Description of the SweLL corpus

Contents

General information 1

Metadata description for the SweLL corpus 3
- Personal metadata 3
- Task metadata 5
- Essay metadata 8
- Schools metadata 8

Manual coding / annotation in the SweLL corpus 9
- Pseudonymization codes 9
- Correction annotation codes 11

General information

Total nr essays: 525 (June, 2, 2020; the collection is not finalized and will grow further)

NOTE! In the SweLL corpus we use the term CORRECTION ANNOTATION instead of a more traditional ERROR ANNOTATION

Description of the project: https://spraakbanken.gu.se/en/projects/swell
The SweLL corpus is maintained at the University of Gothenburg, Språkbanken-Text <https://spraakbanken.gu.se>

Personal data management: Essays were collected following a consent from the learners. The consent allows the use of essays for research by registered (approved) users. Handwritten essays were transcribed using secure encrypted environment (SweLL kiosk). All essays were manually **pseudonymized** (using SweLL kiosk) based on the Pseudonymization guidelines:
https://spraakbanken.github.io/swell-project/Anonymization_guidelines

Mode: All essays were written as an exam, in a classroom. Most of the essays were written by hand and were transcribed later according to the Transcription Guidelines:
https://spraakbanken.github.io/swell-project/Transcription_guidelines

Time constraints and access to **allowed materials** varies between tasks (see details in Task Metadata for each particular task).

Pseudonymized essays were **normalized** and **corr-annotated** by L2 specialists.

To get information on **access** to the SweLL corpus, see the webpage:
https://spraakbanken.gu.se/en/projects/swell/swell4users

To cite the corpus:

Elena Volodina, Lena Granstedt, Arild Matsson, Beáta Megyesi, Ildikó Pilán, Julia Prentice, Dan Rosén, Lisa Rudebeck, Carl-Johan Schenström, Gunlög Sundberg and Mats Wirén (2019). The SweLL Language Learner Corpus: From Design to Annotation. *Northern European Journal of Language Technology,* Special Issue.

Read the article: https://nejlt.ep.liu.se/article/view/1374

Newer articles will be added on completion of the project

Metadata description for the SweLL corpus

Personal metadata

General	
Student ID	**370** unique students (numbers are not final!), e.g. **C16**. Letter prefix (for a school) + a running number
Birth year in 5-year intervals	1950-1954 -- 2000-2004
Gender	Kvinna, Man, Annat, Vill inte säga
Time in Sweden (sum in months)	0 - 315
Native language(s)	**72** unique languages in 101 unique combinations of 1-4 languages
Education	
Elementary education outside Sweden (nr months)	12 - 156
Elementary education in Sweden (nr months)	12 - 132
Introductory education in Sweden (nyanlända) (months)	1 - 36
Gymnasial education outside Sweden (nr months)	2 - 168

Gymnasial education in Sweden (nr months)	1 - 72
Professional education outside Sweden (nr months)	12 - 60
Professional education in Sweden (nr months)	3 - 72
University education outside Sweden (nr months)	12 - 228
University education in Sweden (nr months)	6 - 96
Professional degree	Free text (e.g. Ekonom)
Educational degree	Free text
Education: additional comments	Free text
Language information	
Education in L1 (hemspråk, education in Sweden)	Language name
Length of education in L1, nr months (hemspråk)	1 - 168
Swedish proficiency courses	Self-instruction, formal education
Length of Swedish proficiency courses, nr months	1 - 336
All known languages	List of languages

Other known language(s) except mother tongue(s)	List of languages
Best written language(s)	List of languages
Best spoken language(s)	Language name(s)
Language(s) used with the family	Language name(s)
Language(s) used with friends	Language name(s)
Metacomment	Free comment added by an assistant

Task metadata

Administrative Metadata	
Task ID	Task ID in the corpus. A letter prefix (for a school) + **T**(ask) + a running number, e.g. **AT14**. Total of 38 unique tasks
Semester (time span)	**VT-2018, HT-2018, VT-2019**
Task date	Year-week, e.g. **2018-W20**
Course type / school form	• Ungdomsgymnasiet • Universitetet • Vuxenutbildningen

Course level	• Behörighetsgivande kurs • Förberedande kurs • Grundläggande SVA dk3 • Inplaceringsprov SFI • SFI B / C / D • SVA 2 / 3 • TISUS
Grading scale	• A-F • G/U • SFI inplacering • Uppgiften har inte betyg
Writing task details	
Task type	• Behörighetstest • Formativ skrivuppgift • Inplaceringsprov • Mitterminsprov • Slutprov • Test inför NP (Nationella Prov)
Task - form / mode	Handskriven, Digital
Task duration (in minutes)	25--180
Text type / genre	• Argumenterande • Berättande • Beskrivande • Förklarande • Informellt mejl • Instruerande • Resonerande • Utredande • Återgivande
Task instructions	Free text alt. reference to an attachment
Allowed aids	Bilingual dictionary, Monolingual dictionary, Internet, etc.

Additional material	Free text comment
Additional comments	Free text comment
Additional comment on coursebooks used	Book title(s) / free text
Approximate level	● Nybörjare ● Fortsättning ● Avancerad
Tasks - subject / topic	● 1. Din första arbetsplats alt. Kvinnors arbete; 2. Mejl till kusin på besök i Sverige ● 1. När jag var liten och gick till skolan första gången; 2. Mejl till kusin på besök i Sverige ● Argumenterande text om arbetsmoral ● Argumenterande text/brev ● Berätta hur du bor! ● Berätta utifrån texten "Alma berättar" ● Beskriv - En god relation ● Beskriva - Min första kärlek ● Brott - orsaker och konsekvenser ● Demokratiska val - hur gammal ska man behöva vara för att rösta och varför? ● Diskuterande text om pengars betydelse ● En kulturupplevelse ● En plats du tycker om ● En viktig plats ● Enkel utredande text om litterära teman ● Familjen ● Ge tips och råd ● Ge tips och råd - en anställningsintervju ● Insändare ● Kommunikation och sociala medier ● Mejl till en vän ● Mina första intryck ● Objektivt utredande uppgift ● Om din bostad och om att bo ● Referat av texten "Giftermål ett större steg än barn" ● Skriv en insändare ● Skriv ett brev

	• Skriv ett mail • Skriv om en känd person • Två sätt att uppfostra • Utredande text (pm) övning inför NP • Världens lyckligaste länder • Övnings-pm inför NP
Task-url	In certain cases where handouts were used, attachments are available at the urls.

Essay metadata

Grade	If available, according to the grading scale for a particular Task

Schools metadata

School listing		*Description*
A	Vuxenutbildningscentrum	Inplacering SFI-utbildning: A-D
B	Gymnasieskola	
C	Komvux/SFI	SFI A-D
E	Behörighetsgivande kurser	motsv gymnasiet
F	Behörighetsgivande kurser	motsv gymnasiet
G	SFI-provet	SFI A-D
H	TISUS-prov	motsv gymnasiet
J	Grundläggande vuxenutbildning	

Manual coding / annotation in the SweLL corpus

Pseudonymization codes

Pseudonymization guidelines: https://spraakbanken.github.io/swell-project/Anonymization_guidelines (Needs updating!)

Category type	**Codes**	**Pseudonym / details**
Names	firstname_male firstname_female firstname_unknown initials middlename surname	replaced by an equivalent
Geographical data	city (+*foreign* for non-Swedish ones) area (+*foreign* for non-Swe ones) country geo place region street_nr zip_code	Swedish names replaced with dummy names; All other names with equivalent ones (i.e. cities with other cities)
Institutions	school work other_institution	replaced with equivalents
Transportation	transport_name transport_nr	replace with X-linjen randomly
Age	age_digits age_string	replaced with a random number in the ±2 span from the actual number

Dates	date_digits day month_digit month_word year	11/11/1111 replace randomly 11/11 replace randomly ±2 randomly
Miscellaneous		
	account_nr	
	e-mail	email@dot.com
	extra	
	license_nr *(e.g. cars)*	ABS 000
	other_nr_seq	
	phone_nr	0000-000000
	person_nr	123456-000
	url	url@com
	Zip-code	000 00
Sensitive	edu (education) prof (profession) fam (family members) sensitive (free text with hints on ethnical & sexual info, religious & political views)	Markup only, no replacement

Correction annotation codes

Normalization guidelines: https://spraakbanken.github.io/swell-project/Normalization_guidelines
Correction annotation guidelines: https://spraakbanken.github.io/swell-project/Correction-annotation_guidelines-and-codebook

ORTHOGRAPHY	
O (misspelling)	
O-Cap (capitalization)	
O-Comp (compounding)	
LEXIS (derivational morphology included)	
L (any other lexical problem)	
L-Der (wrong derivation mechanism used)	
L-FL (mix of foreign language in Swedish)	
L-Ref (reference error)	
L-W (wrong word choice)	
MORPHOLOGY (inflectional) ≈ PHRASE LEVEL	
M-Adj/adv (adjective instead of adverb)	
M-Case (case problem)	
M-Def (definiteness problem)	
M-F (problem on form / morphology level)	
M-Gend (problem with gender)	
M-Num (problem with number)	
M-Other (any other, not listed, problem on morphology level)	
M-Verb (any problem on the verb or verb phrase level)	
SYNTAX ≈ CLAUSE LEVEL	
S-Adv (word order, sentence adverbial placement)	
S-Clause (clause vs phrase level)	

S-Comp (phrase vs compound structure)	
S-Ext (extensive change)	
S-FinV (word order, verb placement)	
S-M (missing word)	
S-MSubj (missing subject)	
S-Other (any other problem/syntact. level)	
S-Type (change of construction type/phrase level)	
S-WO (any other word order problem)	
PUNCTUATION	
P (general miss on punctuation)	
P-M (missing)	
P-R (redundant)	
P-Sent (sentence segmentation)	
P-W (wrong)	
OTHER	
C (consistency change)	
Cit-FL (use of foreign language for citation)	
Com! (comment on an essay level)	
OBS! (comment on a token level)	
X (unintelligible)	
Unid (unidentified)	