INSTITUTIONEN FÖR
SVENSKA SPRÅKET

GU-ISS-2021-01

SweLL transcription guidelines, L2 essays

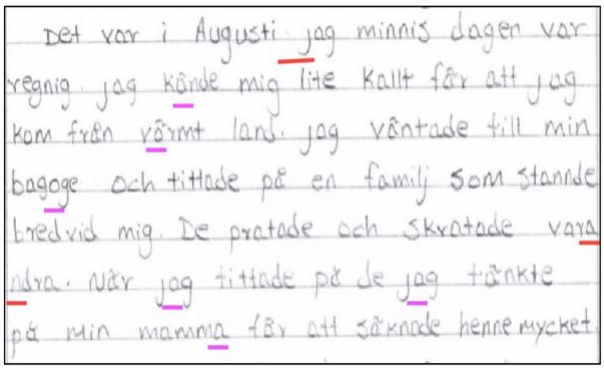
Elena Volodina, Bea Megyesi

Forskningsrapporter från institutionen för svenska språket, Göteborgs universitet
Research Reports from the Department of Swedish

ISSN 1401-5919

SweLL
Transcription guidelines

by Elena Volodina and Beáta Megyesi



Det var i Augusti . jag minnis dagen var
regnig. jag kände mig lite kallt för att jag
kom från vörmt land. jag väntade till min
bagoge och tittade på en familj som stannde
bredvid mig. De pratade och skratade vara
ndra. När jag tittade på de jag tänkte
på min mamma för att säknode henne mycket.

August 2021

The SweLL guideline series:

SweLL Transcription guidelines
by Elena Volodina and Beáta Megyesi

SweLL Pseudonymization guidelines
by Beáta Megyesi, Lisa Rudebeck and Elena Volodina

SweLL Normalization guidelines
by Lisa Rudebeck, Gunlög Sundberg and Mats Wirén

SweLL Correction annotation guidelines
by Lisa Rudebeck and Gunlög Sundberg

Preface

by Elena Volodina, Lena Granstedt, Beáta Megyesi, Yousuf (Samir) Ali Mohammed, Julia Prentice, Lisa Rudebeck, Gunlög Sundberg and Mats Wirén

During years starting 2017-2021 we have been working on setting up the main building blocks for empirically based research on Swedish as a second language which we release under the name of the *SweLL infrastructure*. This work entailed collecting and manually annotating learner written essays, which we refer to as *SweLL-gold corpus*. However, this process turned out to be highly versatile and involved a lot of work "behind the scene". First, to make sure the annotations are reliable, we invested extensive work into developing and documenting a taxonomy of corrections (or errors, a more traditional term used in other projects) and a taxonomy of personally identifiable information (PII, for successful pseudonymization). Second, to make sure that the manual annotation is as consistent as possible, we developed a set of tools to support the annotation itself and the management of the annotation process. Third, to make sure the resulting collection of essays can reach the intended user, we worked on legal aspects of access to the material as well as on visualization of the corpus so that it may be browsed and analyzed statistically, from the point of textual, educational and linguistic characteristics.

The current document is a part of the SweLL guidelines series consisting of four parts which aim to report how we have worked on the material and which decisions we have made. Guidelines are available for each step in the manual annotation process, including:

- Transcription guidelines
- Pseudonymization guidelines
- Normalization guidelines
- Correction annotation guidelines

We specifically described all processes in English to make sure our principles and experience can be of help to people working on other learner infrastructure projects independent of the language.

More information about the metadata used in the corpus and an overview of the taxonomies can be found here: https://spraakbanken.github.io/swell-release-v1/Metadata-SweLL

A short introduction to the SweLL project

SweLL - Swedish Learner Language – is a research infrastructure for Swedish as a second language. It was funded by Riksbankens Jubileumsfond 2017-2020 (IN16-0464:1), and had four participating universities: University of Gothenburg (project leadership), Stockholm University, Uppsala University and Umeå University.

The SweLL infrastructure project had as an aim to lay the fundament for digital Second Language Acquisition research by:
(1) collecting and manually annotating learner essays written by learners of Swedish at different levels of development
(2) developing well-functioning annotation principles, tagsets and processes, and thoroughly describing them
(3) developing and documenting digital tools for processing and storing of learner essays
(4) making the data and tools available through a portal developed for digital resources and tools for second language acquisition research of Swedish

The learner corpus infrastructure SweLL includes:

(1) The SweLL portal that is used for collection, storage and versioning of essays, administration of the annotation process, statistical overview, inter-annotator agreement, import and export of the data.

(2) The SweLL portal hosts a collection of more than 680 essays that have been digitized and manually transcribed from handwritten samples during the course of this project. All essays were pseudonymized to protect the privacy of each individual learner. A larger portion of the essays – 502 texts, the so-called SweLL-gold corpus – were normalized, i.e. re-written in order to fit the norms of standard Swedish by correcting erroneous and deviant language, and each correction was assigned a correction label describing the difference between the learner's version (source text) and the corrected version (target text).

(3) Several other tools are available for future users of the infrastructure:

- SVALA annotation tool for performing manual annotation steps (pseudonymization, normalization, correction annotation) (Wirén et al. 2019)
- Automatic pseudonymizer service (included as a part of the SVALA tool, and available through github for potential extensions or re-use in other projects) (Volodina et al. 2020)

(4) Extensive work was done to document how the learner data were processed, which includes

- selection and documentation of associated metadata (corpus-related, student-related, task-related, school-related and essay-related)
- taxonomies for pseudonymization and correction annotation, and
- guidelines for all (manual) annotation steps (transcription, pseudonymization, normalization and correction annotation)

(5) Thorough work has been carried out to make sure that the GDPR guidelines and ethical principles are followed. In consultation with the university lawyers at the University of Gothenburg, the access principles have been defined and legal basis double-checked. Access to essays can be granted following an application. As of 2021, according to the GDPR, users outside Europe cannot

get immediate access to the data in its entirety. Their applications need to be processed by the university lawyers on a case-to-case basis. Applicants inside EU can get access to the full dataset provided their intended use targets L2-oriented research, development or pedagogical applications.
(6) The data can be browsed through corpus search interface Korp (https://spraakbanken.gu.se/korp/) with specific solutions for L2-material facilitating filtering for e.g. texts written by writers of a certain age, gender, mother tongue, or writers at a certain proficiency level or course, a certain text type – all with a possibility for full-text view.

More information about the project and tools are available at the project page: https://spraakbanken.gu.se/projekt/swell

Acknowledgments

Our gratitude goes to teachers and assessors who supported us during the essay collection stage. They were many, and without their enthusiasm we would not have been able to build any infrastructure today. We are grateful to learners who were positive to allow their texts to be used for teaching, research and development.

We would like to acknowledge a group of advisors, assistants and developers who are not listed as co-authors of this preface, but who have been involved during the different periods of the SweLL project:

- University of Gothenburg: Arild Matsson, Ildikó Pilán, Monica Reichenberg, Dan Rosén, Carl Johan Schenström
- Stockholm University: Sofia Brusling, Sofia Johansson, Miku Westerholm

We would also love to extend our gratitude for the generous financial support provided by the main funder Riksbankens Jubileumsfond, as well as for the indispensable financial support during the last months of finalizing the infrastructure from Språkbanken Text and Swe-Clarin at the University of Gothenburg as well as by the Department of Swedish Language and Multilingualism at Stockholm University.

August 2021

Elena Volodina, University of Gothenburg
Lena Granstedt, Umeå university
Beáta Megyesi, Uppsala university
Yousuf (Samir) Ali Mohammed, University of Gothenburg
Julia Prentice, University of Gothenburg
Lisa Rudebeck, Stockholm University
Gunlög Sundberg, Stockholm university
Mats Wirén, Stockholm university

swell-project

SweLL transcription guidelines, L2 essays

Elena Volodina, Bea Megyesi, June, 04, 2018 - August, 19, 2021;

Online version of this document: https://spraakbanken.github.io/swell-project/Transcription_guidelines

Transcription flow

This document contains instructions for manual conversion of hand-written essays to a digital format, and recommends the following flow:

1. Acquaintance with guidelines (annotators)
2. Transcription workshop (annotators + researchers)
3. Transcription (individual annotators)
4. Cross-consultation, in uncertain cases (annotator-annotator or annotator-researcher)
5. Transcription check (third party)

Acquaintance with the guidelines (1) means actual study of this document, optimally combined with a practical test-case using a number of real-life essays, to see how different questions can be guided.

Transcription workshop (2) is a practical one-day session when several annotators work on actual essays and discuss uncertain cases between themselves and with a responsible researcher. The workshop is aimed at resolving subjective judgements in favour of objective decisions. Optimally, annotators involved in this process can build some network so that when uncertainties arise during their later work, they can ask each other.

Transcription, individual phase (3) is an individual process when each annotator is working on his/her share of essays.

Cross-consultation (4) is a step which can be of use in uncertain cases. We

recommend to get in contact with another assistant or a responsible researcher to double-check any uncertainties.

Transcription check (5) is performed by a third party, e.g. another annotator. During this stage random checks are performed on the transcribed files.

Transcription principles

The two **major principles** for essay transcription are:

- not to correct author's mistakes and
- not to make personal assumptions.

Transcription rules

The following **rules** should apply to transcription of hand-written essays:

(1) Authenticity of writing

Errors should be preserved from the hand-writing, e.g. *no error correction!* (see ③ in Figure 1). Tips: disable spell-checker.

If there is a dubious case - for instance, when you are uncertain whether the learner has written correctly or incorrectly - a *positive assumption* should be made. For example, if it is unclear whether two words have been written as one or as two items (with too little space between them), the "positive" assumption would be that the learner meant to write two items, and in practice in the transcribed format the string should be separated into two words. When it is obvious that the two words are written as one, that should be preserved (typed as one item). (See ④ in Figure 1)

(2) Handwriting

In many cases some basic knowledge of Swedish should help to understand what is written (see ⑤ in Figure 1; as well as Figure 2, lilac underlinings). If the handwriting is

illegible, write $ (dollar-sign) for each character that cannot be understood.

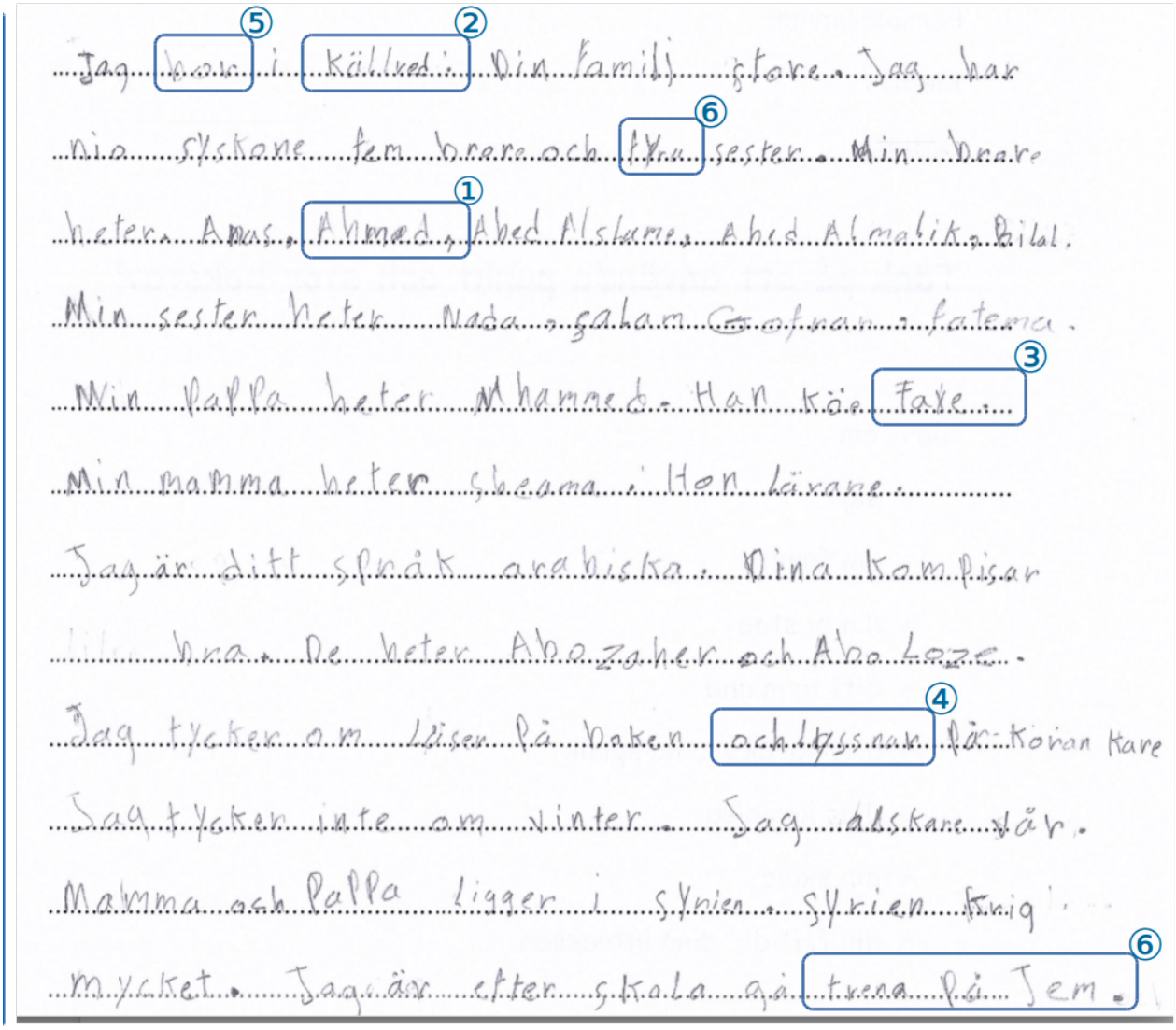


Figure 1. Essay example. Level A1 (beginner), nr tokens: 117, topic: Presentation/Om mig; transcription time: 15 min.

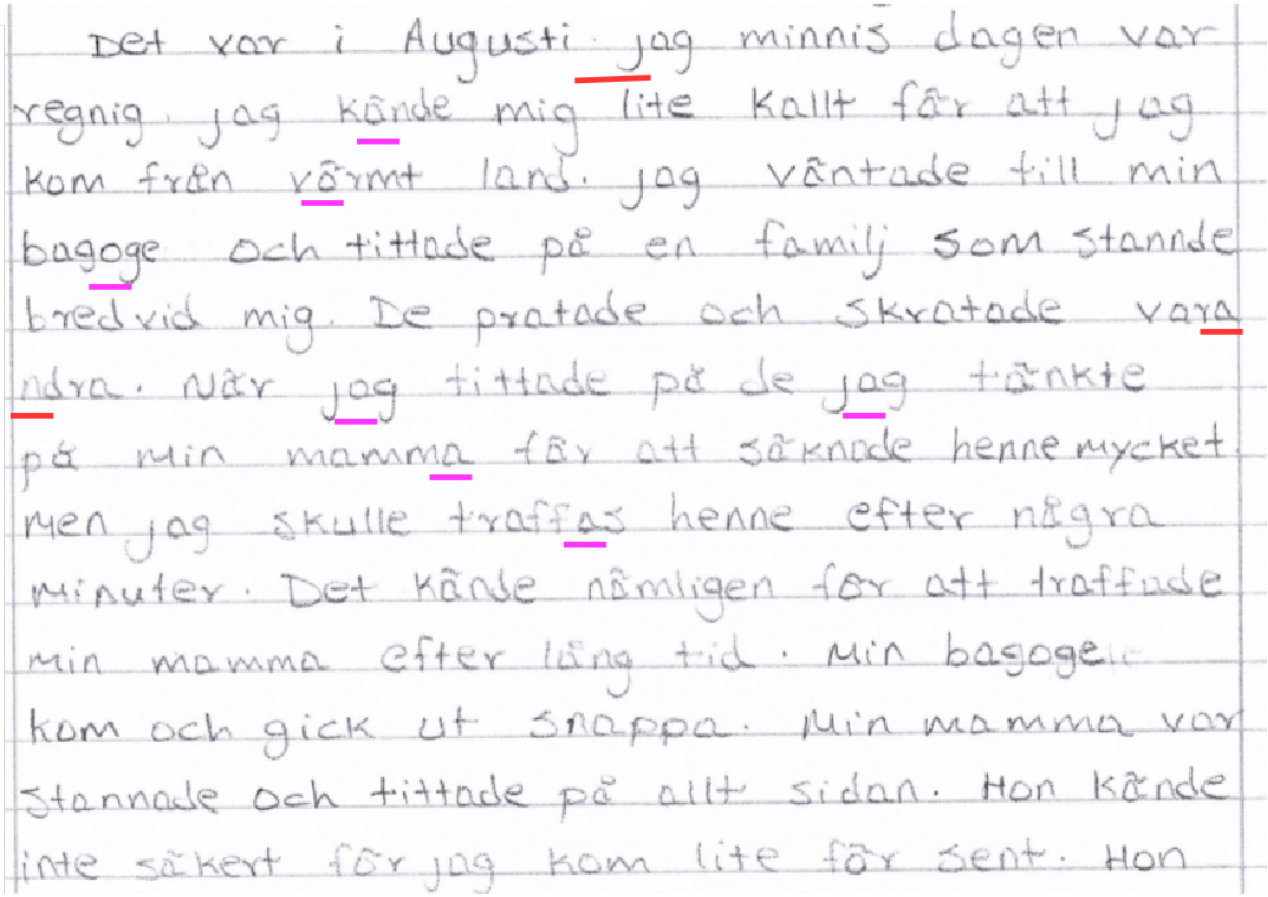


Figure 2. Deciphering letters in student writing: o vs a; capital/non-capital; hyphenation. Level B1 (intermed.), nr tokens: 330, topic: Min första dag i Sverige; transcription time: 34 min.

(3) Non-existent letters

Sometimes students can be very creative and "invent" their own letters (see Figure 3). In the case when you know there is an equivalent letter in another language, use that one, as for example, in the first word in Fig.3: *Sverigë*.

In the case, if there is no way to reflect that letter in writing, choose the closest one in shape, keeping to the positive assumption described in (1). For example, in the second example in Fig.3, the options could be o and å, but å holds the positive assumption, så the transcribed version should contain the word *går*.

If there is no way you can report a corresponding "created" letter, use dollar-sign $ as if it is an unintelligeble letter.

Figure 3. Invented letters

(4) Graphical issues (supra-linguistic features)

1. Insert line breaks (“Enter”) to introduce new paragraphs.
2. Differentiate between capital letters and non-capital ones (see ⑥ in Figure 1).
3. Keep smileys
4. If a student has stricken out some text, that text should not be transcribed.
5. Don’t bother about leaving indentation. We are primarily interested in the contents and the language, not in the graphical reproducibility.
6. All edits, like underlinings, are discarded. We keep only the text.
7. Comments in the margin are a case for interpretation. If it is obvious that the margin text is a part of the running text, it should be added into the essay. Otherwise - not. (We need, though, to make transcribed versions comparable to the digitally-born essays. Here, we need to see whether there are any cases where students have left their comments in the digitally-born versions, e.g. in the form of footnotes.)

Rule of thumb

Re-read each hand-written essay once again and compare with your transcribed version. You will get used to the student’s handwriting by then, and will - probably - understand better what is written. Another reason for re-reading the essays is to double-check that no unintended error-correction is introduced (rules 1-4).

Time estimation

To be able to give some time estimation, we have taken time for transcription of several essays per level. The summary follows in table 1 below.

It will take longer per essay in the beginning, when annotators are not yet confident with the guidelines and the process itself. The time will also depend upon the legibility of handwriting, length of the text and challenges of the writing, i.e. presence of challenging interpretations/uncertainties. Take the time estimations below only as an

approximation.

Average	A1	A2	B1	B2	C1
Characters / essay	459	702	1761	-	-
Words / essay	85	134	331		
Minutes / essay	10	9	24,5		
Words / minute	8,25	15	13,5		

Table 1. Time estimation for essay transcription at different levels

During your work, write down your time per essay in an excel sheet acc. to the example below (Table 2):

Essay-ID	Level	L1	Nr words	Time in minutes	Comment, if necessary

Table 2. Time estimation for annotator work on transcription

Text format / use of a kiosk version of SVALA

The work on hand-written essays is potentially risky, since certain amount of personal information in the text (as well as handwriting itself) may give away a person behind the text. That is why this work has to be performed in a safe environment. We use a specially designed SweLL "kiosk" option for that, which is a special encrypted computer that has extremely limited access to the internet and has pre-installed database for managing essays. See instructions for "kiosk" here: https://spraakbanken.github.io/swell-project/SweLL_kiosk_user_manual – ***Note!*** *In Swedish!*

You can save your work in any format while you are working, but your should deliver it in a plain text format (in unicode utf-8).

Save information about time used for transcribing an essay - for our statistics.

GU-ISS, Forskningsrapporter från Institutionen för svenska språket, är en oregelbundet utkommande serie, som i enkel form möjliggör spridning av institutionens skriftliga produktion. Det främsta syftet med serien är att fungera som en kanal för preliminära texter som kan bearbetas vidare för en slutgiltig publicering. Varje enskild författare ansvarar för sitt bidrag.

GU-ISS, Research reports from the Department of Swedish, is an irregular report series intended as a rapid preliminary publication forum for research results which may later be published in fuller form elsewhere. The sole responsibility for the content and form of each text rests with its author.