines, questions

Digitization in a nutshell

SweLL digitization principles

1. Do not reveal author identity

- * revealing names → replace with *NN*
- * addresses → replace with *NN-street*

2. Do not correct errors

- * if several interpretations possible → make *positive assumption*, i.e. that the learner made no mistake

3. Preserve illegible handwriting

- * each illegible letter → replace with @
- * stricken text → leave out

FACTS

Level: A1

Time: 15 min

Length:

117 tokens

Topic:

Presentation/

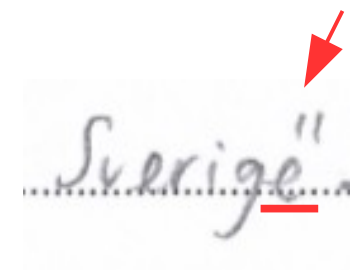
Om mig

Jag bor i Källred. Din familj gästar. Jag har
nio syskon fem bröder och fyra syster. Min bror
heter Amas, Ahmed, Abed Alkame, Abed Almalik, Bilal.
Min syster heter Nada, Sahar, Sofyan, Fatema.
Min pappa heter Mohammed. Han kallar Faxe.
Min mamma heter Sheema. Hon lärare.
Jag är ditt språk arabiska. Dina kompisar
heter bra. De heter Abazaher och Abo Loze.
Jag tycker om läsa på boken och lyssna på Korian Kane
Jag tycker inte om vinter. Jag älskar vår.
Mamma och Pappa ligger i Syrien. Syrien krig
mycket. Jag är efter skola gå trena på Jem.

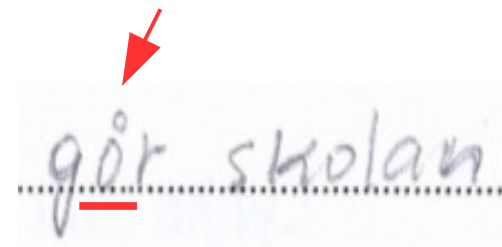
“New” letters

creative student writing

Case 1: If there is an existing letter, use it, e.g. Sverig**ë**



Case 2: If there is no such letter, find the closest graphical correspondence, e.g. **o** or **å**; and apply the rule of positive assumption, e.g. choosing **å** - **går**



Deciphering letters

creative student writing

Det var i Augusti. Jag minns dagen var regnig. Jag kände mig lite kallt för att jag kom från varmt land. Jag väntade till min bagage och tittade på en familj som stannade bredvid mig. De pratade och skrattade varandra. När jag tittade på de jag tänkte på min mamma för att saknade henne mycket. Men jag skulle träffa henne efter några minuter. Det kände nämligen för att träffa min mamma efter lång tid. Min bagage kom och gick ut snappa. Min mamma var stannade och tittade på allt sidan. Hon kände inte säkert för jag kom lite för sent. Hon

Some statistics (small experiment)

	A1	A2	B1	B2	C1
Nr essays	3	3	2		
Nr char	1376	2106	5283		
Nr words	256	402	993		
Time (min, sum)	31	27	73		
Average (min /essay)	10	9	24,5		
Average (char/essay)	459	702	1761		
Av. words/essay	85	134	331		
Av words/min	8,25	14,9	13,6		

SweLL error annotation: considerations

Other error-annotated corpora

closely related languages

- ASK – L2 Norwegian
- Merlin – L2 German, Italian, Czech
- FALKO – L2 German
- ICLE – L2 advanced English
- Cambridge Learner English – L2 English
- COPLES – L2 Portuguese

ASK (L2 Norwegian)

http://hnk.ffzg.hr/bibl/lrec2006/pdf/573_pdf.pdf

- Lexical codes (8)
- Morphological codes (3)
- Syntactic codes (7)
- Punctuation codes (4)
- Unidentified errors (1)

Total: 23 error types

Merlin (L2 German/Czech/Italian)

http://merlin-platform.eu/docs/MERLIN_user-manual-EN.pdf

- G_Grammar [TH1] (21)
- O_Orthography [TH1] (8)
- G_Intelligibility (8)
- V_Vocabulary (10)
- C_Coherence/Cohesion (4)
- S_Sociolinguistic appropriateness (10)
- P_Pragmatics (3)

Total: 64 error types

Approach to error-annotation

essay: <https://goo.gl/bXv3kj>

1. En dag Jag var sjuk och dag gick till sjuhus träffade läkare när dag träffade läkare Jag kände rädd så mycket.
2. Det var Min att går till sjuhsen och Jag väntrummet läkare och läkare ni och Han frågade till Mig och Han tag Medicin till mig och han försöket och Jag tillbakade min hemma och Jag sova Jag var trett sen Jag vaknade Jag åt Mat och Medicin.
3. sen min mamma ring till mig frågade mig Hur var det sen Jag pratade slut och sova förta Jag kom till sverige alla männingko.
4. tittade till mig Jag vet inte också att varför tittad mig.
5. och Jag var hemma Jag var rädd Jag vågade Jag kunde inte ut ur hemma.
6. 1-2 måna Jag kunde gick ut ur hemma.
7. Jag träffade alla männingko mycket Jag träffade kompis och Jag var ollis och Jag åkebe min hemma.
8. sen Jag var ont i huvud och Jag åt Medicin och det inte stutade.
9. i Mårgon Jag gick sjukhus Jag fick Medicin den var en farti@, att Jag inte åkede tillsjukhuset.

Approach to error-annotation

uppsatsen: <https://goo.gl/bXv3kj>

2. Det var Min att går till sjuhsen och Jag väntrummet läkare och läkare ni och Han fårgade till Mig och Han tag Medicin till mig och han försöket och Jag tillbakade min hemma och Jag sova Jag var trett sen Jag vaknade Jag åt Mat och Medicin.

“ . “

error with capitalization or not?

Approach to error-annotation

uppsatsen: <https://goo.gl/bXv3kj>

2. *Det var Min att går till sjuhsen och Jag väntrummet läkare och läkare ni och Han frågade till Mig och Han tag Medicin till mig och han försöket och Jag tillbakade min hemma och Jag sova Jag var trett sen Jag vaknade Jag åt Mat och Medicin.*

“ . “

error with capitalization or not?

Shoud we have “consequence of correction” /
“följdnkorrigerering” error-grupp?

What is it we are error-annotating?

Approach to error-annotation

uppsatsen: <https://goo.gl/bXv3kj>

3. sen min mamma ring till mig fårkade mig Hur var det sen Jag pratade slut och sova förta Jag kom till sverige alla männingko.
4. tittade till mig Jag vet inte också att varför tittad mig.

Intelligencebility?

Approach to error-annotation

uppsatsen: <https://goo.gl/bXv3kj>

6. 1-2 måna Jag kunde gick ut ur hemma.

7. Jag träffade alla männingko mycket Jag träffade kompis och Jag var ollis och Jag åkebe min hemma.

8. sen Jag var ont i huvud och Jag åt Medicin och det inte stutade.

9. i Mårgon Jag gick sjukhus Jag fick Medicin den var en farti@, att Jag inte åkede tillsjukhuset.

Any chance to make THIS correct?

Question 1

uppsatsen: <https://goo.gl/bXv3kj>

1. En dag Jag var sjuk och dag gick till sjuhus träffade läkare när dag träffade läkare Jag kände rädd så mycket.
2. Det var Min att går till sjuhsen och Jag väntrummet läkare och läkare ni och Han frågade till Mig och Han tag Medicin till mig och han försöket och Jag tillbadade min hemma och Jag sova Jag var trett sen Jag vaknade Jag åt Mat och Medicin.
3. sen min mamma ringt till mig färdade mig Hur var det sen Jag pratade slut och sova förta Jag kom till sverige alla männingko.
4. tittade till mig Jag vet inte också att varför tittad mig.
5. och Jag var hemma Jag var rädd Jag vågade Jag kunde inte ut ur hemma.
6. 1-2 måna Jag kunde gick ut ur hemma.
7. Jag träffade alla männingko mycket Jag träffade kompis och Jag var ollis och Jag åkebe min hemma.
8. sen Jag var ont i huvud och Jag åt Medicin och det inte stutade.
9. i Mårgon Jag gick sjukhus Jag fick Medicin den var en farti@, att Jag inte åkede tillsjukhuset.

Is there any reason to error-annotate such an essay?

Question 2

uppsatsen: <https://goo.gl/bXv3kj>

1. En dag Jag var sjuk och dag gick till sjuhus träffade läkare när dag träffade läkare Jag kände rädd så mycket.
2. Det var Min att går till sjuhsen och Jag väntrummet läkare och läkare ni och Han frågade till Mig och Han tag Medicin till mig och han försöket och Jag tillnade min hemma och Jag sova Jag var trett sen Jag vaknade Jag åt Mat och Medicin.
3. sen min mammaring till mig förkade mig Hur var det sen Jag pratade slut och sova förta Jag kom till sverige alla männingko.
4. tittade till mig Jag vet inte också att varför tittad mig.
5. och Jag var hemma Jag var rädd Jag vågade Jag kunde inte ut ur hemma.
6. 1-2 måna Jag kunde gick ut ur hemma.
7. Jag träffade alla männingko mycket Jag träffade kompis och Jag var ollis och Jag åkebe min hemma.
8. sen Jag var ont i huvud och Jag åt Medicin och det inte stutade.
9. i Mårgon Jag gick sjukhus Jag fick Medicin den var en farti@, att Jag inte åkede tillsjukhuset.

What can L2 research learn from this?

ASK

what is lacking

1. foreign morphology (?): *som jag springe**d***
2. foreign phrase-building (?): *De väder... var inte bra*
3. contextual agreement: *...(min **bror**). **Hon**...*
4. non-idiomatic use: *...vi kan ser vilket är **inte bättre än annat***
→ mindre bra? *...hon är meningslös **i världen***
5. preposition errors: *...**i** alla platser* (“wrong word” is too broad)
6. detailed agreement errors (morphology group only?), e.g.
 - tense: *psykiska besvär **har fördubbla** sedan 1996*
 - def-indef / over-definiteness: *min **bussen**, **den** huvud orsaken, **de** väder*
 - noun-adj: *ing**en** problem*
 - number: ***de** kan bli **stressad***

ASK

uncertain

Jag tillbakade min hemma

- *tillbakade* – wrong word? wrong inflection?
wrong part of speech?

Principles of error annotation

http://merlin-platform.eu/docs/MERLIN_user-manual-EN.pdf

1. Add a correctly re-written hypothesis

To guarantee transparency, coherence, and reliability of annotations, it is a good idea first explicitly write a 'target hypothesis' (TH), i.e. a corrected reconstruction of the learner text that a subsequent error annotation can build upon (Reznicek/Lüdeling et al. 2012).

Principles of error annotation

http://merlin-platform.eu/docs/MERLIN_user-manual-EN.pdf

2. Change the original text as little as possible ...

...to create a grammatically and orthographically correct version of the original learner text ('minimal' TH → TH1)

Principles of error annotation

http://merlin-platform.eu/docs/MERLIN_user-manual-EN.pdf

3. Change further the original text to create an acceptable version ('extended' TH → TH2)...

take into account vocabulary, context, socio-linguistic etc. aspects

Annotation tool

normalization, error annotation

<http://demo.spraakdata.gu.se/dan/swell/>



Non-error taxonomy

Something new? Should we?

- competence assessment depends not only on errors, but also successes (acc to CEFR document, Chapter 5, 9). “Can-do” statements are positively expressed!

https://www.coe.int/t/dg4/linguistic/Source/Framework_EN.pdf

- progress is measured not only through errors, mostly through development of new “structures” in the target language (see Processability Theory and its stages)

goo.gl/SjeGmd

- CAF-principles: Complexity, Accuracy, Fluency
- Second Language Acquisition research (based on an extensive review of research literature)
- learner texts (based on an inductive analysis of 10% of all learner texts)

Next steps for the core corpus

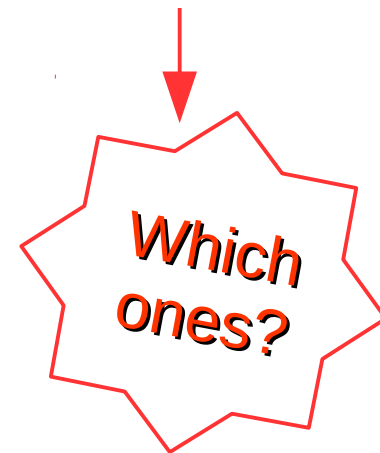
- Preparatory work
 - Legas issues → Ethics by design?
 - Metadata decisions
 - Corpus constitution (L1, topics, levels, courses)
 - Error/deviation taxonomy
 - (“Can-do” taxonomy?)
- Collection
 - School contacts, agreement contracts (translated into the L1s)
 - Collection portal
 - (De-personalized) flow of collection
- Tools for
 - Annotators: transcription, normalization, anonymization, error annotation, visualization and statistics
 - Researchers: searches, statistics, ...
 - Learners: error detection, exercise generation, ...
- XXX

Alternative L2 data

Lark Trills for Language Drills

Text-to-speech technology for language learners

- Dictation and spelling exercise
- Focus on
 - evaluation of the quality of TTS
 - finding ways to give feedback on **spelling errors**

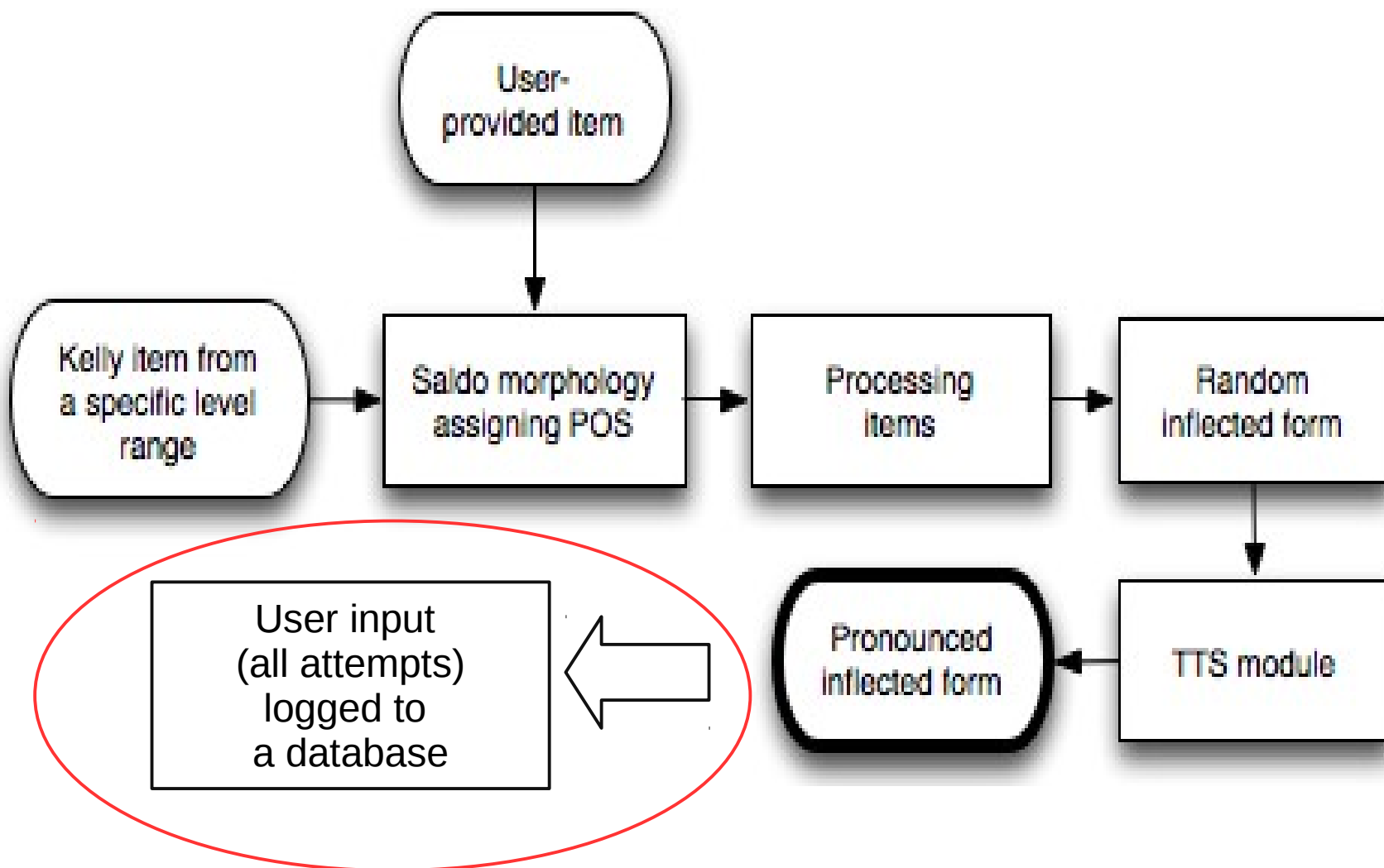




Pipeline



for word & (inflected word) levels



SPEED

SPELLing Error Database

- For each correct item (base form + word class) we store:
 - session ID (no personal data, such as L1)
 - incorrect spelling(s)

L2 spelling error database, SPEED

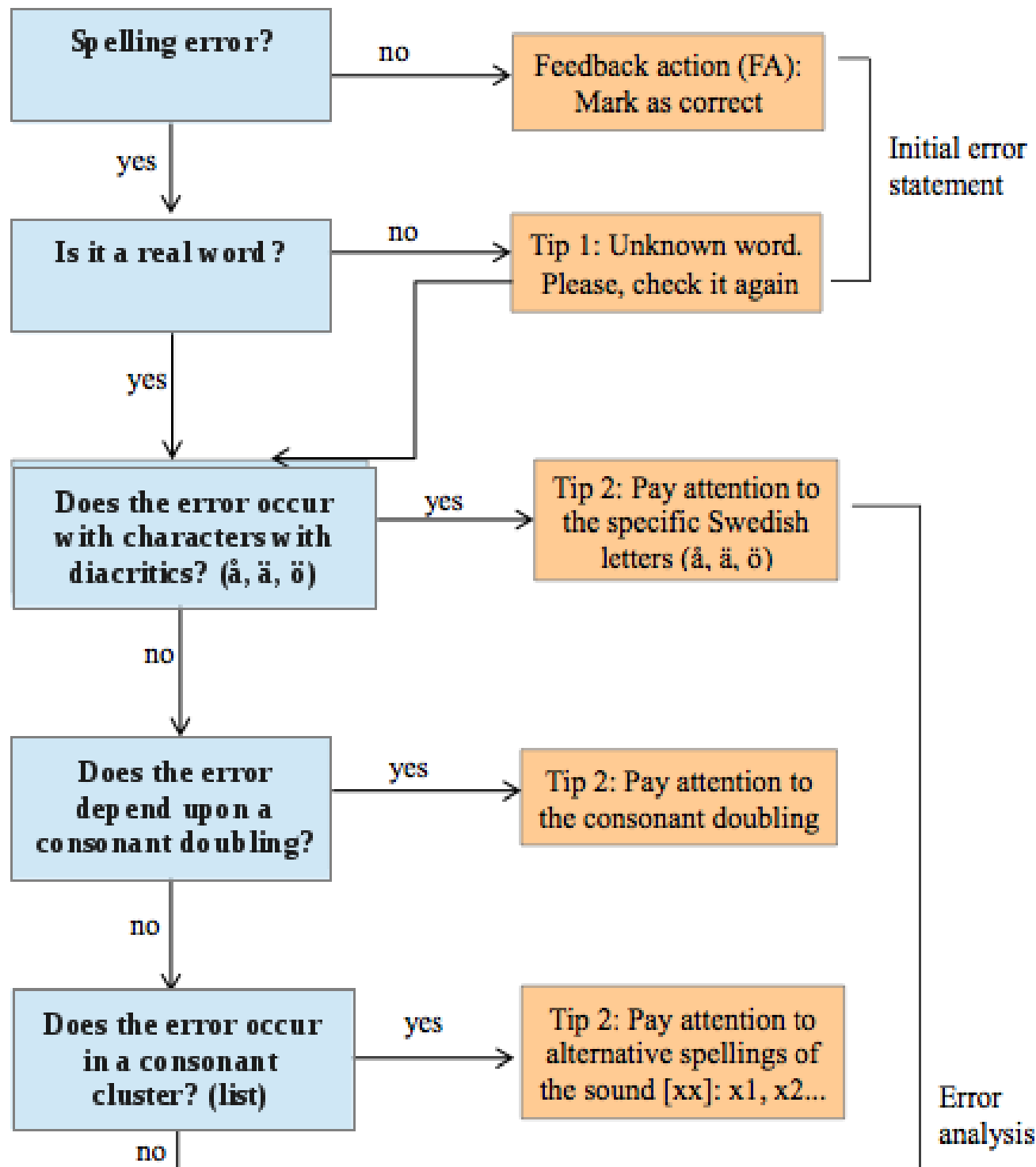
```
- <LexicalEntry uid="LexicalEntry-58d3459f-5acb-43f8-b60e-deb45a986c56">  
  <Sense id="speed--kelly-6950" uid="Sense-b1d45016-bdb5-4584-9ec5-11f780ecbf8a"/>  
  <word lang="swe" pos="AV" uid="word-4d9ed4cb-83b7-4293-a786-ff11f398e2d2">förträfflig</word>  
  <misspelling sessionID="2013-05-13-22-27-28" time="22:58:25" uid="misspelling-3ba31d83-f1ad-4c99-bc33-2c6a3a3c7849">förtrevlig</misspelling>  
- <modification uid="modification-4673c2b2-63a1-4c3c-b2a5-c6a80bc5dd20">  
  <feat att="updatedBy" val="laerka" uid="feat-ec50eb25-d870-4f53-b86c-ed620e3a332c"/>  
  <feat att="modificationDateTime" val="2013-05-13T22:58:26.01+02:00" uid="feat-4032acab-afa3-42c3-b6b0-3f904e345b76"/>  
  <feat att="modificationAccepted" val="pending" uid="feat-2e43c1b4-1106-4364-8df6-3991a4da6578"/>  
  <feat att="modificationComment" val="" uid="feat-22d76dd2-c26f-4ccd-b7a9-e6997082e0cd"/>  
</modification>  
</LexicalEntry>
```

Correct

Logged misspelling

Error data

Error types	Nr,%	Example (correct → *incorrect*)
Competence-based errors	55	
Consonant doubling	28	stoppa → *stopa*
Diacritics (å, ä, ö)	23	högre → *hogre*
Phonetic errors (e.g. voiced vs voiceless)	25	relevans --> *relevanz*
Consonant clusters (phoneme-grapheme mappings, incl. cases of homonyms)	20	skön → *sjön*
Other (unclassifiable)	4	Israel → *visträv*
Performance-based errors	17	
Typos (neighbouring keys, addition, deletion, insertion, replacement)	17	förbättra → *förb'ttra*
Across one word (phrases & sentences)	28	se en bild → *sen bild*



SPEED

SPELLing Error Database

Advantages of collecting a corpus
by applying this method:
participants are quickly attracted,
while cost, time and effort of
collecting a corpus are reduced

THIS is RESEARCH DATA!

And we need more of it!

SweLL focus is on...

- *Main:* L2 essays (writing)
- *Secondary:* **exercise logs (reading and listening comprehension, vocabulary and grammar training)**

L2 “alternative” data

- Logs – acc. to a defined research interest
- Steps:
 - Implement an activity for learners
 - Prepare database for storing (structured) data
 - Implement a way to browse logs, visualize statistics etc
 - If necessary – add extra annotation steps (manual, automatic)

Pilot 1 on L2 “alternative” data

- Identifying most predictive features for a language proficiency level (for diagnostic purposes)
 - Multi-word expressions
 - Syntactic properties (e.g. word order)
 - Knowledge of word morphology (e.g. inflections)



David Alfter

Pilot 1 on L2 “alternative” data

A1	A2	B1	B2	C1
translation	gaps	gaps	matching	gaps
matching	matching	category substitution	free answers	wordbank
gaps	free answers	text questions	gaps	matching

Table 1: Top exercise types by level



David Alfter

L2 “alternative” data (logs)



LÄR språket via **Korpus** Analys

Exercise type evaluation

Bundled gaps (variant 1)

Which word fits into these gaps? Each gap contains the same word. Write the word.

Hennes _____ var på hans lår , gned in värme i hans kalla ben .

En annan taxi tar _____ om skolbarnen .

Novelty hade flera trumf på _____ .

I första _____ har hon spelat dragspel och fiol .

Evaluation

For which levels is this exercise type relevant?

A1 A2 B1 B2 C1

Comments

Pilot 2 on L2 “alternative” data

- Automatic assigning new words to a proficiency level
 - We predict the level automatically
 - Learners (of a known level) get the word in an exercise (or a series of exercises)
 - We see whether learners can cope with it

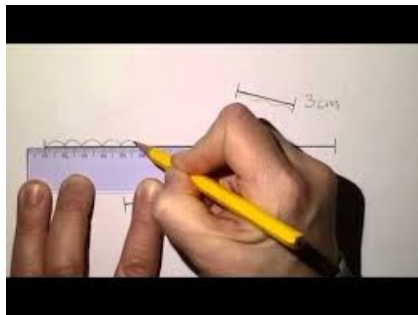


David Alfter

Algorithms and tools



L2 word-level normalization



- **Levenstein distance (as is)**
 - Good for advanced levels (edit distance of 1)
 - Fails at lower levels (with multiple edits)
- **Levenstein distance (for historical texts)**
- **LanguageTool + candidate ranking**
 - 73% correct variant selection
 - Failed to identify 30% of spelling errors





L2 error detection and correction

Within the SweLL infrastructure project



Task

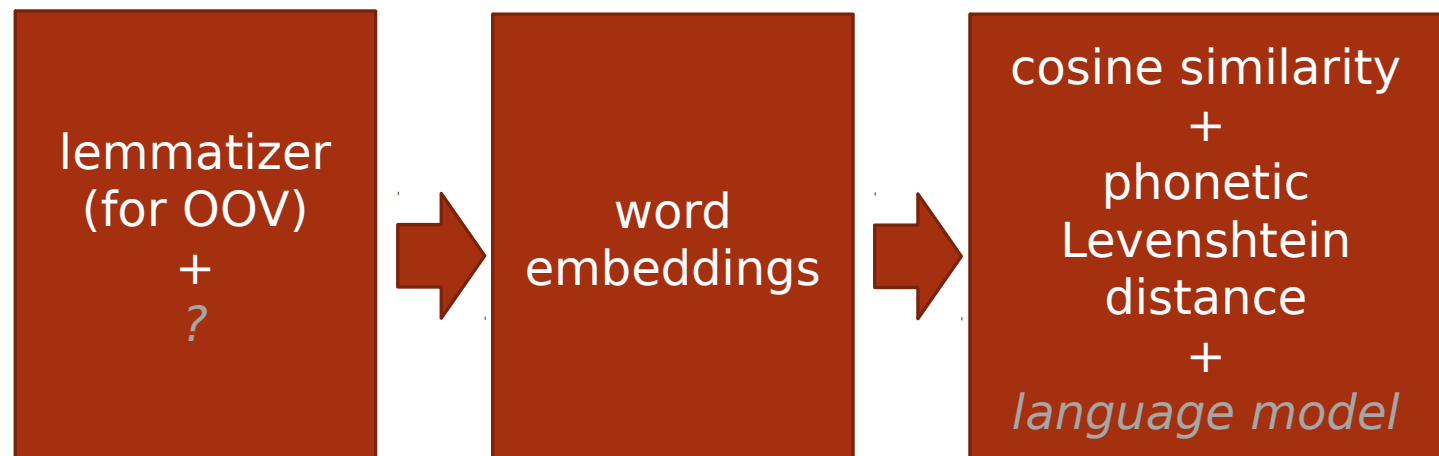


- **Spelling** error correction
- Steps:
 - 1. Detect errors
 - 2. Get correction candidates
 - 3. Choose best candidate (or rank candidates)

Suggested solution


- Steps:
 - 1. Detect errors: non-word errors (= **non-lemmatized** tokens)
 - 2. Get correction candidates: from **word embeddings**
 - 3. Choose best candidate (or rank candidates): **phonetic Levenshtein** distance

DETECT ERRORS GET CANDIDATES SCORE CANDIDATES





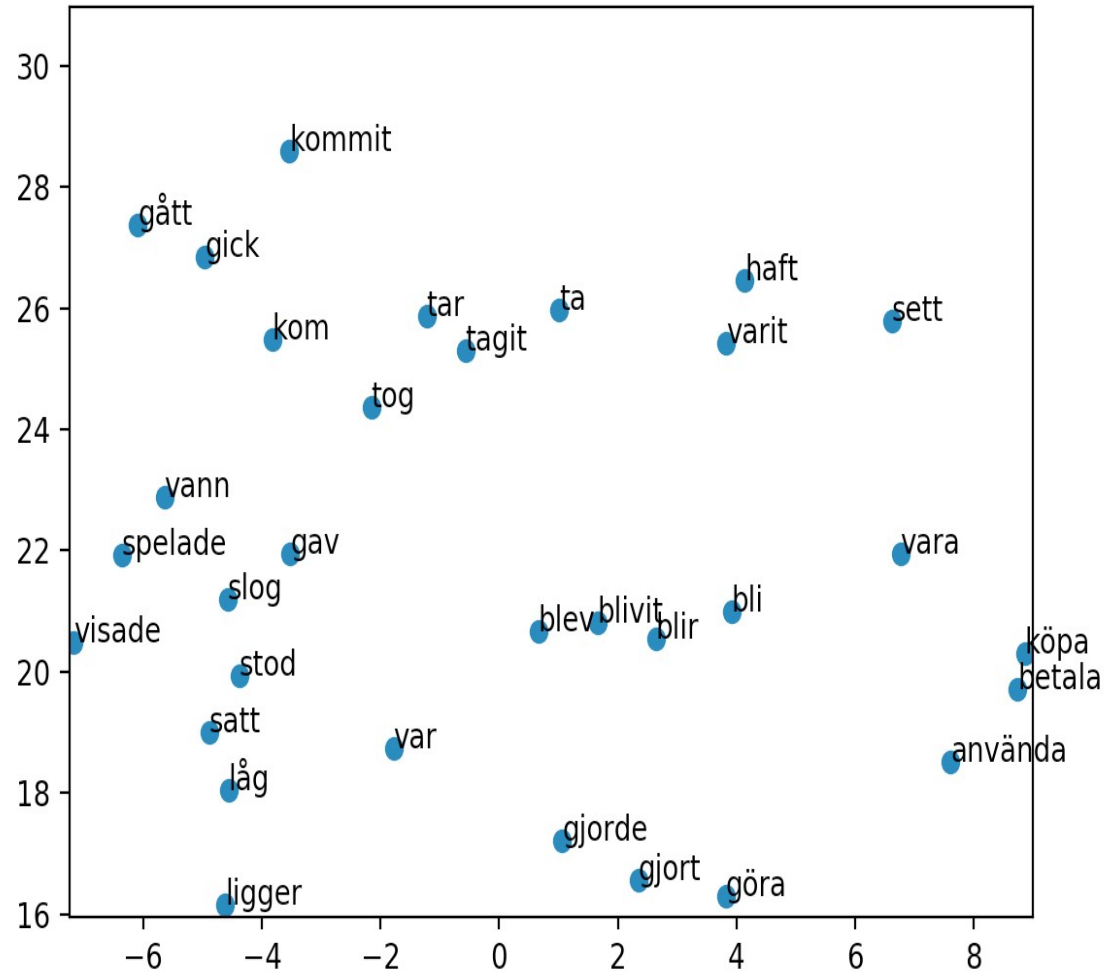
Step 2: Getting correction candidates



Experiments: WEs with character n-grams

- Trained a model using **FastText**
 - Not word as smallest unit (as in word2vec), but **character n-grams**
 - *E.g. sunny* = [sun, sunn,sunny],[sunny,unny,nny]
 - Advantages: possibility to find representations for **new** or **rare words**
 - **Corpora** used: COCTAILL+Läsbart+Åttasidor+SweLL (= **L2** and **easy-to-read** corpora including misspellings) + GP2013 = **3M words**, 62.253 unique
 - Default params except minCount=1 (to retain all misspellings)
- Seems to work **better** for **longer** words (> 3 char.s)
 - by default n-grams are between 3-6 characters
- How to evaluate?

Example of learned representations



Examples for related terms -nouns


- "klädder"
 - klädder 0.862
 - **kläder** **0.796**
 - ridkläder 0.749
 - gläder 0.747
 - dunkläder 0.745
- "moromor"
 - mor 0.983
 - mormmor 0.975
 - morbror 0.941
 - mormorsmor 0.898
 - **mormor** **0.879**

Examples for related terms - verbs

- "traffäs" -> "träffas" not in top 25 nearest neighbors... □
 - traffäs 0.878
 - traffas 0.842
 - traffa 0.818
 - traffar 0.808
 - traffik 0.758
- "hanlar"
 - halar 0.845
 - haar 0.797
 - hanar 0.786
 - himlar 0.784
 - harklar 0.779
 - **handlar 0.773**



Step 3: Selecting best candidates



Grapheme-to-phoneme (G2P) conversion

- Tool: g2p toolkit from **CMUSphinx** (speech recognition)
 - <https://github.com/cmusphinx/g2p-seq2seq>
 - recurrent neural networks (RNN) with long short term memory cells (LSTM)
 - TensorFlow based
 - No alignments needed in the training data
- Training data:
 - **Lexin** used (20.973 graphical form - phonetic transcription **pairs**)
- Performance on test set (ca. 2100 instances): **78.3% accuracy**(with 64 layers)

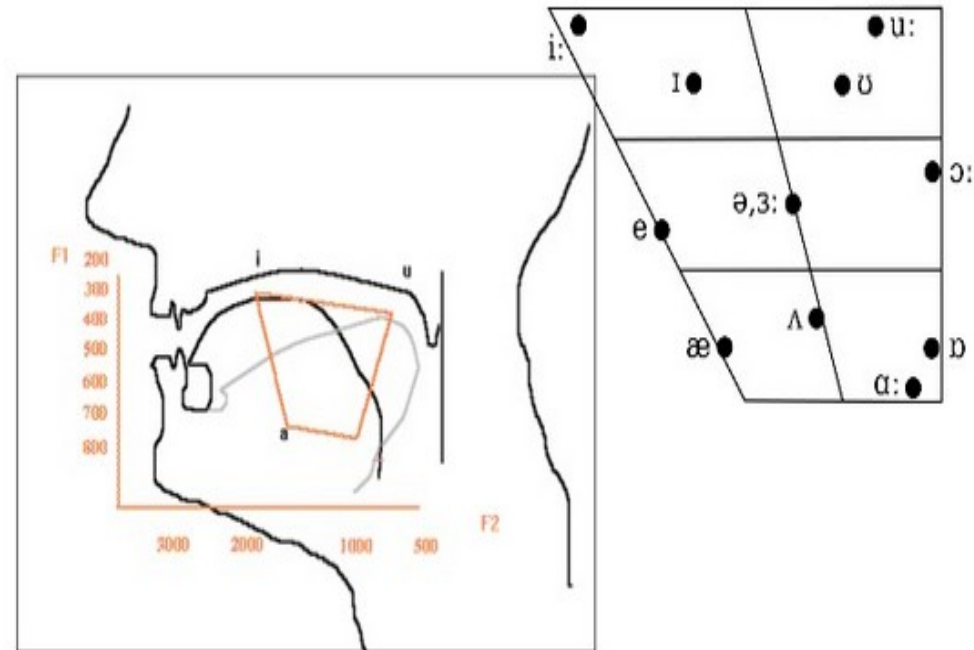
Example (mappings to SAMPA)

```
Creating 2 layers of 64 units.  
Reading model parameters from swe_g2p_2  
> fika  
f i: k a  
> trappuppgången  
t r a p u0 p g 0 N E n  
> låna  
l o: n a  
> märta  
m E t` a  
> slågaskåp  
s l o: g a s k o: p
```

G	H	I
allophone	lexin	sampa
ɑ:	a:	A:
a	a	a
u:	o:	u:
ʊ	o	U
o:	å:	o:
ɔ	å	O
ɘ:	u:	}:
ə	u	u0
e:	e:	e:
ɛ	e	E
i:	i:	i:
I	i	I
y:	y:	y:
Y	y	Y
ɛ:	ä:	E:
ɛ	ä	E
ø:	ö:	2:
ø	ö	2

Phonetic Levenshtein distance

- Phonemes represented along 3 dimensions (numeric or binary features)
 - Vowels: height, backness, roundedness
 - Consonants: place, manner, voice
- E.g. *tr**a**ffas* - *tr**ä**ffas*
 - *LevD*: **1**
 - *LevD-Phon*: **0.125**
- E.g. *tr**o**ffas* - *tr**ä**ffas*
 - *LevD*: **1**
 - *LevD-Phon*: **0.58**
-



Next steps for tool development

- Spelling error detection/normalization:
 - Compounding
 - Splitting
 - Real-word errors (drifting to phrase level analysis)
- Error detection/normalization on phrase level (need to identify which types to target first, e.g.)
 - grammar – which?
 - MWE?
 - Real words out-of-context?
- Anonymization:
 - Names, cities, addresses, dates – what to replace with?
 - Events – which? What to replace with?
- ...

User-oriented tools



Visualization and statistics

exploring data

- Korp for L2 texts – monolingual versus parallel fashion
- Strix – on text level

Can-do taxonomy

Something new? Should we?

- competence assessment depends not only on errors, but also successes (acc to CEFR document, Chapter 5, 9). “Can-do” statements are positively expressed!

https://www.coe.int/t/dg4/linguistic/Source/Framework_EN.pdf

- progress is measured not only through errors, mostly through development of new “structures” in the target language (see Processability Theory and its stages)

goo.gl/SjeGmd

- CAF-principles: Complexity, Accuracy, Fluency
- Second Language Acquisition research (based on an extensive review of research literature)
- learner texts (based on an inductive analysis of 10% of all learner texts)