INSTITUTIONEN FÖR
SVENSKA SPRÅKET

GU-ISS-2021-03

SweLL normalization guidelines

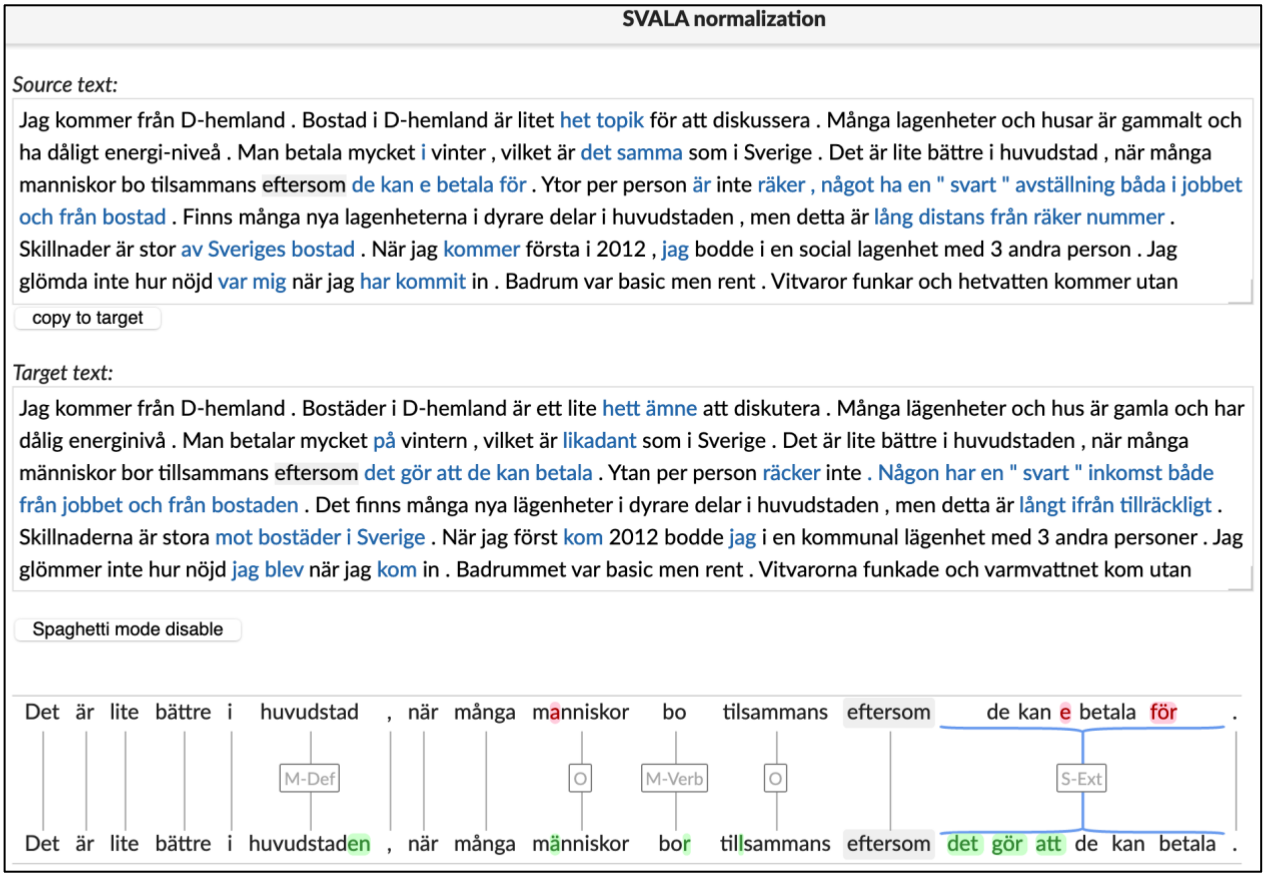
Lisa Rudebeck, Gunlög Sundberg, Mats Wirén

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

ISSN 1401-5919

SweLL
Normalization guidelines

by Lisa Rudebeck, Gunlög Sundberg and Mats Wirén



SVALA normalization

Source text:

Jag kommer från D-hemland . Bostad i D-hemland är litet het topik för att diskussera . Många lagenheter och husar är gammalt och ha dåligt energi-niveå . Man betala mycket i vinter , vilket är det samma som i Sverige . Det är lite bättre i huvudstad , när många manniskor bo tilsammans eftersom de kan e betala för . Ytor per person är inte räker , något ha en " svart " avställning båda i jobbet och från bostad . Finns många nya lagenheterna i dyrare delar i huvudstaden , men detta är lång distans från räker nummer . Skillnader är stor av Sveriges bostad . När jag kommer första i 2012 , jag bodde i en social lagenhet med 3 andra person . Jag glömda inte hur nöjd var mig när jag har kommit in . Badrum var basic men rent . Vitvaror funkar och hetvatten kommer utan

Target text:

Jag kommer från D-hemland . Bostäder i D-hemland är ett lite hett ämne att diskutera . Många lägenheter och hus är gamla och har dålig energinivå . Man betalar mycket på vintern , vilket är likadant som i Sverige . Det är lite bättre i huvudstaden , när många människor bor tillsammans eftersom det gör att de kan betala . Ytan per person räcker inte . Någon har en " svart " inkomst både från jobbet och från bostaden . Det finns många nya lägenheter i dyrare delar i huvudstaden , men detta är långt ifrån tillräckligt . Skillnaderna är stora mot bostäder i Sverige . När jag först kom 2012 bodde jag i en kommunal lägenhet med 3 andra personer . Jag glömmer inte hur nöjd jag blev när jag kom in . Badrummet var basic men rent . Vitvarorna funkade och varmvattnet kom utan

Det är lite bättre i huvudstad , när många manniskor bo tilsammans eftersom de kan e betala för .

Det är lite bättre i huvudstaden , när många människor bor tillsammans eftersom det gör att de kan betala .

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 - **Swe**dish **L**earner **L**anguage – 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

Normalization guidelines

Lisa Rudebeck, Gunlög Sundberg, Mats Wirén (May 2021)

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

Contents

1. The purpose of the normalization

Normalization in SweLL means editing of the original learner text in such a way that the normalized version of the text adheres to standard Swedish text norms. The purpose of the normalization is twofold:

1. To render the text in a version which is amenable to automatic annotation using a standard linguistic analysis pipeline. (For Swedish, such a pipeline is *efselab* or

Sparv. The SweLL data is currently processed with the Sparv-pipeline. It is possible to re-annotate later with other pipelines.)

2. To provide a basis for the correction annotation through an explicit representation of the specific standard version of the text to which the original version of the text is related. Such a representation is highly useful by allowing for many new types of search of the corpus, for instance:

- Finding missing occurrences of a construction in the learner text in the sense that it could or should have been used, for example, when something that ought to have been expressed using a passive was not.
- Finding mismatches between the learner and corrected text, for example, an adjective in the learner text that corresponds to an adverb in the corrected text, or cases when a content word in the original text has been changed in the normalized text.

2. Fundamental values

The normalization is carried out by balancing the following fundamental values:

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

The conflict between adherence to standard Swedish norms (1) and fidelity to the original text (2) is the very basis for the normalization; only in cases where the original text deviates from standard Swedish norms is a normalization called for. But the two sides of the fidelity to the original text may also be conflicting, so that a greater similarity to the original text string may mean a less effective communication of the assumed intended meaning, and vice versa, as the following example illustrates:

- Firsta gång såg jag henne i **gymnastik sckola** (C151CT23)

Here, a strict application of principle 2a (in combination with principle 1), would yield the normalization *gymnastikskola*, but since the rest of the text makes this an unlikely interpretation of the writer's intended meaning, the string *gymnastik sckola* is instead normalized as *gymnasieskola*, which seems to be what the writer meant.

The normalizer should thus strive to create a text version which adheres to the norms of standard Swedish, while staying as close to the original text string as possible and communicating the perceived intended content as effectively as possible. The result of this balancing act can be seen as an *interpretation* or *translation* of the original text into "standard Swedish".

It should be clear that this process of normalization by necessity involves an element of subjectivity; in the typical case there are several possible normalizations of a text. This is a necessary consequence of our choice to provide one single normalization in the tool Svala, in which norms on all linguistic levels are considered simultaneously, from orthography to syntax and sentence-internal semantics. First of all, a normalization which includes wording and morphosyntax highly depends on the normalizer's interpretation of the source text, and even on the basis of a specific interpretation it is far from always the case that there is one normalization which is unequivocally optimal, as the following example may serve to illustrate:

- **Original**: *Och det finns olika former av svarta tjänster* ***exempelvis av detta*** *är svartjobb , svart hyra och svartmarknad* (E7ET8)
- **Normalization 1**: *Och det finns olika former av svarta tjänster,* ***exempelvis*** *svartjobb , svart hyra och svartmarknad*
- **Normalization 2**: *Och det finns olika former av svarta tjänster.* ***Exempel på detta*** *är svartjobb , svart hyra och svartmarknad*

Moreover, it is obviously impossible to base a normalization which includes wording and morphosyntax on a finite number of explicit principles. The most important basis for securing the quality of our normalizations and upholding the fundamental values is, instead, our *methodological practices*. These are presented below, after a section on adherence to the norms of standard Swedish.

We would like to stress that the Svala tool may be used for the visualization and analysis of deviations between a specific source text and *any* normalization of this text. Researchers who wish to use the Svala tool to relate the learner texts to their own normalizations, based on other principles or methods than ours, are free to do so.

3. Adherence to the norms of standard Swedish

The normalized text version should contain no obvious deviations from standard Swedish norms. (There are two exceptions to this: **(1)** Unintelligible strings may be X-marked without being changed, see section 6. **(2)** In some cases, non-Swedish strings are left untranslated and marked Cit-FL, see section 5.)

The norms considered are norms for spelling and inflection of specific words, general punctuation norms, general morphological and syntactic rules and patterns, as well as well-established collocational patterns and norms for the usage range of specific words and expressions. However, the acceptance for unusual expressions is fairly high, and the understanding of "standard Swedish" is quite encompassing, including expressions and constructions which are widespread throughout the Swedish language community within consciously edited texts of any prose genre.

Norms concerning the composition of texts at a discourse level, i.e. beyond those which may be dealt with sentence-internally, are generally not considered. The normalization thus involves no changes of the ordering of sentences or paragraphs, etc., nor any deletions of whole sentences, not even in cases of seemingly unintended repetitions. However, the delimitation of sentences may be changed, for instance by exchanging a conjunction for a sentence-delimiting punctuation mark in a very long list of clauses joined by conjunctions.

While only intra-sentence changes are made, the context provided by the rest of the text is taken into consideration when judging each sentence. For instance, anaphoric expressions and conjunctions may be changed due to intra-sentential relationships. And the interpretation of an expression in one sentence is often affected by contextual information.

In the following, we comment on our implementation of norm-adherence for orthography and inflectional patterns of specific words, and for punctuation and sentence segmentation. When it comes to norms for wording, collocational patterns, and general morphosyntactic rules and patterns, we refer directly to the section on methodological practices. The treatment of non-Swedish words is discussed in a separate section.

3.1 Orthography and inflectional patterns of specific words

Norms for orthography, as well as for inflectional patterns of specific words (i.e. their adherence to a specific conjugation or declination), are generally the most clearly codified and stable ones, only occasionally giving room for subjective judgement or acceptable alternatives. In the minority of cases when alternative spellings or inflectional patterns are widely spread and/or codified in central lexicographic sources (such as *Svenska Akademiens ordlista*), either of these forms are accepted. If the source text contains one in a pair or set of such alternative forms, this form should not be changed on the basis of, for instance, style, frequency, text-internal consistency, or explicit recommendation in normative sources.

Examples of such alternative, and thus equally accepted, orthographic forms are *sen/sedan, mejl/mail, nån/någon, dom/de(m)* and *ska/skall*. And examples of alternative and equally accepted inflectional patterns for specific words are *partner* (null plural)/*partners, kolleger/kollegor, dåligare/sämre, givit/gett* and *lyste/lös*.

3.2 Punctuation and sentence segmentation

Swedish punctuation norms include both stricter and softer ones. One example of a strict punctuation norm is that in a list of items, separated by commas and by a conjunction before the last item on the list, a comma should not occur before the conjunction. Deviations from such strict norms should always be corrected in the normalization, as illustrated in the following example.

- *Jag köpte bananer, apelsiner* **,** *och päron. –> Jag köpte bananer, apelsiner och päron.*

However, the normalization of punctuation also involves alterations of punctuation for the sake of readability, even when no strict norm is involved. This includes, for instance, addition of commas separating long main clauses. A similar approach is taken to the segmentation of sentences; very long sentences may be divided for the sake of readability, as in this example.

- *Efteråt inträffade revolution inom kommunikation område i det tjugonde och tjugoförsta talet som uppfinningen av tv och utveckling av internet i hela världen* ***och*** *denna uppfinningar underlättade den globala kommunikationen så att världen har blivit en liten by –> Därefter inträffade en revolution inom kommunikationsområdet i det tjugonde och tjugoförsta århundradet , såsom uppfinningen av tv och utvecklingen av internet i hela världen. Dessa uppfinningar underlättade den globala kommunikationen så att världen har blivit en liten by*

The acceptance for *satsradning*, i.e. the practice of separating main clauses with a comma (without a conjunction) instead of with a period or another sentence-dividing punctuation mark, varies between genres. On the basis of our encompassing understanding of “standard Swedish” our acceptance for *satsradning* is fairly high. We do not correct instances of *satsradning* solely on the basis of style considerations, but may correct it because the separated clauses are not closely related by a causal relationship or the like, or for the sake of readability.

4. Methodological practices

1. The team of normalizers has been small (4 persons) but includes expertise in linguistic structure, Swedish language norms, and Swedish as a second language.
2. The first 20 normalizations of a normalizer are double-checked altogether by another team-member.
3. Whenever a normalizer comes across a particularly difficult decision, a second opinion from a co-normalizer should be sought. The resulting discussions should be documented for future reference, as a basis for similar normalization decisions as well as for future accounts of the decisions both within and outside the project group.
4. Often an overview and reading of the whole text renders clues for the interpretation of the writer’s intentions or use of vocabulary. A needed and helpful practice can be therefore be to read a text several times, and not normalize sentence by sentence.
5. When several texts belong to the same writing task, and are written by learners in the same group, it is good practice to normalize them together. It gives the normalizer a better understanding of the texts in relation to the given task.

5. Non-Swedish words and sequences

When coming across a word or a sequence of words stemming from a non-Swedish language, the normalizer has the following options:

1. *The word or sequence is left unchanged*.

   1.1 The normalizer judges the word/sequence as having been incorporated into written standard Swedish, and the word/sequence is thus kept unchanged. This judgement is based on the normalizer’s acquaintance with written Swedish, and may, in case of doubt, be informed by a secondary opinion from another member of the team of normalizers, by searches in corpuses or on the Internet, or, in some instances, by information in lexicographic sources. (The fact that a certain word or phrase is not included in dictionaries is in itself not sufficient to judge it as not belonging to standard Swedish.)

1.2 The normalizer judges the word/sequence as a genre appropriate usage of cited foreign language (explicitly signaled citations, code switching etc.). In such cases the word/sequence is left unchanged, but marked with the tag Cit-FL (see the Svala manual). The word/sequence **is not corrected to fit the norms of the source language**.

Judged as appropriate code switching:

- *Badrum var **basic** men rent –> Badrummet var **basic** men rent*
- *gillar du **quiz nights**? –> gillar du **quiz nights**?*

Clearly marked citation of Norwegian passage:

- *I samma artikel skriver Bengt Östling om man läser några webbsidor där norska ungdomar debatterar , förstår man att diskussionen om den obligatoriska nynorskan är inflammerad . “ **Ett språk som holdes kunstig i live gjennom tvan og finansiering gjennom skatt , og sakte men sikkert dör ut ja . Det finns ikke vilje hos folk til å beholde nynorsk** “ , lyder det i ett debattinlägg . –> I samma artikel skriver Bengt Östling att om man läser några webbsidor där norska ungdomar debatterar , förstår man att diskussionen om den obligatoriska nynorskan är inflammerad . “ **Ett språk som holdes kunstig i live gjennom tvan og finansiering gjennom skatt , og sakte men sikkert dör ut ja . Det finns ikke vilje hos folk til å beholde nynorsk** “ , lyder det i ett debattinlägg .*

1.3 If a word or string is recognized as likely belonging to another language, and the language knowledge within the team of normalizers does not suffice to interpret it, no further efforts is made to interpret the word/string. It is left unchanged and marked with the X-tag (unintelligible string, see below).

2. *The word or sequence is translated into Swedish.*

The normalizer does not judge the word/sequence as part of standard Swedish, nor as a genre appropriate usage of cited Swedish language. The normalizer is however able to interpret the word/sequence, and thus translates it into Swedish

6. Unintelligible and unreadable strings

When the normalizer comes across a string which she is unable to interpret, it is marked with the X-tag. The word/passage may either be left unchanged, or the normalizer may provide a guess as to its interpretation in the normalization.

- *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*** *–> Jag gick till dem och frågade om hennes namn och bad dem att inte berätta något .* ***Jag var*** *blyg av mig så jag letade tills jag hittade hennes hem* ***och bad henne och pjäsen idag dig varje morgon jag går att se det och låta dig gå till skolan***

Both of the marked strings are X-marked as unintelligible, but in the first case a guess is provided in the normalization, while the second string is left unchanged.

Note: Since an X-marked passage may be left unchanged in the normalized version, a normalized text may include some passages which do not adhere to the norms of standard Swedish.

Strings which have been marked as unreadable by the transcriber (with "$" representing an unreadable symbol) are treated as any other string; if the surrounding context provides a basis for a sound interpretation, the normalization is based on this interpretation. If the unreadable string is also uninterpretable, it is marked with the X-tag, and the normalizer may either provide a guess in the normalization, or keep the string unchanged.

7. Some special procedures with tokenization and punctuation

The normalization procedure involves some exceptions to the regular use of punctuation and spaces. These exceptions are due to tokenization procedures, and to the fact that many of our originals are hand written, which may make it hard to distinguish a hyphen from a dash, etc. (The effects of tokenization on the handling of the texts in Svala are described in the Svala manual.)

These special procedures with tokenization and punctuation are:

- Errors involving the incorrect placement of a space before a punctuation mark (which is not part of an abbreviation or another one-token string) will not be corrected, since a space is always inserted before such punctuation marks in the process of tokenization.
- Errors involving the lack of a space between a punctuation mark and the following word are corrected in the normalization process (a space is inserted), but will

nevertheless be left untagged in the correction annotation process.

  - Jog kan ante skriv meka ord .tack –> Jag kan inte skriva många ord. Tack.

- Instances where a hyphen has been used in the original text where a dash would be more appropriate are left uncorrected.

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.

GÖTEBORGS
UNIVERSITET