INSTITUTIONEN FÖR
SVENSKA SPRÅKET

GÖTEBORGS
UNIVERSITET

GU-ISS-2021-04

SweLL correction annotation guidelines

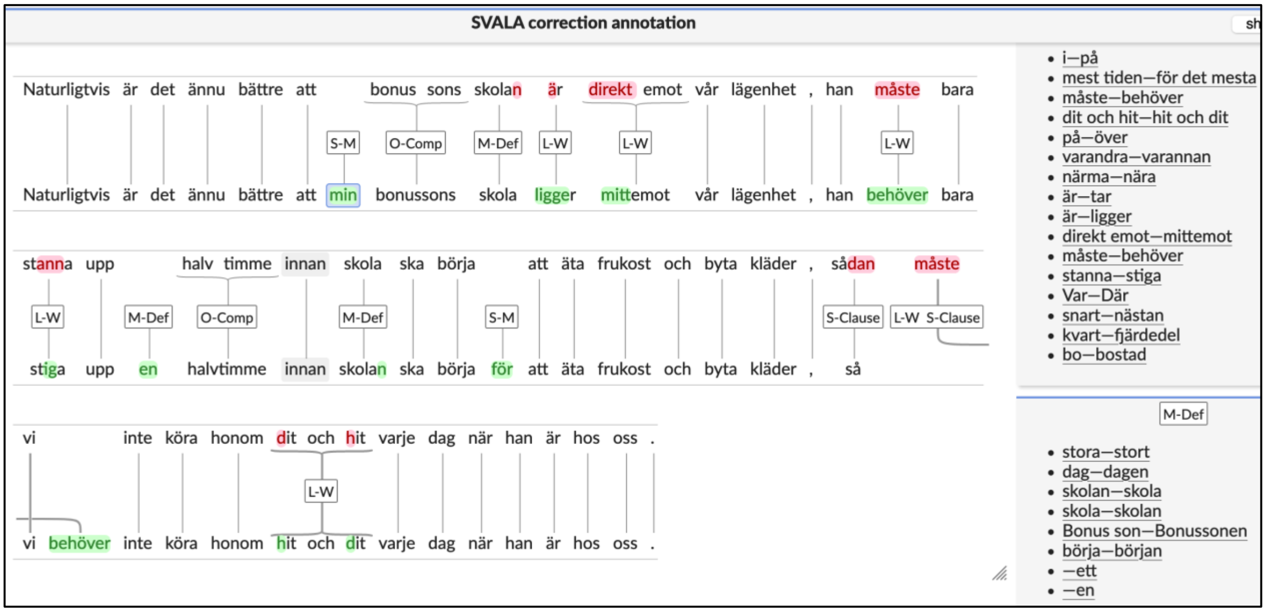
Lisa Rudebeck and Gunlög Sundberg

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

ISSN 1401-5919

SweLL
Correction annotation guidelines

by Lisa Rudebeck and Gunlög Sundberg



SVALA correction annotation

Naturligtvis är det ännu bättre att bonus sons skolan är direkt emot vår lägenhet , han måste bara stanna upp halv timme innan skola ska börja att äta frukost och byta kläder , sådan måste vi inte köra honom dit och hit varje dag när han är hos oss .

Naturligtvis är det ännu bättre att min bonussons skola ligger mittemot vår lägenhet , han behöver bara stiga upp en halvtimme innan skolan ska börja för att äta frukost och byta kläder , så vi behöver inte köra honom hit och dit varje dag när han är hos oss .

- i—på
- mest tiden—för det mesta
- måste—behöver
- dit och hit—hit och dit
- på—över
- varandra—varannan
- närma—nära
- är—tar
- är—ligger
- direkt emot—mittemot
- måste—behöver
- stanna—stiga
- Var—Där
- snart—nästan
- kvart—fjärdedel
- bo—bostad

M-Def

- stora—stort
- dag—dagen
- skolan—skola
- skola—skolan
- Bonus son—Bonussonen
- börja—början
- —ett
- —en

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

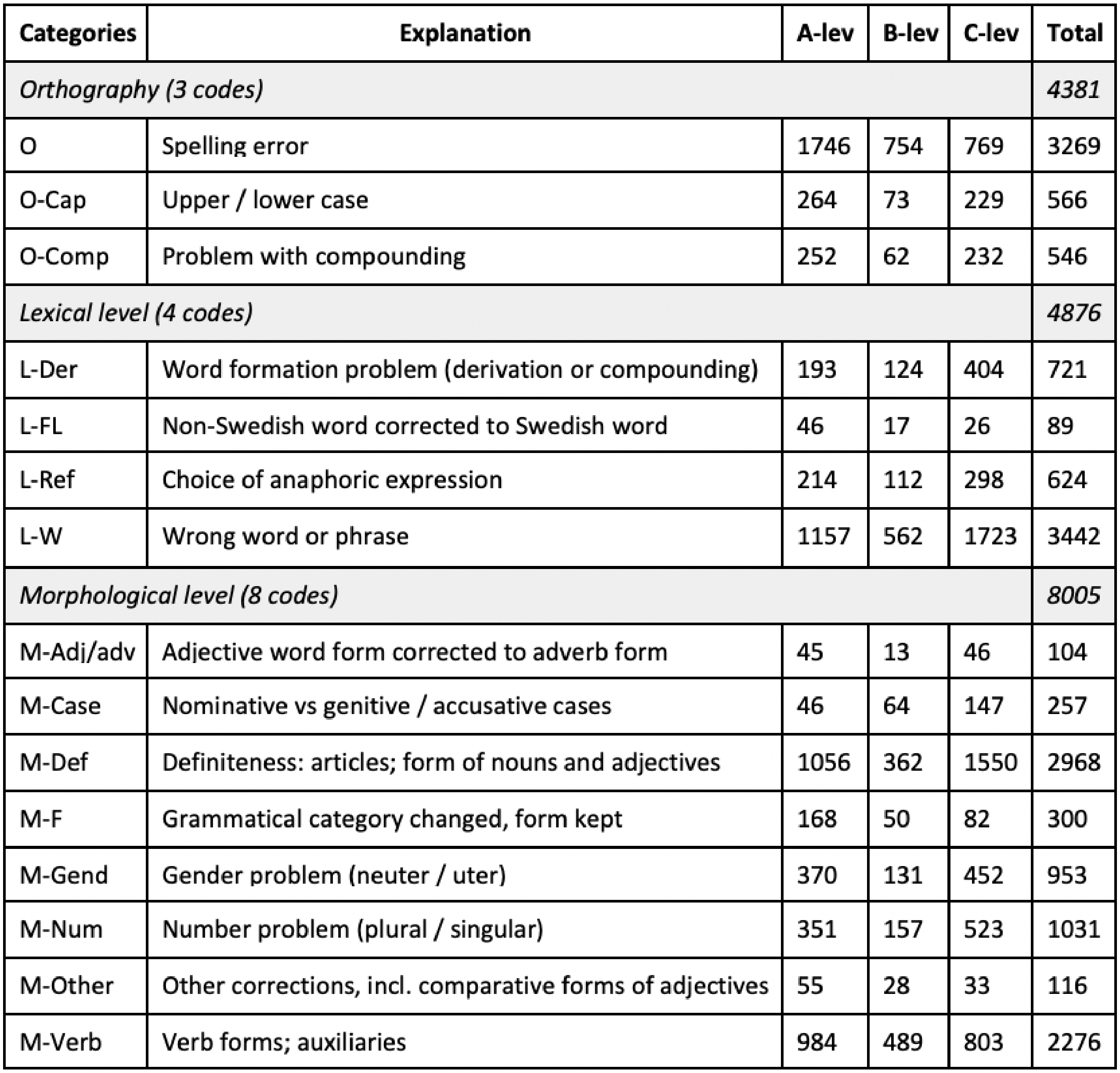
During years starting 2017-2021 we have been working on setting up the main building blocks for empirically based research on Swedish as a second language which we release under the name of the *SweLL infrastructure*. This work entailed collecting and manually annotating learner written essays, which 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:

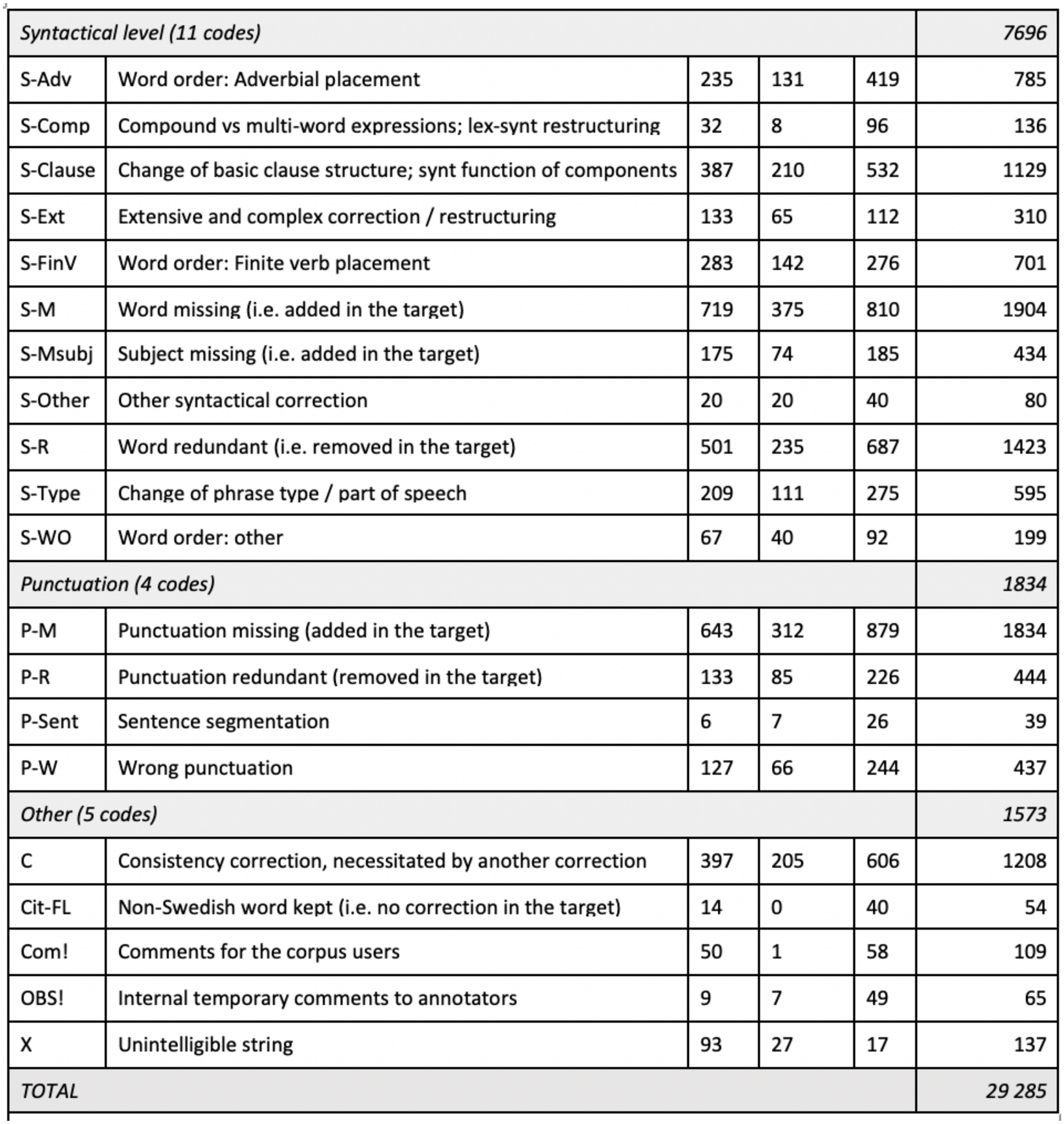
- Transcription guidelines
- Pseudonymization guidelines
- Normalization guidelines
- Correction annotation guidelines

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.

To give you a feel for the SweLL-gold corpus, have a look at the overview table with correction labels below.



Categories	Explanation	A-lev	B-lev	C-lev	Total
Orthography (3 codes)					*4381*
O	Spelling error	1746	754	769	3269
O-Cap	Upper / lower case	264	73	229	566
O-Comp	Problem with compounding	252	62	232	546
Lexical level (4 codes)					*4876*
L-Der	Word formation problem (derivation or compounding)	193	124	404	721
L-FL	Non-Swedish word corrected to Swedish word	46	17	26	89
L-Ref	Choice of anaphoric expression	214	112	298	624
L-W	Wrong word or phrase	1157	562	1723	3442
Morphological level (8 codes)					*8005*
M-Adj/adv	Adjective word form corrected to adverb form	45	13	46	104
M-Case	Nominative vs genitive / accusative cases	46	64	147	257
M-Def	Definiteness: articles; form of nouns and adjectives	1056	362	1550	2968
M-F	Grammatical category changed, form kept	168	50	82	300
M-Gend	Gender problem (neuter / uter)	370	131	452	953
M-Num	Number problem (plural / singular)	351	157	523	1031
M-Other	Other corrections, incl. comparative forms of adjectives	55	28	33	116
M-Verb	Verb forms; auxiliaries	984	489	803	2276



Syntactical level (11 codes)					*7696*
S-Adv	Word order: Adverbial placement	235	131	419	785
S-Comp	Compound vs multi-word expressions; lex-synt restructuring	32	8	96	136
S-Clause	Change of basic clause structure; synt function of components	387	210	532	1129
S-Ext	Extensive and complex correction / restructuring	133	65	112	310
S-FinV	Word order: Finite verb placement	283	142	276	701
S-M	Word missing (i.e. added in the target)	719	375	810	1904
S-Msubj	Subject missing (i.e. added in the target)	175	74	185	434
S-Other	Other syntactical correction	20	20	40	80
S-R	Word redundant (i.e. removed in the target)	501	235	687	1423
S-Type	Change of phrase type / part of speech	209	111	275	595
S-WO	Word order: other	67	40	92	199
Punctuation (4 codes)					*1834*
P-M	Punctuation missing (added in the target)	643	312	879	1834
P-R	Punctuation redundant (removed in the target)	133	85	226	444
P-Sent	Sentence segmentation	6	7	26	39
P-W	Wrong punctuation	127	66	244	437
Other (5 codes)					*1573*
C	Consistency correction, necessitated by another correction	397	205	606	1208
Cit-FL	Non-Swedish word kept (i.e. no correction in the target)	14	0	40	54
Com!	Comments for the corpus users	50	1	58	109
OBS!	Internal temporary comments to annotators	9	7	49	65
X	Unintelligible string	93	27	17	137
TOTAL					*29 285*

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

Correction annotation guidelines

Contents

1 Introduction

1.1 The purpose of the correction annotation

The purpose of the correction annotation is to make the learner corpus searchable for different types of deviations from a standard language norm. The annotation of the learner texts according to the correction taxonomy of the SweLL project, made with the annotation tool Svala, is hence an important step in making the texts in the corpus analyzable for research and educational purposes.

Back to the menu

1.2 The purposes of this document

This document has three interconnected purposes:

- It documents choices made by the SweLL project group regarding the principles of the correction annotation.
- It provides guidance for annotators on how to apply the SweLL annotation taxonomy in order to ensure the greatest possible inter-annotator agreement.
- It provides a description of the correction annotation for users of the Swell corpus.

Back to the menu

1.3 The organization of the document

The rest of this document contains the following six main sections:

2. **The most fundamental principles and terms**, explaining what a correction is, and the basic terms *correction*, *unit*, *link* and *tag*.
3. **The general structure of the taxonomy**, providing a description of the general structure of the taxonomy, short descriptions of each tag, indications of tag relationships requiring special attention, and a sketched algorithm for the choice of tag.
4. **General directions for the correction annotation**, explaining the basic practices for using the Svala annotation tool, and providing directions for the annotation which are more general than the usage of specific tags.
5. **Descriptions of each tag**, given in the order they appear in the Svala annotation tool.
6. **Discriminating between specific correction categories**, providing guidance for discriminating between specific pairs or small groups of tags.

7. **Other categorization issues**, including some categorization issues involving more complex relationships between several tags.

Back to the menu

2 The most fundamental principles and terms

2.1 What is a correction? What is annotated?

By *correction* we mean a difference between the original learner text and the normalized version of the text. The correction annotation is thus a categorization of such differences.

This means that the correction annotation only indirectly indicates properties of the original learner text. What is directly indicated is *the relationship between the original version and the normalized version of the text*. The correction annotation is thus highly dependent on the preceding normalization.

While the purpose of the correction annotation is to make the texts searchable for deviations from a standard language norm, such deviations are only possible to categorize on the basis of assumptions of the learner's intended content, along with judgements of the suitability of standard language expressions communicating that content. This means that *any* annotation of "deviations", "errors" etc. in learner texts is actually an annotation of a *relationship* between (a segment of) the analyzed text and an assumed standard language version. By the choice of the term *correction annotation* (rather than, for instance, *error annotation*) the SweLL project emphasizes these conditions of learner language analysis. The normalized text versions make explicit this necessary assumption about the specific standard version of the text to which the original text is related. (The principles underlying the normalization are described in the normalization guidelines.)

Back to the menu

2.2 Correction, unit, link and tag

A **correction** may consist of:

- an *addition* of a unit; the unit is only present in the normalized text
- a *deletion* of a unit; the unit is only present in the original text
- a *movement* of a unit; the unit is present both in the original text and the normalized text, but it is placed differently relative to the surrounding text
- a *change* of a unit; a unit in the normalized text is *a corrected version* of a corresponding unit in the original text

A **unit** normally consists of one token – *a token unit* – but it may also consist of a group of tokens – a *group unit* (see section 4.4.1).

The correction annotation is created in Svala by marking **links** between the original text and the normalized text. **Tags** are placed on to the links, representing the categories in the SweLL correction annotation taxonomy, which is presented and described in major parts of this document. A link may be tagged with one or several tags.

The links on which the tags are placed are visualized as lines in the Svala annotation tool. In the case of a *movement* or *change* correction, the link runs between a unit in the original text and the corresponding unit in the normalized text. In the case of a *deletion* correction, the link runs between the unit in the original text and “nothing” – i.e. an empty space between the tokens in the normalized text. In the case of an *addition* correction, the link runs between the unit in the normalized text and “nothing”, in a corresponding fashion. See the *Svala guidelines* for further descriptions of the practical implementation of the correction annotation in Svala.

Back to the menu

3 The general structure of the taxonomy

3.1 The main correction categories

The SweLL correction annotation taxonomy contains five main categories:

- **Orthographic corrections (O)**: “regular spelling corrections”, corrections between upper and lower case, and corrections of the use of spaces and hyphens between words.
- **Lexical corrections (L)**: 1) corrections of the choice of word and 2) corrections of the internal morphological structure of word stems – i.e. the formation of words through compounding and derivation.
- **Morphological corrections (M)**: corrections concerning *inflectional morphology*. This category also covers some extra-morphological corrections which are closely related to inflectional forms and grammatical categories involving inflections.
- **Punctuation corrections (P)**: corrections of the choice of punctuation marks as well as the adding or removal of punctuation marks, and also instances of merging or splitting sentences.
- **Syntactical corrections (S)**: corrections regarding the structure of multi-word phrases and clauses, including corrections of missing or redundant words, word order, the choice between a compound and a multi-word expression, and more complex syntactic corrections such as corrections between a finite clause structure and an infinitive phrase.

In addition to the correction tags included in these five main correction categories, the Svala tool provides six other tags, including tags for corrections made as a consequence of other corrections (**C**), corrections not covered by any of the defined correction categories (**Unid**), unintelligible strings (**X**), strings cited from a foreign language (**Cit-FL**), and finally two tags for notes and comments – **OBS!** for internal work notes, and **Com!** for comments intended for the corpus users.

Back to the menu

3.2 Overview of the taxonomy

The following table provides an overview of all tags.

The left-most column lists the tag names in the order in which they appear in the Svala annotation tool, which is also the order in which they are presented in section 5 (*Descriptions of each tag*). This means that the tags are primarily sorted according to the five major categories: The tags in the group of orthographic corrections (O-tags) come first, and the other four major categories in alphabetical order: lexical corrections (L-tags), morphological corrections (M-tags), punctuation corrections (P-tags), and syntactical corrections (S-tags). After these five major categories follows the group of miscellaneous other tags. Within each correction category, the tags are listed in alphabetical order.

The second column in the table provides a short description of the correction type.

In some cases, the tags are hierarchically ordered, so that a tag with lower priority is only applied in those instances when the tag with higher priority cannot be applied. This hierarchical ordering is indicated in the two columns *Higher priority than* and *Lower priority than*.

The discrimination between certain pairs or groups of tags requires special attention. Most of these discrimination issues are handled in section 6 (*Discriminating between specific correction categories*), but some are also dealt with in section 7 *Other categorization issues*. References to the relevant sections are provided in the column *Sections dealing with relevant discrimination issues*.

Tag name	**Short description of the correction type**	**Higher priority than**	**Lower priority than**	**Sections dealing with relevant discrimination issues**
O	spelling	L-FL	L-Der	O vs punctuation tags (6.1) Spaces, hyphens and dashes (O-Comp, P-M, P-R, P-W) (7.4)
O-Cap	upper/lower case			
O-Comp	spaces and hyphens between words	P-M P-R P-W		O-Comp vs S-Comp (6.2) Verbal particles and reflexives (O-Comp, S-Comp, L-Der, L-W, S-M, S-R) (7.3) Spaces, hyphens and dashes (O-Comp, P-M, P-R, P-W) (7.4)
L-Der	word formation (derivation and compounding)	O L-FL	M-Adj/adv	L-Der vs L-W (6.3) L-Der vs M-Verb (6.4) L-Der vs S-Comp (6.5) Verbal particles and reflexives (O-Comp, S-Comp, L-Der, L-W, S-M, S-R) (7.3)

L-FL	non-Swedish word corrected to Swedish word		O L-Der	Non-Swedish words and sequences (O, Cit-FL, L-Der, L-FL, L-W) (7.2)
L-Ref	choice of anaphoric expression	M-Gend M-Num L-W	M-Def	
L-W	wrong word or phrase		L-Ref	L-Der vs L-W (6.3) L-Der vs M-Verb (6.4) L-W vs M-Case (6.6) L-W vs M-Verb (6.7) L-W vs S-Comp (6.9) Non-Swedish words and sequences (O, Cit-FL, L-Der, L-FL, L-W) (7.2) Verbal particles and reflexives (O-Comp, S-Comp, L-Der, L-W, S-M, S-R) (7.3)
M-Adj/adv	adjective form of word corrected to adverb form	L-Der M-Gend		
M-Case	nominative vs genitive; nominative vs accusative			L-W vs M-Case (6.6)
M-Def	definiteness: articles; forms of nouns and adjectives	L-Ref S-M S-R		M-Def vs M-Num, S-R and S-M (6.10) C vs M-Def, M-Verb and M-Num (6.21)
M-F	grammatical category kept, form changed		M-Num	
M-Gend	gender		L-Ref M-Adj/adv	
M-Num	number	M-F	L-Ref	M-Def vs M-Num, S-R and S-M (6.10) C vs M-Def, M-Verb and M-Num (6.21)
M-Other	other morphological corrections, including change between different comparational forms of adjectives		All other M-categories	

M-Verb	verb forms; use of *ha, komma* and *skola* auxiliaries	S-M S-R		L-Der vs M-Verb (6.4) L-W vs M-Verb (6.7) C vs M-Def, M-Verb and M-Num (6.21) Clause corrected or created? (M-Verb, S-Clause, S-M, S-Msubj) (7.1)
P-M	punctuation missing (added)		O-Comp (spaces and hyphens)	O vs punctuation tags (6.1) Spaces, hyphens and dashes (O-Comp, P-M, P-R, P-W) (7.4)
P-R	punctuation redundant (removed)		O-Comp (spaces and hyphens)	O vs punctuation tags (6.1) Spaces, hyphens and dashes (O-Comp, P-M, P-R, P-W) (7.4)
P-Sent	sentence segmentation			
P-W	wrong punctuation		O-Comp (spaces and hyphens)	O vs punctuation tags (6.1) Spaces, hyphens and dashes (O-Comp, P-M, P-R, P-W) (7.4)
S-Adv	adverbial placement	S-FinV S-WO		S-Adv, S-FinV, S-WO: Exceptions to the default ranking of the word order tags (6.11) S-Clause vs word order tags (6.12)
S-Clause	change of basic clause structure: syntactic function of components, hierarchical clause structure			S-Clause vs word order tags (6.12) S-Clause vs S-Comp (6.13) S-Clause vs S-Ext (6.14) S-Clause vs S-M (6.15) S-Clause vs S-Msubj (6.16) S-Clause vs S-R (6.17) Clause corrected or created? (M-Verb, S-Clause, S-M, S-Msubj) (7.1)
S-Comp	compound vs multi-word expression, and other restructuring of the same lexical morphemes within a phrase			O-Comp vs S-Comp (6.2) L-Der vs S-Comp (6.5) L-W vs S-Comp (6.9) S-Clause vs S-Comp (6.13) Verbal particles and reflexives (O-Comp, S-Comp, L-Der, L-W, S-M, S-R) (7.3)

S-Ext	extensive and complex correction			S-Clause vs S-Ext (6.14) S-Ext vs S-M (6.15) S-Ext vs X (6.19)
S-FinV	finite verb placement	S-WO	S-Adv	S-Adv, S-FinV, S-WO: Exceptions to the default ranking of the word order tags (6.11) S-Clause vs word order tags (6.12)
S-M	word missing (added)		M-Def M-Verb S-Msubj	M-Def vs M-Num, S-R and S-M (6.10) S-Clause vs S-M (6.15) S-Ext vs S-M 6.18) Clause corrected or created? (M-Verb, S-Clause, S-M, S-Msubj) (7.1) Verbal particles and reflexives (O-Comp, S-Comp, L-Der, L-W, S-M, S-R) (7.3)
S-Msubj	subject missing (added)	S-M		S-Clause vs S-Msubj (6.16) Clause corrected or created? (M-Verb, S-Clause, S-M, S-Msubj) (7.1)
S-Other	other syntactical correction		All other S-categories	
S-R	word redundant (removed)		M-Def M-Verb	M-Def vs M-Num, S-R and S-M (6.10) S-Clause vs S-R (6.17) S-Type vs S-M and S-R (6.20) Verbal particles and reflexives (O-Comp, S-Comp, L-Der, L-W, S-M, S-R) (7.3)
S-Type	change of phrase type/part of speech			S-Type vs S-M and S-R (6.20)
S-WO	word order, other		S-Adv S-FinV	S-Adv, S-FinV, S-WO: Exceptions to the default ranking of the word order tags (6.11) S-Clause vs word order tags (6.12)
C	consistency correction, necessitated by other correction			

Cit-FL	non-Swedish word **kept**, i.e. **not corrected**			Non-Swedish words and sequences (O, Cit-FL, L-Der, L-FL, L-W) (7.2)
Com!	comments for the corpus user			
OBS!	internal and temporary comments for the annotators			
Unid	unidentified correction		All other categories	
X	unintelligible string			S-Ext vs X (6.19)

Back to the menu

3.3 Tag choice algorithm

While the correction types are primarily presented as belonging to the five major linguistically defined categories (orthographic, lexical, morphological, punctuation-related, and syntactical corrections), they may also be categorized based on the more straight-forward relationship between the original text and the normalized text: as *additions*, *deletions*, *movements* or *changes* of units (see section 2.2).

To some extent, the corrections within these categories may be hierarchically ordered, on different levels of priority, so that certain tags should only be considered in those cases when a tag with a higher priority is not applicable. In this section, such an algorithm for the choice of tag is sketched.

It must however be stressed that the algorithm presented below is a simplification of the actual decision flow, since in reality tags of different kinds also “compete” with each other to a certain extent; the same correction may for instance be considered both for an addition category and for a change category. Moreover, the limits between the categories are not as clear cut as a straight-forward application of the decision flow algorithm requires. A somewhat more complex picture of the relationships between the correction categories is presented in sections 3.2 (*Overview of the taxonomy*) and 5 (*Descriptions of each tag*), and more nuanced descriptions of how to discriminate between the categories are given in sections 6 (*Discriminating between specific correction categories*) and 7 (*Other categorization issues*).

With that said, here is the algorithm sketch:

When choosing a tag, the following four alternatives should first be considered (in the order listed):

1. Is the correction a follow-up correction of another correction? → C
2. Is the original unit unintelligible? → X
3. Is the original unit a non-Swedish word or phrase (which is not interpretable as a misspelled Swedish word or phrase or as a non-standard composition of Swedish morphemes)? → L-FL
4. Is the correction extensive, involving a greater than usual element of subjectivity or randomness, so that the syntactic structure of the normalized unit may rather be described as *created* than as *corrected*? → S-Ext

If none of the four cases above holds, proceed as follows:

5. Is the correction an *addition*? → See *Additions* (3.3.1)
6. Is the correction a *deletion*? → See *Deletions* (3.3.2)
7. Is the correction a *movement* (word order change)? → See *Movements (word order changes)* (3.3.3)
8. Is the correction a *change*?
    a. Is at least one of the units involved (the original or the normalized one) a multi-word unit? → See *Changes involving a multi-word unit in either the original version, the normalized version or both* (3.3.4.1)
    b. Is the correction a change between two *one-word units* (the original unit is not necessarily a *one-token* unit since it may also be a word which is incorrectly written as two tokens)?
        i. Does the change involve *inflectional morphology* or a stem changed which is due to a change of a grammatical quality which is usually expressed with inflectional morphology (such as *går* →*gick* or *mus* →*möss*)? → See *One-word unit changed to one-word unit, change of inflectional morphology or functionally parallel stem change* (3.3.4.2)
        ii. Is the *word stem* exchanged (except according to i) or morphologically restructured? → See *One-word unit changed to one-word unit, word stem exchanged or morphologically restructured* (3.3.4.3)
        iii. Is the *spelling*, including *the use of spaces and hyphens* and *the use of upper/lower case letters*, changed? → See *One-word unit changed to a one word unit, spelling (including uses of spaces and hyphens and token-internal punctuation)* (3.3.4.4)
    c. Is at least one of the units a punctuation mark? → See *Changes where at least one of the units is a punctuation mark* (3.3.4.5)

Back to the menu

3.3.1 Additions

1 **Priority one**
- The addition increases the number of clauses → **S-Clause**

2 **Priority two**; alternatives with equal priority
- Definite or indefinite article added → **M-Def**
- *Ha, skola* or *komma* auxiliary added → **M-Verb**
- The infinitive marker *att*, used in a construction with a *komma* auxiliary, added → **M-Verb**
- Punctuation mark added → **P-M**
- *Subjunktion* ‘subordinating conjunction’ or other *bisatsinledare* ‘subordinating connector’ added → **S-Clause**
- Subject added (into a clause which exists in the original version) → **S-Msubj**

3 **Otherwise → S-M**

3.3.2 Deletions

1 **Priority one**
- The deletion involves a reduction of the number of clauses → **S-Clause**

2 **Priority two**; alternatives with equal priority:
- Definite or indefinite article deleted → **M-Def**
- *ha, skola* or *komma* auxiliary deleted → **M-Verb**
- the infinitive marker *att* used in a construction involving a *ha, skola* or *komma* auxiliary deleted → **M-Verb**
- Punctuation mark deleted → **P-R**
- *Subjunktion* 'subordinating conjunction' or other *bisatsinledare* 'subordinating connector' deleted → **S-Clause**

3 **Otherwise → S-R**

3.3.3 Movements (word order changes)

1 **Priority one**
- The unit has been moved as part of a change between a main clause structure and a subordinate clause structure *which is also indicated by other means* (as the removal or the addition of a *subjunktion* 'subordinating conjunction') *or* the unit has been moved because its syntactic function has changed → **S-Clause**

2 **Priority two**
- Choose the tag which gives one rather than two word order changes.

3 **Priority three**
- The *fundament* 'pre-finite element' is exchanged for textual rather than clause-internal reasons → **S-WO**

4 **Priority four**
- Adverbial moved → **S-Adv**

5 **Priority five**
- Finite verb moved → **S-FinV**

6 **Otherwise → S-WO**

3.3.4 Changes

3.3.4.1 Changes involving a multi-word unit in either the original version, the normalized version or both

1 **Priority one**
- The change is limited to the mere addition of a space (or several spaces) *between the words in a multi-word-phrase* → **O-Comp**

2 **Priority two**
- The change involves a change of the primary syntactic function of a word/phrase *or* the number of clauses is changed → **S-Clause**

3 **Priority three**; alternatives with equal priority:

- At least one of the units (the original one or the normalized one) is (an attempt at) a word or a fixed phrase, and at least one lexical morpheme is changed → **L-W**
    - The syntactic category/part of speech is changed → The *additional* tag **S-Type** is applied
- The normalized unit is a morphosyntactic restructuring of the lexical morphemes which are present in the original unit → **S-Comp**

4 **Otherwise → S-Other**

3.3.4.2 One-word unit changed to one-word unit, change of inflectional morphology or functionally parallel stem change

1 **Priority one**; alternatives with equal priority:

- Change of an adjective, noun, article or pronoun
    - **Priority one**: Definiteness change → **M-Def**
    - **Priority two**: An anaphoric expression has been changed to suit its correlate and textual position → **L-Ref**
    - **Priority three**: An adjective is changed from a form which is not used as an adverb to a form which is used as an adverb → **M-Adj/adv**
    - **Priority four**; alternatives with equal priority:
        - Case change → **M-Case**
        - Gender change → **M-Gend**
        - Number change → **M-Num**
- Change of a verb form (which involves a change of function) → **M-Verb**

2 **Priority two**: A declension/conjugation form (typically a suffix) which is used to express a specific grammatical category (e.g. plural or present tense) has been corrected to a form belonging to another declension/conjugation expressing the same grammatical category → **M-F**

3 **Otherwise → M-Other**

3.3.4.3 One-word unit changed to one-word unit, word stem exchanged or morphologically restructured

1 **Priority one**; alternatives with equal priority:

- One *ha, skola* or *komma* auxiliary has been exchanged for another *ha, skola* or *komma* auxiliary → **M-Verb**
- An anaphoric expression has been changed to suit its correlate and textual position → **L-Ref**
- Change between an indefinite and a definite article → **M-Def**

2 **Priority two**; alternatives with equal priority:

- The root morpheme(s) of the category-defining part of the stem is kept → **L-Der**
    - The syntactic category/part of speech is changed → The *additional* tag **S-Type** is applied (*in addition to L-Der*)
- The root morpheme(s) of the category-defining part of the stem is changed → **L-W**
    - The syntactic category/part of speech is changed → The *additional* tag **S-Type** is applied (*in addition to L-W*)

3.3.4.4 *One-word unit changed to a one word unit, spelling (including uses of spaces and hyphens and token-internal punctuation)*

1 **Priority one**; alternatives with equal priority:

- The change is limited to the mere usage of spaces and hyphens *between the words in a compound* → **O-Comp**
- Regular spelling correction (letters changed) → **O**
- Addition or removal of token-internal punctuation → **O**
- Change between upper and lower case letters → **O-Cap**

3.3.4.5 *Changes where at least one of the units is a punctuation mark*

1 Alternatives with equal priority:

- Punctuation mark changed to another punctuation mark → **P-W**
- Change between a conjunction and a sentence-separating punctuation mark → **P-Sent**

Back to the menu

4 General directions for the correction annotation

This section provides directions for the annotation which are more general than the usage of specific tags.

4.1 Some errors and corrections are not tagged

4.1.1 Some errors are not tagged

A consequence of the fundamental principle of annotating *corrections*, understood as differences between the original text and *one specific* interpretation of this text, rather than “errors in general”, is that certain clear deviations from the norms of written standard Swedish in the original texts are left without annotations – because they are not deviations *in relation to the normalized text*. This occurs for instance when a misspelled word in the original text has been corrected to another word altogether. Such a correction will be annotated as an instance of a faulty choice of word (L-W), and since the word in the original text cannot be analyzed as a misspelling of the word in the normalized text, the spelling error will be left without annotation.

- Du kan göra manga **nognting** → Du kan göra många **saker**

  The misspelling of *någonting* is not tagged, since *nognting* is related to *saker* and not to *någonting*.

4.1.2 Only visible corrections are tagged

Another fundamental principle of the correction annotation is that only “superficially visible” corrections are tagged – i.e. corrections which are reflected by differences between the original string and the normalized string. Here are some examples of the consequences of this principle:

In the following example, *det* is inserted as an expletive subject into a clause (and tagged S-Msubj). This correction also means that the subject of the original clause (marked with italics) is changed to an *egentligt subjekt* ('object-positioned subject'). The latter change is however not reflected in any additional difference (i.e. on top of the addition of *det*) between the original string and the normalized string. The change of the subject of the original string to an *egentligt subjekt* in the normalized string is thus left without any tag.

- på andra sidan finns *människor som har så mycket pengar att de kan köpa halva världen* → Å andra sidan finns **det** *människor som har så mycket pengar att de kan köpa halva världen*

In the following example a passive construction is changed to an active construction (and tagged as S-Clause). Consequently, the subordinate clause (marked with italics) is changed from an *egentligt subjekt* to an object. This changed is left untagged.

- I texten **sägs det av henne** *att hon efterlyser de som enbart talar finska och är positiva till svenskan och tvåspråkigheten* . → I texten **säger hon** *att hon efterlyser de som enbart talar finska och är positiva till svenskan och tvåspråkigheten* .

Back to the menu

4.2 Positive assumption principle

The *positive assumption principle*, which is a crucial principle for the transcription of the essays, is a working principle also for the correction annotation. This means that, when in doubt about which of two tagging alternatives to choose, the alternative implying the "smallest correction" should be chosen. For instance, when it seems equally likely that a correction is due to a spelling mistake as to a word choice mistake, the O tag should be chosen rather than the L-W tag, since correcting the spelling of a word may be considered less invasive than exchanging the word altogether. Other examples are provided below.

- Många unga **arbetssökanden** → många unga **arbetssökande**

  The target form *arbetssökande* is a plural indefinite form. The original form may be interpreted either as a singular definite form or as an incorrect plural indefinite form, following the pattern for similar neuter nouns, such as *ett sökande, flera sökanden*. In accordance with the positive assumption principle, the assumption is here that the writer attempted at a plural indefinite form, and the correction is tagged M-F rather than M-Num and M-Def.

- Jag blir **förvånat** → **förvånad**

  The form *förvånat* is assumed to be an attempted participle form with the wrong gender rather than a supine form, and the correction is thus tagged M-Gend rather than L-Der.

4.2.1 Annotation of strings including unreadable signs in transcribed texts

Signs which are judged unreadable in handwritten originals are transcribed as "$". When possible (i.e. when the string as a whole is nevertheless interpretable) such signs are eliminated in the normalized version of the text. When annotating such corrections, the *positive assumption* principle should be applied. In other words:

If it is possible that the original string was correct, assume that it was correct and leave the correction (i.e. the difference between the original string including the "$" and the normalized string without the "$") without *any* tag representing a correction category. A *Com!* tag is however applied, with the Edge comment "Original assumed to be correct." Cf. *Com! (comments for the corpus users)*.

This holds in the following cases:

- The $ has simply been removed in the normalized version; it is possible that $ marks a character which was intended to be marked as removed in the original.
    - *ja$g* → *jag*
- The $ corresponds to exactly one letter in the normalized version; it is possible that $ marks the letter to which it has been changed in the normalized version.
    - *hi$$a* → *hitta*

Only if it is not possible that the original string was correct should the correction be annotated with a suitable tag.

Back to the menu

4.3 Several tags on the same link

In many instances, the same unit is corrected in multiple ways reflecting separate correction categories, without one correction being included in or implied by another (cf. 4.4.2 *Corrections of tokens included in group units*). In these cases, the link should be marked with one tag for each correction type.

- Both spelling and the use of a hyphen between two words in a compound is changed, the link is tagged with both O and O-Comp:
    - **energi-niveå** → **energinivå**
- Both spelling and definiteness of a noun is changed, the link is tagged with both O and M-Def:
    - Finns många nya **lagenheterna** i dyrare delar i huvudstaden → Det finns många nya **lägenheter** i dyrare delar i huvudstaden
- Both definiteness and number of a noun is changed, the link is tagged with M-Def and M-Num:
    - Alla sitter bakom sina **skärmens** sken → Alla sitter bakom sina **skärmars** sken
- Both case and placement of a pronoun is changed, the link is tagged with M-Case and M-WO:
    - Man kan inte vänta att lägga **de** upp på social medier → Man kan inte vänta med att lägga upp **dem** på sociala medier

There are also some instances when one single corrected element reflects two correction categories simultaneously. Also in such cases the link should be tagged with tags for the both correction types. For instance, in the following example, the correction from the infinitive phrase *köpa mat* to the compound noun *mataffär* is simultaneously a correction of the choice of word (L-W) and a correction of the phrase type (S-Type), and the link is tagged with both of these tags:

- Tycker om min plats har en **köpa mat** ett , litet centrum en förskolan en vårdcentralen → Jag tycker om platsen där jag bor , den har en **mataffär** , ett litet centrum , en förskola , en vårdcentral

Similarly, the following example is tagged with both L-Der (word formation correction) and S-Type (phrase type correction):

- Jag ska **solen** och bada → Jag ska **sola** och bada

In the following example the adjective *nationell* in the original text has a gender form (non-neuter) which is incongruent with the noun it modifies (*språk*, neuter), and its singular form does not match the co-ordinated subject (*svenska och finska*) with which the whole predicative NP (*nationell språk*) is associated.

- Sedan 1919 bliv både svenska och finska **nationell** språk i Finland → Sedan 1919 är både svenska och finska **nationella** språk i Finland

In cases like this, arguments could be made for the solution to tag this correction both as M-Num and M-Gend (since the correction eliminates both the gender error and the number error), but it could also be argued that such changes should only be marked with the M-Num tag (since the plural adjective is undifferentiated for gender). We have however decided for the former solution: Corrections of this kind are tagged both with M-Num and M-Gend.

Back to the menu

4.4 Group units

4.4.1 Grouping of tokens

Corrections are by default tagged token by token. Group units are only created in the following cases:

1. A *movement* correction concerns a phrase made up of several tokens.

Grouping of moved units may be made for all the word order correction categories, i.e. S-Adv, S-FinV and S-WO. Whenever a moved token is part of a multi-token phrase which is moved altogether, the whole phrase should be grouped. The only exception is when a single token included in such a moved phrase is also corrected, see *Corrections of tokens included in group units* (4.4.2) below.

Grouping of moved units is made also in the case of coordinated phrases being moved altogether:

- Jag kunde inte att komma för **jobbar och pluggar** jag tillsamman → Jag kunde inte komma för att jag både **jobbar och pluggar**

2. **A *change* correction concerns strings without a token-to-token correspondence between the original and the normalized version, i.e. at least one of the corresponding units (the original unit or the normalized unit) is made up by more than one token.**

Grouping of changed units is primarily made for the following correction types:

- O-Comp:
  - Jag kände mig **jätte konstig → jättekonstig**
- L-FL:
  - **gas “bojler” → gaskokare**
- L-W:
  - Jag maner att om vi **har önskemål** kan vi göra nånting utan pengar → Jag menar att om vi **vill** kan vi göra nånting utan pengar
  - Traditioner ger människorna tid att **stå still** och fundera över livet → … att **stanna upp** och fundera över livet
  - Konsekvenserna man skulle få ifall undervisning i svenska blev frivillig så skulle mer än hälften av finska befolkningen **avskaffa** svenskan som modersmålsundervisning → Konsekvenserna man skulle få ifall undervisning i svenska blev frivilligt är att då skulle mer än hälften av den finska befolkningen **välja bort** svenskan som modersmålsundervisning
- S-Clause
  - I texten **sägs det av henne** att hon efterlyser de som enbart talar finska → I texten **säger hon** att hon efterlyser de som enbart talar finska
  - I artikeln … skriver Catharina Söderbergh om tre personer som har sagt sin åsikt om tvångssvenskan och **hur svenskan befinner sig** → I artikeln … skriver Catharina Söderbergh om tre personer som har sagt sin åsikt om tvångssvenskan och **läget för svenskan**
  - **Att växa** upp som en flicka så var det väldigt många orättvisor man fick vänja sig vid. -→ **När man växte** upp som en flicka så var det väldigt många orättvisor man fick vänja sig vid.
- S-Comp
  - **det vardagliga livet → vardagslivet**
  - Enligt Hyltenstam så kan minoritetsspråk räddas om man **inblandar** dem äldre som kan språket → Enligt Hyltenstam så kan minoritetsspråk räddas om man **blandar in** de äldre som kan språket
  - Skillnader är stor av **Sveriges bostad** → Skillnaderna är stora mot **bostäder i Sverige**
- S-Ext
  - **Det därför tycker jag om det är** simma lungt och blåser → **Jag tycker om det för att jag tycker om** att simma lugnt och när det blåser

- Så hur mycket pojkarna bettalad **värt inte deras äppelträd** eftersom trädet var viktig för dem → Så hur mycket pojkarna betalade **var mindre än vad deras äppelträd var värt** eftersom trädet var viktigt för dem .

- S-Type (and L-W)
  - jag behover pengar f$r **liv** och betalning av min hus → jag behöver pengar för **att leva** och betala för mitt hus

Note: M-Def corrections and M-Verb corrections involving more than one token in one or both text versions are **not grouped**, but tagged token by token.

- Slutligen tror jag att sociala medier **blev** en essentiell del → Slutligen tror jag att sociala medier **har blivit** en essentiell del

  A link between *blev* and *blivit* is tagged with M-Verb, as well as a separate link between *har* and "nothing".

- Tyvärr **ska** jag inte komma på kursen → Tyvärr **kommer** jag inte **att** komma på kursen

  A link between *ska* and *kommer* is tagged with M-Verb, as well as a separate link between *att* and "nothing".

- När kommer **den buss** → När kommer **bussen**

  A link between *buss* and *bussen* is tagged M-Def, as well as a separate link between *den* and "nothing".

- Han har egen **rummet** och jag egen → Han har **ett** eget **rum** och jag **ett** eget

  A link between *rummet* and *rum* is tagged M-Def, as well as separate links between each instance of the added indefinite article *ett* and "nothing".

4.4.2 Corrections of tokens included in group units

Corrections of tokens which are included in a corrected group unit may be of three kinds, which are handled in different ways:

1 The correction of the included token is part of the correction which motivated the forming of the group unit.

→ The token correction is not tagged separately, but is considered included in the tag placed on the link from the group unit.

- **min plats → platsen där jag bor**

All words involved in this correction (*min plats* and *platsen där jag bor*) are grouped together, and the correction as a whole is tagged S-Clause. Accordingly, *min* is not tagged as S-R, *jag* is not tagged as S-Msubj, *där* and *bor* are not tagged as S-M, and the change from *plats* to *platsen* is not tagged as M-Def.

- **Att växa** upp som en flicka så var det väldigt många orättvisor man fick vänja sig vid → **När man växte** upp som en flicka så var det väldigt många orättvisor man fick vänja sig vid

Here, the change of structure from an infinitive phrase to a finite clause is analysed as an instance of S-Clause, and this tag is placed on the link between the group unit *Att växa* in the original text and the group unit *När man växte* in the normalized text. *Man* is thus not tagged as S-Msubj, the change from *växa* to *växte* is not tagged as M-Verb, etc.

- Finns många nya lagenheterna i dyrare delar i huvudstaden , men detta är **lång distans från räker nummer** . → Det finns många nya lägenheter i dyrare delar i huvudstaden , men detta är **långt ifrån tillräckligt**

This correction is analysed as an instance of L-W (wrong word or phrase). While the adverb *långt* in the normalized text could be seen as corresponding to the adjective *lång* in the original text, this correction is included in the exchange of the whole phrase. These two tokens (*lång* and *långt*) are thus not aligned separately, and the correction is not tagged as a separate correction.

2 The correction of the included token is not part of the correction which motivated the forming of the group unit.

a The correction of the token may still be considered a correction of the whole group unit, since the token is the head of the grouped phrase.

→ The additional correction is marked by placing an additional tag on the group unit, which is not broken up.

- Skillnader är stor av **Sveriges bostad** → Skillnaderna är stora mot **bostäder i Sverige**

The forming of the group unit is motivated by the analysis of the change from *Sveriges bostad* to *bostäder i Sverige* as an instance of S-Comp, and that tag is placed on the connecting link. The change from singular to plural of *bostad* affects the number of the whole phrase, and although this correction would not in itself motivate the forming of a group unit, the M-Num tag is also placed on the group link.

b The correction of the token specifically concerns the individual token – not the whole group unit

→ The group unit cannot be visualized as a group, but needs to be broken up.

The individually corrected token is tagged with the tag which concerns that token specifically.

The part of the broken group unit which contains the longest string or the phrase head is tagged with the tag which concerns the correction of the whole group unit.

The other part(s) of the broken group unit (including the individually corrected token) are marked with C.

- Hon försätter skriva om **att inte avskaffa den obligatoriska svenskan på skolan** är fördel för ungdomar → Hon fortsätter skriva om att **det** är en fördel för ungdomar **att inte avskaffa den obligatoriska svenskan i skolan**

The infinitive phrase *att inte avskaffa den obligatoriska svenskan på* (→ *i*) *skolan* is moved as a group unit. However, the change from *på* to *i* calls for breaking up the group in the Svala visualization. Thus, *att inte avskaffa den obligatoriska svenskan* is visually grouped together and the link is tagged with S-WO. The link between *på* and *i* is marked with L-W, but also with C, to indicate that the movement of this word is part of the movement of the already tagged part of the phrase. The link from *skolan* to *skolan* is exclusively tagged with C.

- Jag hoppas att du **intressant för din ny livs** → Jag hoppas att du **tycker att ditt nya liv är intressant**

tycker att in the normalization forms a unit which is tagged with S-Ext (the link runs between this normalized unit and "nothing"); *din* is linked with *ditt* and tagged M-Gend and C; *ny* is linked with *nya* and tagged M-Def and C; *livs* is linked with *liv* and tagged M-Case and C; *intressant för* is linked with *är intressant* and tagged C.

Back to the menu

4.5 Document comments

The *Document comment* field provides an opportunity to comment on deviations from the standard norm regarding text properties which cannot be adequately reflected by tags on individual links. The field may also be used for any kind of general comment on the text which the annotator regards as essential for the future corpus user.

This field may for instance be used when the verb tense choices in the text are inconsistent at a global text level, but when corrections of individual verb forms have generally not been made, since there is consistency more locally in the text.

Back to the menu

5 Descriptions of each tag

In the following the annotation categories and their tags will be presented in the order in which they appear in the Svala annotation tool – i.e. the O tags first, followed by the four other main correction categories in alphabetic order (L, M, P, S), and, finally, the remaining tags under the heading *Other tags*.

5.1 O – Orthographic corrections

The O tags represent the orthographic correction category. It includes three sub-categories.

5.1.1 O (regular spelling correction)

Tag name	Short description of the correction type	Higher priority than	Lower priority than	Sections dealing with relevant discrimination issues
O	spelling	L-FL	L-Der	O vs punctuation tags (6.1) Spaces, hyphens and dashes (O-Comp, P-M, P-R, P-W) (7.4)

The plain O tag is used for regular spelling corrections, i.e. when the string of letters is different in the original text and the normalized text, due to a spelling mistake.

- Det **tåg** 6 timmar från Teheran till Göteborg → **tog**
- Det finns **monga** affärer → **många**

The O tag is also used for instances when token-internal punctuation marks have been removed or added. (See *6.1 O vs punctuation tags*.)

- **t.v.** → **tv**

  (The intended use of *t.v.* here is the noun *teve*, not the phrase *tills vidare*.)

- **vdar** → **vd:ar**

5.1.2 O-Cap (upper/lower case)

Tag name	Short description of the correction type	Higher priority than	Lower priority than	Sections dealing with relevant discrimination issues
O-Cap	upper/lower case			

The O-Cap tag is used for corrections regarding the choice between upper and lower case letters.

- Det fanns en affär och vi gick dit. **vi** köpte flera saker → Det fanns en affär och vi gick dit. **Vi** köpte flera saker.
- På **Måndag** är det den 3 **Mars**. → På **måndag** är det den 3 **mars**.

5.1.3 O-Comp (spaces and hyphens between words)

Tag name	**Short description of the correction type**	**Higher priority than**	**Lower priority than**	**Sections dealing with relevant discrimination issues**
O-Comp	spaces and hyphens between words	P-M P-R P-W		O-Comp vs S-Comp (6.2) Verbal particles and reflexives (O-Comp, S-Comp, L-Der, L-W, S-M, S-R) (7.3) Spaces, hyphens and dashes (O-Comp, P-M, P-R, P-W) (7.4)

The O-Comp tag is used for corrections which involve the removal of a space between two words which have been interpreted as making up a compound in the normalized text version:

- Jag kände mig **jätte konstig** → **jättekonstig**
- Det ligger ett **kultur hus** nära min bostad → **kulturhus**

The tag is also used for the adding of a space between two words:

- för två **årsen** → för två **år sen**
- **andravärldskriget** → **andra världskriget**

And it is used for corrections regarding the use of hyphens in compounds:

- **24 åringar** → **24-åringar**
- **buss-stationer** → **busstationer**
- **tv program** → **tv-program**
- **fantasy och äventyrsroman** → **fantasy- och äventyrsroman**
- han har jämförts **" tvångsvenska " debatten** i Finland med **nynorska debatten** i Norge → han har jämfört **"tvångssvenska"-debatten** i Finland med **nynorska-debatten** i Norge

In the following example, the change from a comma to a long dash between *Finland* and *Sverige* is tagged with P-W, and the change from a space to a hyphen between *Sverige* and *historien* is tagged with O-Comp:

- Författaren börjar med att beskriva det långa **Finland , Sverige historien** → Författaren börjar med att beskriva den långa **Finland–Sverige-historien**

Note (two tags): Corrections which involve both a removal of a space and a change of the form of the first part of the compound, e.g. through the adding or the removal of a *foge*-s, are tagged both with O-Comp and L-Der:

- **Kommunikation förändring → Kommunikationsförändring**
- **sports människor → sportmänniskor**

Note (discrimination): The O-Comp tag should only be used for corrections concerning the mere orthographic rendering, with or without a space or a hyphen, of a compound or a multi-word expression, and not for corrections which are rather to be interpreted as involving an actual alternation between a compound and a multi-word expression. The latter case is covered by the S-Comp tag. (See *6.2 O-Comp vs S-Comp*.)

Back to the menu

5.2 L – Lexical corrections

The L tags represent the lexical correction category. It includes four sub-categories.

5.2.1 L-Der (word formation)

Tag name	Short description of the correction type	Higher priority than	Lower priority than	Sections dealing with relevant discrimination issues
L-Der	word formation (derivation and compounding)	O L-FL	M-Adj/adv	L-Der vs L-W (6.3) L-Der vs M-Verb (6.4) L-Der vs S-Comp (6.5) Verbal particles and reflexives (O-Comp, S-Comp, L-Der, L-W, S-M, S-R) (7.3)

The L-Der tag represents the correction category *deviant word formation*. It is used for corrections of the internal morphological structure of word stems, both with regard to compounding and to derivation.

The L-Der tag is exclusively used for links between *one-word units* (not necessarily one-token units, since a word may mistakenly be written as two tokens). The normalized word has kept at least one root morpheme from the original word, **and** another morpheme has been removed, added, exchanged or had its form altered.

- A derivational affix has been removed, added or corrected:
  - **förstöra → störa**
  - **ändring → förändring**

    - De är **stressiga** på grund av studier → De är **stressade** på grund av studier

- A verbal particle of a compound form of a particle verb has been removed, added, or corrected (Cf. 7.3 *Verbal particles and reflexives (O-Comp, S-Comp, L-Der, L-W, S-M, S-R)*.):

    - att debatten i båda länder är jätte lik varandra , bara hantering **avskiljer** → att debatterna i de båda länderna är jättelika varandra , bara hanteringen **skiljer**

    - tre olika **ställningar** som finns om svenska inom Finland → tre olika **inställningar** som finns till svenska i Finland

- A verbal particle has been exchanged for a derivational affix:

    - Internet **uppmanar** vår förståelse → Internet **utmanar** vår förståelse

        - **efterlyser → belyser**

- The form of the first part of a compound has been corrected (for instance by adding or removing a "foge-s"):

    - **tvångsvenska → tvångssvenska**

    - **politiksområede → politikområde**

    - **grottaflicka → grottflicka**

    - **solenkräm → solkräm**

- The first word of the compound has been exchanged, while the second, category-defining word, has been kept (cf. *6.3 L-Der vs L-W*):

    - **dagsskolan → förskolan**

Note (two tags): When the correction of a word tagged with L-Der involves a change of phrase type or part of speech, the correction is tagged with S-Type, in addition to L-Der.

- **norska** bokmål → **norskt** bokmål

- jag talar **kurdisk** → jag talar **kurdiska**

- en **nybyggnad** lägenhet → en **nybyggd** lägenhet

- Jag ska **solen** och bada → Jag ska **sola** och bada

- Jag behöver pengar för liv och **betalning** av min hus → Jag behöver pengar för att leva och **betala** mitt hus

- Eller det skapar kontakter och **trivs**? → Eller skapar de kontakter och **trivsel**?

- något ha en " svart " avställning **båda** i jobbet och från bostad → Någon har en " svart " inkomst **både** från jobbet och från bostaden

Note (priority): Corrections of an adjective form to an adverb form of the same word should be tagged with the specific tag M-Adj/adv rather than with L-Der.

Note (discrimination): Participle forms of verbs are treated as derivations rather than as inflections. Corrections between participle forms and supine forms are thus tagged L-Der (and S-Type) rather than M-Verb. (See 6.4 *L-Der vs M-Verb*.)

- vi har precis **flyttad** till Norrby → vi har precis **flyttat** till Norrby

Note (discrimination): Corrections of verbal particles which are part of phrasal forms of particle verbs are not tagged with the L-Der tag, but with the L-W tag. (See 7.3 *Verbal particles and reflexives (O-Comp, S-Comp, L-Der, L-W, S-M, S-R)*.)

5.2.2 L-FL (foreign word corrected to Swedish)

Tag name	Short description of the correction type	Higher priority than	Lower priority than	Sections dealing with relevant discrimination issues
L-FL	non-Swedish word corrected to Swedish word		O L-Der	Non-Swedish words and sequences (O, Cit-FL, L-Der, L-FL, L-W) (7.2)

The L-FL tag is used for words from a foreign (non-Swedish) language which have been corrected to a Swedish word. It may also be applied to words which have certain non-Swedish traits due to influence from a foreign language.

- Det fanns flera rum, två kök, **balkony** och trädgård → **balkong**
- Jag och min **family** → **familj**
- som beror på historiska **event** → vilket beror på historiska **händelser**
- Bostad i D-hemland är litet het **topik** för att **diskussera** → Bostäder i D-hemland är ett lite hett **ämne** att **diskutera**

The L-FL tag is used for corrections with the following characteristics:

- The word (or word-string) in the original text is recognizable as a word from a foreign language *to the annotator* (who is obviously not equally proficient in all languages). Alternatively, *the annotator* recognizes certain non-Swedish traits in the word which are due to influence from a foreign language.
- The word is not recognizable as a Swedish word. Alternatively: The word-string is not recognizable as a string of Swedish words. (See 7.2 Non-Swedish words and sequences (O, Cit-FL, L-Der, L-FL, L-W).)
- The word in the original text may or may not be correctly formed and used according to the norms of the assumed influencing language. For instance, the form *ticker*, corrected to *biljetter*, is tagged L-FL, on the basis of the assumption that it is the English word *tickets* which has led to the form *ticker*.

5.2.3 L-Ref (anaphoric expressions)

Tag name	Short description of the correction type	Higher priority than	Lower priority than	Sections dealing with relevant discrimination issues
L-Ref	choice of anaphoric expression	M-Gend M-Num L-W	M-Def	

The L-Ref tag is used for anaphoric expressions (particularly pronouns and pronominal adverbs) which have been corrected in order to have the grammatical form (gender, number, reflexive/non-reflexive), semantic content (masculine/feminine, directional/locational etc.), and specificity which suits its correlate and its textual position.

- Jag har en bror, **hon** heter xx → **han** heter xx
- Det var lätt att förstå **alla** → Det var lätt att förstå **allt**
- Innan trädet blev borta dem fick fars äpplen och **sin** fru lagade saft och mos av äpplena → Innan trädet var borta fick de fars äpplen och **hans** fru lagade saft och mos av äpplena
- Oavsett hur än pappa anstränger **honom**… → Oavsett hur mycket pappan än anstränger **sig** …
- Du måste röra på **sig** → Du måste röra på **dig**
- Östling skriver om politiker som tycker att obligatorisk språkundervisningen och prov på nynorska krävs för att försäkra ett minoritets språk som har ett gamal kulturarv knutet till **den** . → Östling skriver om politiker som tycker att obligatorisk språkundervisning och prov på nynorska krävs för att skydda ett minoritetsspråk som har ett gammalt kulturarv knutet till **sig** .

In cases like the following example, the L-Ref tag has higher priority than the M-Num tag:

- Stämmer det att våra sociala media skapar individualism och ensamhet? Eller skapar **det** kontakter och trivs? → Eller skapar **de** kontakter …

In cases like the following examples the L-Ref tag has higher priority than the M-Gend tag:

- Jag tycker om Gotland och **den** ligger i Sverige → Jag tycker om Gotland , och **det** ligger i Sverige
- “ Finns det något som är mer värt än pengar ? ” Jag vet faktiskt inte , **den** är en svår fråga . → “ Finns det något som är mer värt än pengar ? ” Jag vet faktiskt inte , **det** är en svår fråga .
- Om man jobbar åtta timmar varje vardag för att få pengar , sedan är det lite hycklande att säga något som , ” Ja , det finns jo viktigare saker i livet ! ” **Den** är en tredjedel av din dag ! → Om man jobbar åtta timmar varje vardag för att få pengar , är det sedan lite hycklande att säga något som : ” Ja , det finns ju viktigare saker i livet ! ” **Det** är en tredjedel av din dag !

In cases like the following examples the L-Ref tag has higher priority than the L-W tag:

- I artikeln … Cecilia Christner Raid om varför man ens lär sig svenska idag i Finland **som** beror på historiska event . → I artikeln … skriver Cecilia Christner Riad om varför man ens lär sig svenska idag i Finland , **vilket** beror på historiska händelser .
- Jag ska resa **där** ikväll → Jag ska resa **dit** ikväll
- efter den händelsen **som** Anna blev skadad av Elsas magiska kraft → efter den händelsen **då** Anna blev skadad av Elsas magiska kraft
- Haga, **var** jag bor idag → Haga, **där** jag bor idag

The L-Ref tag may also be used when a noun which is used anaphorically has been exchanged for a pronoun, or the other way around, in order for the specificity of the anaphoric expression to suit its textual position:

- **Han** blir bortgift mot viljan med sin kusin i pappas land … → **Killen** blir bortgift mot sin vilja med sin kusin i pappans land …
- Heter Jag Karin och Jag har änmält mig till danskursen och bitalad för det men tyvär fick Jag inte tid för att koma . Jag kunde inte att komma för jobbar och pl$ggar Jag tillsamman och **det** passar inte med min tiden → Jag heter Karin och jag har anmält mig till danskursen och betalat för den , men tyvärr hade jag inte tid att komma . Jag kunde inte komma för att jag både jobbar och pluggar , och **kursen** passar inte med mina tider

Note (priority): Changes between indefinite and definite forms of nouns are tagged M-Def rather than L-Ref, even in cases when the change is due to textual circumstances for the use of the noun as an anaphor.

Note (discrimination): The L-Ref tag is only used for *corrections* of anaphoric expressions, not for additions of them. Added pronouns (or other anaphoric expressions), like *den* in the example below, are tagged S-M.

- ta emot → ta emot **den**

5.2.4 L-W (wrong word or phrase)

Tag name	Short description of the correction type	Higher priority than	Lower priority than	Sections dealing with relevant discrimination issues
L-W	wrong word or phrase		L-Ref	L-Der vs L-W (6.3) L-W vs M-Case (6.6) L-W vs M-Verb (6.7) L-W vs S-Comp (6.9) Non-Swedish words and sequences (O, Cit-FL, L-Der, L-FL, L-W) (7.2) Verbal particles and reflexives (O-Comp, S-Comp, L-Der, L-W, S-M, S-R) (7.3)

The L-W tag represents the correction category *wrong word or phrase*. It is used when a word or phrase in the original text has been replaced by another word or phrase in the normalized version. At least one of the units (the original unit or the normalized unit) is a word or a fixed phrase (the original unit may be *an attempt at* a word or a fixed phrase).

The L-W tag is placed on units which are *exchanged* rather than *corrected*. This principle underlies many of the more specific principles for discriminating between the L-W tag and other tags, referred to in the table segment above. It is also an underlying principle for some decisions about whether to group or not to group the words in a corrected/exchanged phrase. See the note below on grouping.

If both the original unit and the normalized unit are one-word units, the root morpheme(s) of the category-defining part of the stem is exchanged (see *6.3 L-Der vs L-W* for further explanations and examples). This typically means one of the following:

- A word is completely exchanged (no common morpheme).
- The second word of a compound is exchanged.

If at least one of the strings is a multi-word unit, at least one lexical morpheme has to be exchanged (cf. *6.9 L-W vs S-Comp*).

One word replaced by one word:

- Uppfinningen som **transformerade** hela kommunikationsområdet → Uppfinningen som **förändrade** hela kommunikationsområdet
- På det sättet kan kommunen **motionera** alla medborgare att träna → På det sättet kan kommunen **motivera** alla medborgare att träna

- vi solade och badade mycket och träffäde några **personer** från flera länder → Vi solade och badade mycket och träffade **människor** från flera länder

- Alla blir **busiga** med sina sociala medier → Alla blir **upptagna** med sina sociala medier

  (See 7.2 Non-Swedish words and sequences (O, Cit-FL, L-Der, L-FL, L-W) for further comments on this example.)

- Jag bor i D-område **på** en lägenhet → Jag bor i D-område **i** en lägenhet

One word replaced by a multi-word expression:

- Med detta som bakgrund så kan man konstatera att majoriteten av **Finland** på olika missnöjesnivåer avskyr inlärning av svenska på grund- och gymnasieskolor → Med detta som bakgrund så kan man konstatera att majoriteten av **Finlands befolkning** , på olika missnöjesnivåer , avskyr inlärning av svenska på grund- och gymnasieskolor

Multi-word expression replaced by one word:

- Jag maner att om vi **har önskemål** kan vi göra nånting utan pengar → Jag menar att om vi **vill** kan vi göra nånting utan pengar

Multi-word expression replaced by another multi-word expression:

- Finns många nya lagenheterna i dyrare delar i huvudstaden , men detta är **lång distans från räker nummer** → Det finns många nya lägenheter i dyrare delar i huvudstaden , men detta är **långt ifrån tillräckligt**

  (*This correction is also tagged as S-type, see below.*)

Phrasal verb replaced by another phrasal verb, both verb and particle exchanged:

- Traditioner ger människorna tid att **stå still** och fundera över livet → … att **stanna upp** och fundera över livet

Phrasal verb replaced by another phrasal verb, different verb but the same particle; only the verb is tagged, not the particle:

- men på debattsidor debatterar de oftast språkfrågan eftersom dem inte **hittar** på något annat att diskutera om → men på debattsidor debatterar de oftast språkfrågan eftersom de inte **kommer** på något annat att diskutera om

Verbal particle replaced by another verbal particle, but verb kept; only the verbal particle is tagged, not the verb:

- Han torkade **bort** bordet → Han torkade **av** bordet

Compound particle verb replaced by a phrasal verb, both verb and particle replaced:

- så skulle mer än hälften av finska befolkningen **avskaffa** svenskan som modersmålsundervisning → då skulle mer än hälften av den finska befolkningen **välja bort** svenskan som modersmålsundervisning

Phrasal verb replaced by a simple verb:

- Men jag bestämde mig själv för att **hänga ihop** båda att studera och jobba → Men jag bestämde mig själv för att **kombinera** att både studera och jobba

Note (grouping): The L-W tag is used for corrections when one unit has been exchanged for another unit; if a unit involved (the original unit, the normalized unit, or both) consists of more than one word, the words should be grouped to form a group unit.

When identifying the units the general condition that the L-W tag is placed on units which are exchanged rather than corrected should be kept in mind. This means that when a phrase has been changed in such a way that the lexical word(s) is kept but a grammatical word is exchanged, the phrase is considered *corrected* while the grammatical word is exchanged, and the L-W tag should be placed only on the exchanged grammatical word, not on the whole phrase. The phrase should accordingly not be treated as a group unit in these cases, even if the whole phrase (in one of its versions) may be considered a fixed phrase.

- såsom våra ”moderna” språk är ibland en blandning av flera utrotade minoritetsspråk som **under** tiden skapade något nytt och unik → såsom våra ”moderna” språk ibland är en blandning av flera utrotade minoritetsspråk som **med** tiden skapade något nytt och unikt

  Both *under tiden* and *med tiden* are fixed phrases, but *under tiden* is (according to the principle of positive assumption) considered as an attempt at the fixed phrase *med tiden*. The phrase is thus *corrected* through the *exchange* of the preposition, and the expressions are therefore not treated as group units; the exchanged unit is the preposition (*under* has been replaced with *med*), and the L-W tag is thus placed only on the preposition.

- man väljer att använda engelska stället **av** svenska → man väljer att använda engelska i stället **för** svenska

  Here, *stället av* in the original text is considered an attempt at the fixed phrase *i stället för*. This phrase is *corrected* by the *exchange* of the preposition *av* for the preposition *for*, and with the addition of the preposition *i*. Since the phrase as a whole is corrected rather than exchanged, the phrase is not grouped to form a group unit. The L-W tag is placed on the exchanged preposition, i.e. on a link between *av* and *för*. The added *i* is tagged S-M.

Note (two tags): When a correction tagged with L-W involves a change of phrase type/part of speech, the correction is also marked with the additional tag S-Type.

- plötsligt bröt muren **sinsemellan** → plötsligt rasade muren **mellan dem**

- jag behover pengar f$r **liv** → jag behöver pengar för **att leva**

- Tycker om min plats har en **köpa mat** ett , litet centrum en förskolan en vårdcentralen → Jag tycker om platsen där jag bor , den har en **mataffär** , ett litet centrum , en förskola , en vårdcentral

- Man kan promonera **lång tid** finns det blåser → Man kan promenera **länge** när det blåser

- **Några tider** vi kan titta och lyssna på hur funkar det → **Ibland** kan vi titta och lyssna på hur det funkar
- jag bor **in** lägenhet plan ett → Jag bor **i** en lägenhet på plan ett

Back to the menu

5.3 M – Morphological corrections

The M tags represent the category *morphological corrections*. It covers corrections related to inflections. This includes primarily corrections of individual inflectional forms, but in some cases also corrections of more complex grammatical constructions closely related to inflectional forms. The latter concerns basic definiteness constructions (see 5.3.3) periphrastic comparative and superlative adjective constructions (see 5.3.4 and 5.3.7), and tense-related verbal constructions involving auxiliaries (see 5.3.8).

The category of morphological corrections includes eight sub-categories.

5.3.1 M-Adj/adv (adjective corrected to adverb)

Tag name	**Short description of the correction type**	**Higher priority than**	**Lower priority than**	**Sections dealing with relevant discrimination issues**
M-Adj/adv	adjective form of word corrected to adverb form	L-Der M-Gend		

The M-Adj/adv tag is used for corrections of an adjective to its t-form, when the t-form is called for due to the adjective being used as an adverb.

- Användaren mår **dålig** → **dåligt**
- Hur påverkar detta våra månden **långsiktig?** → **långsiktigt**
- Jag tror att man måste hitta balansen mellan rikdom och andra saker som är **riktiga** viktiga → Jag tror att man måste hitta balansen mellan rikedom och andra saker som är **riktigt** viktiga

The M-Adj/adv is also used for similar changes, when an adjective or adjective-like word is changed to a morphologically closely related but distinct adverb form:

- När jag kommer **första** i 2012 , jag bodde i en social lagenhet med 3 andra person → När jag **först** kom 2012 bodde jag i en kommunal lägenhet med 3 andra personer

Moreover, the M-Adj/adv tag is used when an adjectival form of the word *liten*/*litet* is changed to the adverb form of the same word, i.e. *lite*. Although the form *litet* is occasionally used as an adverb in standard Swedish, it is too archaic to be used in most of the Swell text genres, and adverbial uses of the form *litet* are thus normally corrected to the form *lite* during normalization. Such changes are also tagged with M-Adj/adv:

- Bostad i D-hemland är **litet** het topik för att diskussera → Bostäder i D-hemland är ett **lite** hett ämne att diskutera

Note (not two tags): Since the change from adjective to adverb is already indicated by the tag M-Adj/adv itself, the tag S-Type is not needed in these cases.

5.3.2 M-Case

Tag name	Short description of the correction type	Higher priority than	Lower priority than	Sections dealing with relevant discrimination issues
M-Case	nominative vs genitive; nominative vs accusative			L-W vs M-Case (6.6)

The M-Case tag is used for corrections regarding the choice of case form for nouns (nominative vs genitive) and pronouns (nominative vs accusative).

- 50 **kilometer** avstånd → **kilometers**
- Som kan bidra till **samhället** utveckling → **samhällets**
- Det ger **man** en positiv energi → Det ger **en** en positiv energi
- Folk hinner inte prata med **de** → Folk hinner inte prata med **dem**
- **Dem** ska kunna kommunicera → **De** ska kunna kommunicera
- Ingen förstår **hon** → Ingen förstår **henne**

Note (discrimination): When the form *dem* is changed to the form *de* used as a definite article, the correction is tagged as L-W and S-Type, not as M-Case (see *6.6 L-W vs M-Case*).

5.3.3 M-Def (definiteness)

Tag name	Short description of the correction type	Higher priority than	Lower priority than	Sections dealing with relevant discrimination issues
M-Def	definiteness: articles; forms of nouns and adjectives	L-Ref S-M S-R		M-Def vs M-Num, S-R and S-M (6.10) C vs M-Def, M-Verb and M-Num (6.21)

The M-Def tag is used for corrections regarding definiteness constructions. The kinds of corrections involved are:

- Change between indefinite and definite forms of nouns (*bok* vs *boken*; *rum* vs *rummet*)
- Change between, removal of and addition of indefinite and definite articles (*en*, *ett*; *den*, *det*, *de*)

- Change between indefinite and definite forms of adjectives, participles and adjective-like pronouns (*stor/stort* vs *stora*; *skriven/skrivna* vs *skrivna*; *egen/eget* vs *egna*)

Examples:

- Vi gick till McDonalds och åt **maten** → **mat**
- Jag trivs bättre på **jobb** här → **jobbet**
- Jag har läst ditt **mejlet** → ditt **mejl**
- De **gränserna** som fanns mellan kvinnor och män har nästan försvunnit → De **gränser** som fanns
- När kommer **den buss** → När kommer **bussen**
- Då kommer svenska språket försvinna ur Finska samhället → Då kommer svenska språket försvinna från **det** finska samhället
- Han har egen **rummet** och jag egen → Han har **ett** eget **rum** och jag **ett** eget
- min **stor** dag → min **stora** dag
- den så **kallad** tvångssvenskan → den så **kallade** tvångssvenskan

Note (grouping, consequence change): M-Def corrections involving more than one token in one or both text versions are tagged token by token, with a separate M-Def tag on each token (no C-tag). (See *4.4.1 Grouping of tokens* and *6.21 C vs M-Def, M-Verb and M-Num*.)

Note (category extension): The M-Def tag is used for removals and additions of articles, even when the correction is due to a change between the use of a noun as a countable vs a non-countable noun.

- Jag tycker att du kan baka kakor för dina arbetskamrater och bjuda dem för **en** fika → Jag tycker att du kan baka kakor till dina arbetskamrater och bjuda dem på fika

Note (discrimination): When the addition or removal of a singular indefinite article (*en* or *ett*) is due to a change of number rather than a change of definiteness, the article should be tagged M-Num rather than M-Def, S-R or S-M. (See 6.10 *M-Def vs M-Num, S-R and S-M*.)

5.3.4 M-F (wrong form, correct grammatical category)

Tag name	Short description of the correction type	Higher priority than	Lower priority than	Sections dealing with relevant discrimination issues
M-F	grammatical category kept, form changed		M-Num	

The M-F tag is used when a declension/conjugation form (typically a suffix) which is used to express a specific grammatical category (e.g. plural or present tense) has been corrected to a form belonging to another declension/conjugation expressing the same grammatical category.

The tag is used for the following correction types:

Nouns:

- from one to another plural-associated form (suffix or stem changed form): **båter** → **båtar**, **huvudproblemer** → **huvudproblem, basketlager** → **basketlag, sätter** → **sätt, musar** → **möss**

Verbs:

- from one to another form associated with present tense: jag **lager** mat → jag **lagar** mat
- from one to another form associated with past tense: **böjade** → **böjde**, **sjungde** → **sjöng**
- from one to another supine-associated form: **kunnit** → **kunnat**, **drickit** → **druckit**

Adjectives:

- from one to another comparative-associated form: **högare** → **högre**, **långare** → **längre**
- from one to another superlative-associated form: **högast** → **högst**, **ungast** → **yngst**
- from/to a periphrastic to/from an inflectional comparative or superlative construction: I Sverige är **mer kallt** än i xx → I Sverige är det **kallare** än i xx

Pronouns:

- Enligt Hermanssons artikel debatten om huruvida nynorska skulle behållas eller om **dens** öde skulle vara upp till folket och marknaden att bestämma över resulterar i blandade åsikter → Enligt Hermanssons artikel resulterar debatten om huruvida nynorskan ska bevaras eller om **dess** öde ska vara upp till folket och marknaden att bestämma över i blandade åsikter
- Jag bo i ett lägenhet med min sambo och **våras** yng dotter → Jag bor i en lägenhet med min sambo och **vår** lilla dotter

Note (priority): Unsuffixed noun forms will be interpreted as singular when corrected to a suffixed plural form, although unsuffixed plurals exist. Corrections like *område* → *områden* will thus be tagged with the M-Num tag and not with the M-F tag.

5.3.5 M-Gend (gender)

Tag name	**Short description of the correction type**	**Higher priority than**	**Lower priority than**	**Sections dealing with relevant discrimination issues**
M-Gend	gender		L-Ref M-Adj/adv	

The M-Gend tag is used to mark corrections of gender forms (neuter vs non-neuter) of nouns, articles, adjectives, and pronouns.

- Det är **en** mycket **trevlig** område → Det är **ett** mycket **trevligt** område
- **Köken** är **stor** → **Köket** är **stort**
- **Den** viktigaste är att … → **Det** viktigaste är att …
- **Den** första telefonnät → **Det** första telefonnätet
- Kommunikationen är **möjligt** → Kommunikationen är **möjlig**
- ifall undervisning i svenska blev **frivillig** → **frivilligt**
- grönländska blev **marginaliserad** → grönländska blev **marginaliserat**
- Han har **egen** rummet och jag **egen** → Han har ett **eget** rum och jag ett **eget** (cf. the same example in section 5.3.3)

Note (category extension): The M-Gend tag is only used for corrections between neuter and non-neuter forms. Corrections from the distinctly masculine form of adjectives -*e* to the general/feminine form -*a* are tagged M-Other rather than M-Gend. (See *5.3.7 M-Other*.)

Note (priority): Changes from the non-neuter form to the neuter form of adjectives which are due to the adjective being used as an adverb should be tagged M-Adj/adv rather than M-Gend.

Note (priority): Gender corrections of pronouns which are due to their anaphoric reference are covered by the L-Ref tag.

5.3.6 M-Num (number)

Tag name	Short description of the correction type	Higher priority than	Lower priority than	Sections dealing with relevant discrimination issues
M-Num	number	M-F	L-Ref	M-Def vs M-Num, S-R and S-M (6.10) C vs M-Def, M-Verb and M-Num (6.21)

The M-Num tag is used to mark number corrections of nouns, articles, adjectives, participles, and pronouns with adjective-like functions.

- Stress kan komma i många **form** → Stress kan komma i många **former**
- Alla sitter bakom sina **skärmens** sken → Alla sitter bakom sina **skärmars** sken
- Så fort bilder är **tagen** → Så fort bilderna är **tagna**
- De blir **offrat** → De blir **offrade**

Note (two tags): When an adjective or adjective-like attribute in an NP is changed from singular to plural and the gender of the original singular adjective is not congruent with the main word, the changed adjective should be tagged with both the M-Num tag and the M-Gend tag, since the correction eliminates both the number error and the gender error (cf. 4.3 Several tags on the same link):

- Sedan 1919 bliv både svenska och finska **nationell** språk i Finland → Sedan 1919 är både svenska och finska **nationella** språk i Finland

Note (consistency change): When several words in the same NP, including the main word, are changed with regard to number, the tag M-Num should only be placed on the main word. The other words are tagged with C, since these corrections are a consequence of the number change of the main word. (See *6.21 C vs M-Def, M-Verb and M-Num*.)

- därför kommer det att leda till **ett negativt konsekvens** → därför kommer det att leda till **negativa konsekvenser**

  The number change from *konsekvens* to *konsekvenser* is tagged M-Num. The change from *negativt* to *negativa* is tagged with C (as consistency change) and with M-Gend (since the gender error of the original *negativt* is also eliminated through this change, cf. the note above). The removal of the article *ett* is, for the same reasons, also tagged with both C and M-Gend.

Note (priority): Number corrections of pronouns which are due to their anaphoric reference are covered by the L-Ref tag.

Note (discrimination): When the addition or removal of a singular indefinite article (*en* or *ett*) is due to a change of number rather than a change of definiteness, the article should be tagged M-Num rather than M-Def, S-R or S-M. (See 6.10 *M-Def vs M-Num, S-R and S-M*.)

5.3.7 M-Other

Tag name	**Short description of the correction type**	**Higher priority than**	**Lower priority than**	**Sections dealing with relevant discrimination issues**
M-Other	other morphological corrections, including change between different comparational forms of adjectives		All other M-categories	

The M-Other tag is used for corrections involving inflectional morphology for which none of the other M tags are suited. This includes corrections between the comparational forms of adjectives, including corrections between non-morphologically related words functioning as different comparational forms of the same adjective (e.g. *dålig* and *sämre* or *många* and *fler*):

- Morfar ville visa de att det finns nånting som är **viktigaste** → Morfar ville visa dem att det finns nånting som är **viktigare**
- det finns **många** nackdelar än fördelar med att avskaffa den obligatorisk svenskan i Finland → det finns **fler** nackdelar än fördelar med att avskaffa den obligatoriska svenskan i Finland

The M-Other tag is also used for corrections from the distinctly masculine form *-e* of adjectives (and participles and adjective-like pronouns) to the general/feminine form *-a*:

- **förre** året anmälde jag till en kurs för simna → **förra** året anmälde jag mig till en simkurs
- min **käre** mamma → min **kära** mamma

  (Since the masculine form *-e* is never obligatory, corrections from the feminine/common form *-a* to the masculine form are not made during the normalization, and thus do never occur in the correction annotation process.)

Examples of other uses of the M-Other tag:

- flera **folker** → mycket **folk**

  The *folker* form is clearly an attempt at creating a plural form with the plural suffix *-er*, but 1) the correct usage of *folk* here is a singular usage, and 2) the plural form of *folk* is not *folker* but *folk*. This means that neither the M-Num tag may be used (since *folker* is not a plural form of *folk*), nor the M-F tag (since the correct usage of *folk* here is a singular usage).

5.3.8 M-Verb

Tag name	**Short description of the correction type**	**Higher priority than**	**Lower priority than**	**Sections dealing with relevant discrimination issues**
M-Verb	verb forms; use of *ha, komma* and *skola* auxiliaries	S-M S-R		L-Der vs M-Verb (6.4) L-W vs M-Verb (6.7) Clause corrected or created? (M-Verb, S-Clause, S-M, S-Msubj) (7.1)

The M-Verb tag covers corrections regarding inflectional verb forms and basic tense-related constructions involving *ha, skola* and *komma* auxiliaries, i.e.:

- Changes between the following verb forms: infinitive (*dansa*, *äta*), supine (*dansat*, *ätit*), present (*dansar*, *äter*), past (*dansade*, *åt*), imperative (*dansa*, *ät*), and s-form variants of these forms (*dansas, ätas* etc.).
- Additions and removals of forms of *ha*, *skola* and *komma*, when these verbs are used as auxiliaries.
- Changes from one *ha, skola* or *komma* auxiliary to another.
- Additions or removals of *att* which are due to changes of tense-related constructions involving auxiliary uses of the verbs *ha*, *skola* and *komma*.

Examples

- Jag **trivs** där för att jag hade fler vänner → Jag **trivdes** där för att jag hade fler vänner
- Jag tycker om **bor** tillsammans → Jag tycker om **att bo** tillsammans

- Världen **utvecklar** i teknologin i varje minut → Världen **utvecklas** i teknologin i varje minut
- Slutligen tror jag att sociala medier **blev** en essentiell del → Slutligen tror jag att sociala medier **har blivit** en essentiell del
- Det betyder inte att vi glomma bort vår tradition → Det betyder inte att vi **ska** glömma bort vår tradition
- Om ni behöver något mer för att jag **kan** få pengarna tillbaka bara skriv till mig → Om ni behöver något mer för att jag **ska kunna** få pengarna tillbaka, bara skriv till mig!
- Tyvärr **ska** jag inte komma på kursen → Tyvärr **kommer** jag inte **att** komma på kursen
- Min mormor har **kommat att bli** jättesjuk → Min mormor har **blivit** jättesjuk

Note (grouping, consistency): M-Verb corrections involving more than one token in one or both text versions are tagged token by token (no C tag and no grouping). (See *4.4.1 Grouping of tokens* and *6.21 C vs M-Def, M-Verb and M-Num*.)

Note (discrimination): Participle forms of verbs are treated as derivations rather than as inflections. Corrections between participle forms and verb forms (e.g. supine) forms are thus tagged L-Der (and S-Type) rather than M-Verb. (See 6.4 L-Der vs M-Verb.)

- vi har precis **flyttad** till Norrby → vi har precis **flyttat** till Norrby
- Lillasyster blir **skada** → Lillasyster blir **skadad**
- Jag är **imponerar** av henne → Jag är **imponerad** av henne

Back to the menu

5.4 P – Punctuation corrections

The P tags represent the category of *punctuation corrections*, including instances of merging or splitting sentences. It has four sub-categories.

5.4.1 P-M (missing punctuation)

Tag name	Short description of the correction type	Higher priority than	Lower priority than	Sections dealing with relevant discrimination issues
P-M	punctuation missing (added)		O-Comp (spaces and hyphens)	O vs punctuation tags (6.1) Spaces, hyphens and dashes (O-Comp, P-M, P-R, P-W) (7.4)

The P-M tag is used for corrections involving the addition of a punctuation mark.

- Jag har fyra **barn två** pojkar och två flickor → Jag har fyra **barn, två** pojkar och två flickor

- År 1972 togs den sista manuella växeln ur **bruk anger** Björkkvist → År 1972 togs den sista manuella växeln ur **bruk, anger** Björkkvist
- Det ger en positiv **energi därmed** kan man bli av med stress → Det ger en positiv **energi. Därmed** kan man bli av med stress
- Efter fem dagar kom till mig och **sa vad** vill du mig och **jag sa att jag** älskar **dig** → Efter fem dagar kom hon till mig och **sa: Vad** vill du **mig? Och** jag **sa: Jag** älskar **dig.**
- Min mamma lagar så god **mat jag** trivs med hennes mat → Min mamma lagar så god **mat. Jag** trivs med hennes mat

5.4.2 P-R (redundant punctuation)

Tag name	Short description of the correction type	Higher priority than	Lower priority than	Sections dealing with relevant discrimination issues
P-R	punctuation redundant (removed)		O-Comp (spaces and hyphens)	O vs punctuation tags (6.1) Spaces, hyphens and dashes (O-Comp, P-M, P-R, P-W) (7.4)

The P-R tag is used for corrections involving the removal of a punctuation mark.

- Jag minns inte **exakt . vad** det var med det var gott → Jag minns inte **exakt vad** det var men det var gott

5.4.3 P-Sent (sentence segmentation)

Tag name	Short description of the correction type	Higher priority than	Lower priority than	Sections dealing with relevant discrimination issues
P-Sent	sentence segmentation			

The P-Sent tag is used for corrections involving splitting a sentence or merging two sentences into one, when this correction involves more than the pure insertion or removal of a punctuation mark – in the typical case the adding or removal of a conjunction.

- Jag heter xxx och jag pratar arabiska och engelska och lite **svenska och vi** har tre rum i huset → Jag heter xxx och jag pratar arabiska och engelska och lite **svenska . Vi** har tre rum i huset

  In this example, the P-Sent tag is placed on a link between *och* in the original text and the period in the normalized text. The link between *vi* and *Vi* is tagged with C, as a consistency correction.

5.4.4 P-W (wrong punctuation)

Tag name	Short description of the correction type	Higher priority than	Lower priority than	Sections dealing with relevant discrimination issues
P-W	wrong punctuation		O-Comp (spaces and hyphens)	O vs punctuation tags (6.1) Spaces, hyphens and dashes (O-Comp, P-M, P-R, P-W) (7.4)

The P-W tag is used when a punctuation mark (other than a hyphen or a dash) in the original text has been replaced with another punctuation mark in the normalized text.

- Mina barn heter **Peter; Maria; Elisabeth** och John → Mina barn heter **Peter, Maria, Elisabeth** och John
- Efteråt blir drottningen (mamman) rasande och då prinsessan kan inte stå ut med henne **längre, hon** ger iväg … → Efteråt blir drottningen (mamman) rasande och då kan prinsessan inte stå ut med henne **längre. Hon** ger sig iväg …
- **5.5** procent → **5,5** procent

Back to the menu

5.5 S – Syntactical corrections

The S tags represent the syntactical correction category. It contains eleven sub-categories.

5.5.1 S-Adv (adverbial placement)

Tag name	Short description of the correction type	Higher priority than	Lower priority than	Sections dealing with relevant discrimination issues
S-Adv	adverbial placement	S-FinV S-WO		S-Adv, S-FinV, S-WO: Exceptions to the default ranking of the word order tags (6.11) S-Clause vs word order tags (6.12)

The S-Adv tag is used for corrections involving the placement of an adverbial.

- Jag **ofta** är vaken länge på kvällarna → Jag är **ofta** vaken länge på kvällarna
- Å ena sidan **idag** har → Å ena sidan har **idag**

- Jag är jättetrött eftersom jag sover **inte** på nätterna → Jag är jättetrött eftersom jag **inte** sover på nätterna
- som i sin tur har **tydligt** påverkat → som i sin tur **tydligt** har påverkat
- och **med hela världen** dela sina idéer och tankar → och dela sina idéer och tankar **med hela världen**
- om hur **under 600 år** Finland var en del av Sverige → om hur Finland var en del av Sverige **under 600 år**
- Jag tycker om **jättemycket** → Jag tycker **jättemycket** om

5.5.2 S-Clause (basic clause structure)

Tag name	**Short description of the correction type**	**Higher priority than**	**Lower priority than**	**Sections dealing with relevant discrimination issues**
S-Clause	change of basic clause structure: syntactic function of components, hierarchical clause structure			S-Clause vs word order tags (6.12) S-Clause vs S-Comp (6.13) S-Clause vs S-Ext (6.14) S-Clause vs S-M (6.15) S-Clause vs S-Msubj (6.16) S-Clause vs S-R (6.17) Clause corrected or created? (M-Verb, S-Clause, S-M, S-Msubj) (7.1)

The S-Clause tag is used for corrections involving changes of the most basic clause structure. The corrections in this category may be divided into the three following main types, concerning 1) type of clause, 2) number of clauses, and 3) internal clause structure.

1 The basic clause type (main clause, subordinate clause or main clause with question structure) is changed or specified:

   - A main clause is changed to a subordinate clause, for instance by adding a *subjunktion* 'subordinating conjunction':
      - Det ligger utanför Centum och en fördet är att det finns biblitoteket och det finns många skolor → Det ligger utanför Centum och en fördel är att det finns ett bibliotek och **att** det finns många skolor
      - Konsekvenserna man skulle få ifall undervisning i svenska blev frivillig **så** skulle mer än hälften av finska befolkningen avskaffa svenskan som modersmålsundervisning och istället fokusera på finska då dem sällan använder svenskan → Konsekvenserna man skulle få ifall undervisning i svenska blev

frivilligt är **att då** skulle mer än hälften av den finska befolkningen välja bort svenskan som modersmålsundervisning och istället fokusera på finska då de sällan använder svenskan

The added *subjunktion* 'subordinating conjunction' *att* indicates that the clause has been changed to a subordinate clause, and *att* is thus tagged S-Clause, while the change of *så* to *då* is tagged C, as a consequence correction. (The added finite verb *är* is also tagged S-Clause, since this addition involves an addition of a clause in the clausal hierarchy. See the same example under 2, below.)

- A subordinate clause is changed to a main clause, for instance by removing a *subjunktion* 'subordinate conjunction':
  - **Om** du skapar något för att inte känner dig ensam → Du kan göra något för att inte känna dig ensam (Cf. *6.17 S-Clause vs S-R*.)
- An "ordinary" main clause is changed to a main clause with question structure:
  - Jag vill att vet vad är problemet som jag inte kan gå till kursen **det finns** ingen plats eller **ni har** stoppade den här kurs och ni ska inte har det igen → Jag vill veta vad det är för problem som gör att jag inte kan gå på kursen : **Finns det** ingen plats eller **har ni** stoppat den här kursen och ska inte ha den igen ?

    In this case, the changed placement of the finite verb relative to the subject is marked with an S-Clause tag on the finite verb instead of an S-FinV tag.
- The relationship between two clauses is specified, for instance by the addition of a *subjunktion* 'subordinating conjunction' or another *bisatsinledare* ('subordinating connector'):
  - Sådana känslor gör användaren mår dåligt → Sådana känslor gör **att** användaren mår dåligt
  - Man kan promonera lång tid finns det blåser → Man kan promenera länge **när** det blåser

2 The structure of a phrase or a clause is changed in a way which involves a change of the number of clauses involved, for instance:

- The structure of a noun phrase is changed by changing a *framförställt attribut* ('preceding attribute') to an *efterställt attribut* ('succeeding attribute') in the form of a relative clause:
  - **min plats → platsen där jag bor** (Cf. 7.1 *Clause corrected or created? (M-Verb, S-Clause, S-M, S-Msubj)*.)
- An attribute in the form of a relative clause is changed to an attribute in the form of a *particip* 'participle':

    - Människor använder pengar eller nånting **som är lika till dem** från alltid → Människor har alltid använt pengar eller nånting **liknande**

- A finite clause is changed to a noun phrase:

    - tre personer som har sagt sin åsikt om tvångssvenskan och **hur svenskan befinner sig** → tre personer som har sagt sin åsikt om tvångssvenskan och **läget för svenskan**

- A finite clause (or a clear attempt at a finite clause, as in the following example) is changed to an infinitive phrase:

    - Men **om man** åka båt bätre för att båten är jatte stor → Men **att** åka båt är bättre för att båten är jättestor

- An infinitive phrase is changed to a finite clause:

    - **Att växa** upp som en flicka så var det väldigt många orättvisor man fick vänja sig vid. → **När man växte** upp som en flicka så var det väldigt många orättvisor man fick vänja sig vid (See 7.1 *Clause corrected or created? (M-Verb, S-Clause, S-M, S-Msubj)*.)

    - Riad pratar om den minskande procent av svensktalare i Finland och hur det leds till **några att tycka** att svenska ska inte vara ett officiellt språk längre → Riad pratar om den minskande procenten svensktalande i Finland och hur detta leder till **att några tycker** att svenska inte ska vara ett officiellt språk längre

- A clause is created by adding (at least) a finite verb to a "potential clause" which has no verb at all (finite or infinite) in the original text. (Cf. *7.1 Clause corrected or created?*)

    - Hon fördelen med att behålla den obligatoriska svenskan →
      Hon **beskriver** fördelen med att behålla den obligatoriska svenskan

    - Konsekvenserna man skulle få ifall undervisning i svenska blev frivillig så skulle mer än hälften av finska befolkningen avskaffa svenskan som modersmålsundervisning och istället fokusera på finska då dem sällan använder svenskan → Konsekvenserna man skulle få ifall undervisning i svenska blev frivilligt **är** att då skulle mer än hälften av den finska befolkningen välja bort svenskan som modersmålsundervisning och istället fokusera på finska då de sällan använder svenskan

      (See also the same example above, under 1.)

  When the creation of the clause involves the addition of a subject, on top of the addition of a finite verb, the added subject is also tagged with S-Clause (cf. 7.1 *Clause corrected or created? (M-Verb, S-Clause, S-M, S-Msubj)*.)

    - Jag studerar på eftermiddag Sfi och förmiddag praktik → Jag studerar sfi på eftermiddagen och på förmiddagen **har jag** praktik

        - Dem som tycker att avskaffa den obligatoriska svenskan … → De som tycker att **man ska** avskaffa den obligatoriska svenskan…

3 The structure of a clause is changed in a way which involves changing the primary syntactic function (subject, finite verb, object, *egentligt subjekt* 'object-positioned subject' and predicative) of one or more of the words/phrases involved, for instance:

- Subject changed to *egentligt subjekt* 'object-positioned subject':

    - Däremot vet ungarna att arbetslivet är annörlunda och **att vara engagerad** är nödvändigt → Däremot vet ungarna att arbetslivet är annorlunda och att **det** är nödvändigt **att vara engagerad**

      The inserted expletive subject *det* is tagged S-Msubj. The movement of the phrase *att vara engagerad* indicates its change of function from subject to *egentligt subjekt* 'object-positioned subject', and the link between this unit in the original text and the same unit in the normalized text is thus tagged S-Clause. (Cf. 6.16 *S-Clause vs S-Msubj* and 6.11 *S-Clause vs word order tags*.)

- Changes between a passive construction and an active construction:

    - I texten **sägs det av henne** att hon efterlyser de som enbart talar finska → I texten **säger hon** att hon efterlyser de som enbart talar finska

      A link is created between the group unit *sägs det av henne* and the group unit *säger hon*. This link is tagged S-Clause, and this categorization covers all of the corrections involved, none of which should be tagged separately (cf. *4.4.2 Corrections of tokens included in group units*).

- Subject or object changed to a prepositional complement in an adverbial:

    - **du** blir bättre → det blir bättre **för dig**

      A link is created between *du* and *för dig*. This link is tagged S-Clause. The inserted *det* subject is tagged S-Msubj. (Cf. *6.16 S-Clause vs S-Msubj*.)

    - I Norge har politiker laggt ner stora summar av pengar för att försvara språket och håller **det** fast som ett obligatorisk språk för att hålla det levande bland ungdomar → I Norge har politiker lagt ner stora summor pengar för att försvara språket och håller fast **vid det** som ett obligatoriskt språk för att hålla det levande bland ungdomar

- Object predicative changed to an attribute of an object NP:

    - Hon skriver också om att ha svenska **som obligatorisk** i Finland visar att alla är ju finnar och att " finlandssvenska " är bara påhittat → Hon skriver också om att det att ha **obligatorisk** svenska i Finland visar att alla är ju finnar och att " finlandssvenska " är bara påhittat

5.5.3 S-Comp (compound vs multi-word expression)

Tag name	**Short description of the correction type**	**Higher priority than**	**Lower priority than**	**Sections dealing with relevant discrimination issues**
S-Comp	compound vs multi-word expression, and other restructuring of the same lexical morphemes within a phrase			O-Comp vs S-Comp (6.2) L-Der vs S-Comp (6.5) L-W vs S-Comp (6.9) S-Clause vs S-Comp (6.13) Verbal particles and reflexives (O-Comp, S-Comp, L-Der, L-W, S-M, S-R) (7.3)

The S-Comp tag is used for corrections involving the restructuring of the same lexical words within a multi-word expression. It may also be used for changes between a multi-word expression and a one-word expression, when the one-word expression contains a derivational morpheme with the same semantic function as one of the words in the multi-word expression.

Corrections regarding the choice between a compound and a multi-word expression:

- **det vardagliga livet → vardagslivet**
- **livsskillnaden → skillnaden i liv**
- **svenska undervisningar → svenskundervisning**
- **avsnittet av texten → textavsnittet**
- **mamma hund → hundmamman**
- **litteraturverket → litterära verket**
- Om det händer att det finns planhalvor så kan det bero på blyghet, osäkerhet, rädsla eller **socialfobi → social fobi**
- Hejdlös **sociala medier användning** orsakar ensamhet → Hejdlös **användning av sociala medier** orsakar ensamhet

  (See *6.2 O-Comp vs S-Comp* for further discussion of the two examples above with *social*.)

Corrections regarding the choice between a compound and a phrasal structure for the combination of a verb and a verbal particle:

- Enligt Hyltenstam så kan minoritetsspråk räddas om man **inblandar** dem äldre som kan språket → Enligt Hyltenstam så kan minoritetsspråk räddas om man **blandar in** de äldre som kan språket

Other corrections involving the restructuring of the same lexical morphemes within a phrase:

- Skillnader är stor av **Sveriges bostad** → Skillnaderna är stora mot **bostäder i Sverige**
- Hon var **yngre än jag i ett år** → Hon var **ett år yngre än jag**
- Cecila Christner skriver om hur det svenska språket i skolor blir kallad tvångsvenska **i samma tid** svenska har blivit icke populärt i Finland → Cecilia Christner skriver om hur det svenska språket i skolorna blir kallat tvångssvenska **samtidigt** som svenska har blivit impopulärt i Finland
- **dit och hit → hit och dit**

Changes between a multi-word expression and a one-word expression, when the one-word expression contains a derivational morpheme with the same semantic function as one of the words in the multi-word expression:

- svenska har blivit **icke populärt** i Finland → svenska har blivit **impopulärt** i Finland

5.5.4 S-Ext (extensive correction)

Tag name	**Short description of the correction type**	**Higher priority than**	**Lower priority than**	**Sections dealing with relevant discrimination issues**
S-Ext	extensive and complex correction			S-Clause vs S-Ext (6.14) S-Ext vs S-M (6.15) S-Ext vs X (6.19)

The S-Ext tag is used for extensive, often complex, corrections. The syntactic structure of the normalized text segment may rather be described as *created* than as *corrected*, and the correction often also involves additions or removals of lexical words. The original text gives a fair indication of an intended meaning (otherwise the correction would be X-marked), but in most S-Ext cases it gives a poor basis for assuming a specific syntactic goal structure and/or a specific wording. In some cases, the wording and the syntactic goal structure may be fairly self-evident, but the S-Ext tag is chosen simply because the correction is too extensive or too complex to be covered by the other S-tags in any meaningful way.

While all normalizations are to be considered interpretations of the original texts, the S-Ext tag indicates that the range of possible normalizations is especially wide, and that the particular normalization chosen thus involves a greater element of subjectivity or randomness.

- **Det därför tycker jag om det är** simma lungt och blåser → **Jag tycker om det för att jag tycker om att** simma lugnt och när det blåser
- Så hur mycket pojkarna bettalad **värt inte deras äppelträd** eftersom trädet var viktig för dem -> Så hur mycket pojkarna betalade **var mindre än vad deras äppelträd var värt** eftersom trädet var viktigt för dem .
- 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**

- Men **har det hänt något problem med mig** → Men **jag har fått ett problem**
- att ta det lugnt och njuta av livet och allt som sker i de, inte bara pengarna → att ta det lugnt och njuta av livet och allt som sker i det, inte bara **tänka på** pengar (cf. *6.18 S-Ext vs S-M*)
- Riad skriver om ett positiv framtid för svenska i Finland på allt folk har gemensamt kulturell → Riad skriver om en positiv framtid för svenskan i Finland **med tanke** på allt folk har gemensamt kulturellt (cf. *6.18 S-Ext vs S-M*)

Note: When categorizing a correction as S-Ext, more specific syntactic corrections involved in the correction (additions or removals of words, word order changes, etc.) are not individually tagged. (Cf. *4.4.2 Corrections of tokens included in group units*.)

5.5.5 S-FinV (placement of finite verb)

Tag name	**Short description of the correction type**	**Higher priority than**	**Lower priority than**	**Sections dealing with relevant discrimination issues**
S-FinV	finite verb placement	S-WO	S-Adv	S-Adv, S-FinV, S-WO: Exceptions to the default ranking of the word order tags (6.11) S-Clause vs word order tags (6.12)

The S-FinV tag is used for corrections of a finite verb placement, unless the correction concerns the ordering between the finite verb and an adverb, in which case the correction is tagged with the S-Adv tag.

- Vor du **bor**? → Var **bor** du?
- I morgon jag **åker** → I morgon **åker** jag
- När jag ätiti, jag **sover** → När jag ätit, **sover** jag
- Eller det **skapar** kontakter och trivs? → Eller **skapar** de kontakter och trivsel?
- Jag har lärt mig hur **ska** man behandla en människa → Jag har lärt mig hur man **ska** behandla en människa

5.5.6 S-M (word missing, other)

Tag name	Short description of the correction type	Higher priority than	Lower priority than	Sections dealing with relevant discrimination issues
S-M	word missing (added)		M-Def M-Verb S-Msubj	M-Def vs M-Num, S-R and S-M (6.10) S-Clause vs S-M (6.15) S-Ext vs S-M (6.18) Clause corrected or created? (M-Verb, S-Clause, S-M, S-Msubj) (7.1) Verbal particles and reflexives (O-Comp, S-Comp, L-Der, L-W, S-M, S-R) (7.3)

The S-M tag is used when a word is missing in the original text and has been added in the normalized version. This includes the addition of reflexives and verbal particles. (See 7.3 *Verbal particles and reflexives (O-Comp, S-Comp, L-Der, L-W, S-M, S-R)*.)

- Jag trivs mycket bo med dem → Jag trivs mycket **med att** bo med dem
- men känner bara fysiskt närvarande → men känner **sig** bara fysiskt närvarande
- Jag slår ord i ordboken när jag inte vet → Jag slår **upp** ord i ordboken när jag inte vet
- Det är värt → Det är värt **det**
- bara två veckor sen → **för** bara två veckor sen
- stället av → **i** stället för
- Jag gillar inte tapeterna i min kusins lägenhet men han inte byta dem eftersom det är hyresrätt → ... han **kan** inte byta dem ...

  (In this case the finite verb is tagged S-M rather than S-Clause, since the intention of a clause is indicated by the presence of both a subject and an infinite verb in the original text, cf. 7.1 *Clause corrected or created? (M-Verb, S-Clause, S-M, S-Msubj)*.)

Note (priority, 1): Added articles should be tagged M-Def rather than S-M.

Note (priority, 2): Added *ha*, *komma* and *skola* auxiliaries should be tagged M-Verb rather than S-M.

Note (priority, 3): Added subjects should be tagged S-Msubj rather than S-M.

Note (discrimination, 1): Additions of *subjunktioner* (‘subordinating conjunctions’) are tagged S-Clause rather than S-M. (See *6.15 S-Clause vs S-M*.)

Note (discrimination, 2): Finite verbs and subjects which are added as part of creating **a whole new clause** in the normalization should be tagged S-Clause rather than S-M. (See *7.1 Clause corrected or created?*)

Note (discrimination, 3): When the addition of a singular indefinite article (*en* or *ett*) is due to a change of number, the article should be tagged M-Num rather than M-Def or S-M. (See 6.10 *M-Def vs M-Num, S-R and S-M*.)

5.5.7 S-Msubj (subject missing)

Tag name	**Short description of the correction type**	**Higher priority than**	**Lower priority than**	**Sections dealing with relevant discrimination issues**
S-Msubj	subject missing (added)	S-M		S-Clause vs S-Msubj (6.16) Clause corrected or created? (M-Verb, S-Clause, S-M, S-Msubj) (7.1)

The S-Msubj tag is used to mark corrections involving the addition of a subject which is missing in a clause in the original text. This includes cases when the pronoun/*subjunktion* 'subordinating conjunction' *som* is inserted as a subject. It also includes typical instances when a *det* has been inserted due to *platshållartvånget* 'the placeholder requirement'.

- Tycker om min plats → **Jag** tycker om platsen där jag bor
- Regnar ute → **Det** regnar ute
- Det är inte bara arbetet och arbetslivet kan ge stress → Det är inte bara arbetet och arbetslivet **som** kan ge stress
- Det är viktigt att veta vad händer → Det är viktigt att veta vad **som** händer
- I annan text … beskriver dem tre olika ställningar som finns om svenska inom Finland → I en annan text … beskriver **författaren** de tre olika inställningar som finns till svenska i Finland
- Finns många nya lagenheterna i dyrare delar i huvudstaden → **Det** finns många nya lägenheter i dyrare delar i huvudstaden

The S-Msubj tag should be placed on a *det* which has been inserted as a subject, even when the original clause contains a subject (see *6.16 S-Clause vs S-Msubj*), i.e. also in cases like the following:

- *Det* is inserted as a *formellt subjekt* ('formal subject') in the normalized text, and the subject in the original text is changed to an *egentligt subjekt* ('object-positioned subject'):
    - på andra sidan finns *människor som har så mycket pengar att de kan köpa halva världen* → Å andra sidan finns **det** *människor som har så mycket pengar att de kan köpa halva världen*

    Det is tagged with S-Msubj. In this particular example, no additional visible correction corresponding to the change of the original subject to an *egentligt subjekt* (‘object-positioned subject’) is made, and this correction is therefore not tagged at all (cf. *4.1.2 Only visible corrections are tagged*).

    - För att sammanfatta och svara på frågan om *likheter och skillnader* finns mellan nynorskans ställning i Norge och svenkans ställning i Finland → För att sammanfatta och svara på frågan om **det** finns *likheter och skillnader* mellan nynorskans ställning i Norge och svenskans ställning i Finland

      The inserted subject *det* is tagged S-Msubj. The movement of the phrase *likheter och skillnader* indicates its change of function from subject to *egentligt subjekt* ‘object-positioned subject’. The link between this unit in the original text and the same unit in the normalized text is thus tagged S-Clause. (Cf.6.12 *S-Clause vs word order tags*.)

- A heavy subject is moved to a position further back in the clause, and *det* is inserted in the subject position:

    - Hon försätter skriva om *att inte avskaffa den obligatoriska svenskan på skolan* är fördel för ungdomar → Hon fortsätter skriva om att **det** är en fördel för ungdomar *att inte avskaffa den obligatoriska svenskan i skolan*

      The inserted *det* is tagged as S-Msubj and the moved subject is tagged as S-WO.

- *Det* (possibly followed by *att*) is inserted as a subject before an infinitive phrase already functioning as a subject (and the infinitive phrase becomes an attribute to *det*):

    - Hon skriver också om att *ha svenska som obligatorisk i Finland* visar att alla är ju finnar och att " finlandssvenska " är bara påhittat . -> Hon skriver också om att **det att** *ha obligatorisk svenska i Finland* visar att alla är ju finnar och att " finlandssvenska " är bara påhittat .

      det is tagged as S-Msubj, *att* is tagged S-M. The function shift of the infinitive phrase is an “invisible” correction, and is thus not tagged at all (cf. *4.1.2 Only visible corrections are tagged*).

Note (discrimination):

- The S-Msubj tag should only be applied in those cases when the clause is already present in the original text. When a subject has been added as part of a correction involving the creation of a clause which was not present in the original text, the subject should be tagged as S-Clause, just as the rest of the added or changed elements involved in the creation of the clause. (See 7.1 *Clause corrected or created? (M-Verb, S-Clause, S-M, S-Msubj)*.)
- The S-Msubj tag should however be applied to a *det* which is added as a *formellt subjekt* (‘formal subject’) to an already present clause even when this clause already contains another subject. See examples above and *6.16 S-Clause vs S-Msubj.*

5.5.8 S-Other

Tag name	Short description of the correction type	Higher priority than	Lower priority than	Sections dealing with relevant discrimination issues
S-Other	other syntactical correction		All other S-categories	

The S-Other tag is used for syntactic corrections not covered by any of the other S-tags. It includes:

- A negated pronoun is exchanged for a negation and a non-negated pronoun:
  - jag önskar att det finns **inget** problem att få min pengar tillbaka → jag hoppas att det **inte** finns **något** problem med att få mina pengar tillbaka

    (Two S-Other tags are applied: One on a link between *inget* and *något*, and one between *inte* and "nothing".)
- A double negation is corrected to a simple negation:
  - seden har jag fått att anm$la till andra kurs *utan* betala **ingeting** → sedan har jag fått anmäla mig till en annan kurs *utan* att betala **någonting**

5.5.9 S-R (word redundant)

Tag name	Short description of the correction type	Higher priority than	Lower priority than	Sections dealing with relevant discrimination issues
S-R	word redundant (removed)		M-Def M-Verb	M-Def vs M-Num, S-R and S-M (6.10) S-Clause vs S-R 6.17) S-Type vs S-M and S-R (6.20) Verbal particles and reflexives (O-Comp, S-Comp, L-Der, L-W, S-M, S-R) (7.3)

The S-R tag is used when a word is redundant in the original text and has been removed in the normalized version. This includes the removal of reflexives and verbal particles. (See 7.3 *Verbal particles and reflexives (O-Comp, S-Comp, L-Der, L-W, S-M, S-R)*.)

- Man behöver inte **att** klä på sig → Man behöver inte klä på sig
- Det är ett personligt ansvar **för** att välja → Det är ett personligt ansvar att välja

- X-stad är närmare **till** X-land än X-stad → X-stad är närmare X-land än X-stad
- De första mobiltelefonerna kom **i** 1957 → De första mobiltelefonerna kom 1957
- De nya kommunikationssätten har medfört **med** stora möjligheter → De nya kommunikationssätten har medfört stora möjligheter
- sociala medier blev en essentiell del av våra liv som vi inte kan slänga **det** i den moderna världen → sociala medier blev en essentiell del av våra liv som vi inte kan slänga i den moderna världen
- De promenerade **sig** i parken → De promenerade i parken
- Hon gav **bort** honom en blomma → Hon gav honom en blomma

Note (priority, 1): Removed articles should be tagged M-Def rather than S-R.

Note (priority, 2). Removed *ha, komma* and *skola* auxiliaries should be tagged M-Verb rather than S-R.

Note (discrimination, 1): Removed *subjunktioner* 'subordinate conjunctions' should be tagged S-Clause rather than S-R. (See *6.17 S-Clause vs S-R*.)

Note (discrimination, 2): Removals of words which are part of restructuring a phrase or a clause in a way which removes a clause altogether should also be tagged S-Clause rather than S-R. (See *6.17 S-Clause vs S-R*.)

Note (discrimination, 3): When the removal of a singular indefinite article (*en* or *ett*) is due to a change of number, the article should be tagged M-Num rather than M-Def or S-R. (See 6.10 *M-Def vs M-Num, S-R and S-M*.)

5.5.10 S-Type (change of phrase type/part of speech)

Tag name	**Short description of the correction type**	**Higher priority than**	**Lower priority than**	**Sections dealing with relevant discrimination issues**
S-Type	change of phrase type/part of speech			S-Type vs S-M and S-R (6.20)

The S-Type tag is used when a word or phrase has been changed to a word with another part of speech or another phrase type. The S-Type tag is usually combined with an L-Der or an L-W tag.

Example, S-Type and L-W

- Tycker om min plats har en **köpa mat** ett , litet centrum en förskolan en vårdcentralen → Jag tycker om platsen där jag bor , den har en **mataffär** , ett litet centrum , en förskola , en vårdcentral

  See 5.2.4 L-W (wrong word or phrase) for more examples.

Example, S-Type and L-Der

- Jag ska **solen** och bada → Jag ska **sola** och bada

  See *5.2.1 L-Der (word formation)* for more examples.

Note (discrimination): Whenever the syntactic category of a phrase is changed *only* by the addition or removal of one or more words, the S-M and S-R tags should be used rather than the S-Type tag, see *6.20 S-Type vs S-M and S-R*.

5.5.11 S-WO (word order, other)

Tag name	Short description of the correction type	Higher priority than	Lower priority than	Sections dealing with relevant discrimination issues
S-WO	word order, other		S-Adv S-FinV	S-Adv, S-FinV, S-WO: Exceptions to the default ranking of the word order tags (6.11) S-Clause vs word order tags (6.12)

The S-WO tag is used for word order corrections which are not covered by the S-Adv or the S-FinV categories.

In corrections regarding the relative placement of a phrasal head and a modifying element (for instance a noun and its attribute), the modifying element should be marked rather than the phrasal head (e.g. *min* rather than *bostad* in the example below):

- Bostaden **min** → **Min** bostad
- ögon **grå** → **grå** ögon
- en plats **lite kallt** → en **lite kall** plats

In cases when the word order change may be interpreted as moving an element "out of" a normally fairly fixed structure, that element should be marked rather than an element included in the more fixed structure (*dem* rather than *upp* in the example below).

- Man kan inte vänta att lägga **de** upp på social medier → Man kan inte vänta med att lägga upp **dem** på sociala medier

In other cases the placement of the tag may be chosen freely between the elements which have been moved relative to each other, but the readability of the resulting visualization should be taken into consideration.

- Om man är otrogen, skulle **påverkas deras kärleksrelation** → Om man är otrogen, skulle deras **kärleksrelation påverkas**

- **Tidigare Finlands** president → **Finlands tidigare** president

5.6 Other tags

The final group of tags, collected under the heading *Other*, contains six tags used for various purposes:

- The C tag is used for corrections which are necessitated by other corrections, but which do not reflect mistakes in the original text.
- The Cit-FL tag is used for segments of foreign language which have not been corrected during normalization.
- The Com! and OBS! tags are used for notes and comments – the Com! tag for comments intended for the future corpus user, and the OBS! tag for work notes on pending analyses meant for internal project use.
- The Unid tag is used for "unidentified" corrections, i.e. corrections which are not covered by any of the correction categories defined in the taxonomy.
- The X tag is used for unintelligible strings.

Four of these tags, namely Cit-FL, Com!, OBS! and X are available already during the normalization process, and are normally inserted already by the normalizer.

5.6.1 C – Consistency corrections

Tag name	**Short description of the correction type**	**Higher priority than**	**Lower priority than**	**Sections dealing with relevant discrimination issues**
C	consistency correction, necessitated by other correction			C vs M-Def, M-Verb and M-Num (6.21)

The C tag represents consistency corrections, a category which covers necessary (follow-up) corrections in the normalized text that come as a result of a previous correction, i.e. originally there was no mistake in the segment, but due to an introduced correction a follow-up correction is necessary in the segment. By using this tag we indicate that the error was not made originally by the learner.

In some instances it may not be self-evident which one of two related corrections that should be considered as necessitating the other, but by marking one of them with C we avoid marking a single mistake in the original text as two.

- **Bostaden** min → Min **bostad**

  The shift of word order is marked as a word order correction (S-WO). The change from definite form (*bostaden*) to indefinite form (*bostad*) of the noun is made necessary because of the shift of word order, and is thus marked as a consistency correction (C).

- Det ger en en positiv **energi därmed** kan man bli av med stress → Det ger en en positiv **energi . Därmed** kan man bli av med stress

  The insertion of the full stop is tagged as a punctuation correction (P-M). The capitalization of the following D is made necessary because of the insertion of the full stop, and is thus marked as a consistency correction (C).

- **Min** bostanden ser ut som lilla centrum med vlere affere → **Mitt** bostadsområde ser ut som ...

  The change from *bostanden* to *bostadsområde* is marked as a lexical correction (L-W) and a morphological correction (M-Def). The gender change of the pronoun, from *min* (non-neuter) to *mitt* (neuter) is made necessary because of the change of word (from the non-neuter *bostad* to the neuter *bostadsområde*), and is thus marked as a consistency correction (C).

- För det första **är** vi **går** skolan från dagsskolan till allika nivå för att vi hittar jobb → För det första **går** vi i skolan från förskolan till olika nivåer för att vi ska hitta ett jobb

  The removal of *är* is marked as S-R, while the movement of *går* to the finite verb position (where it replaces the likewise finite verb *är*) is tagged with C.

- När jag kommer **första** i 2012 , jag bodde i en social lagenhet med 3 andra person → När jag **först** kom 2012 bodde jag i en kommunal lägenhet med 3 andra personer

  The link between *första* and *först* is tagged with M-Adj/adv (see this section) **and** with C, which indicates that the word order change is a consequence of the change from adjective to adverb. (An alternative normalization would be to keep the adjective form *första* and add the noun *gången*.)

- Jag stoppade i lång tid att saga till mina vänner i D-hemland hur jag bo här , **vad** extra vi ha här , e.g. tvättstuga → Jag slutade för länge sen att säga till mina vänner i D-hemland hur jag bor här , **vilka** extra **saker** vi har här , t.ex. tvättstuga

  The addition of *saker* is tagged as S-M, and the correction of *vad* to *vilka* is tagged with C, since it is the congruence with *saker* which determines the choice of pronoun.

Note: The C tag is not used when multiple tokens are involved in a definiteness correction of an NP, or when multiple tokens are involved in an M-Verb correction. It is however used in certain cases when several tokens are involved in a number change of an NP. (See 6.21 C vs M-Def, M-Verb and M-Num.)

5.6.2 Cit-FL (cited foreign word judged acceptable in the normalization)

Tag name	**Short description of the correction type**	**Higher priority than**	**Lower priority than**	**Sections dealing with relevant discrimination issues**
Cit-FL	non-Swedish word **kept**, i.e. **not corrected**			Non-Swedish words and sequences (O, Cit-FL, L-Der, L-FL, L-W) (7.2)

The Cit-FL tag is used for foreign (non-Swedish) words, phrases or text segments, which have been kept by the normalizer since their usage has been judged acceptable given the norms of the text type in question. This may be the case for instance for explicitly marked citations or intentional code switching appropriate for the genre. The Cit-FL tag is thus used to mark words and text segments which have *not* been corrected in the normalized version, but which nevertheless are not passable as standard Swedish. The Cit-FL tag is usually added already during normalization.

Note that the only requirement for applying this code is that the word or text segment is recognizable as another language than Swedish, and that the choice to use this other language is judged appropriate for the genre and the text. No judgement or correction of the word or text segment is made relative to the norms of the foreign language in questions. For instance, spelling mistakes are left uncorrected.

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 "** , låter det i ett debattinlägg .

5.6.3 Com! (comments for the corpus users)

The Com! tag is connected to an *Edge comment* field, which is open for freely composed comments. It is available already during the normalization process.

The Com! tag is used for comments on specific tokens or text sequences which are relevant for future users of the corpus, and which are thus meant to be kept in the published corpus. (Comments regarding the text as a whole or recurring properties in the text may be added in the Document comment field, see 4.5 *Document comments*.)

The Com! tag may for instance be used to mark text segments which are copied from the task formulation. If a significant portion of the text consists of copied text, this should preferably be indicated also in the Document comment field, in addition to the edge comment field connected to the Com! tag.

The Com! tag is also used when a string which in the transcribed original includes a “$” (representing an unreadable character) has been assumed to be correct, and thus has not been annotated with a tag representing any correction category. In such cases the comment “Original assumed to be correct” is written in the edge comment field. Cf. *4.2.1 Annotation of strings including unreadable signs in transcribed texts*.

5.6.4 OBS! (notes and pending analyses)

The OBS! tag is, like the Com! tag, connected to an Edge comment field, and is available already during the normalization process.

The OBS! tag is used to mark pending analyses to which the annotator wants to return, remarks which the normalizer wishes to pass on to the correction annotator, etc.

5.6.5 Unid (unidentified correction)

The Unid tag is used for any type of correction which cannot be covered by any of the correction categories defined in the taxonomy.

5.6.6 X (unintelligible string)

Tag name	**Short description of the correction type**	**Higher priority than**	**Lower priority than**	**Sections dealing with relevant discrimination issues**
X	unintelligible string			S-Ext vs X (6.19)

The X tag is used to mark unintelligible strings in the original text. The tag is available and usually added already during the normalization process. The marked original string may be left unchanged in the normalized version, or the normalizer may replace it with a more or less wild guess of the intended message.

The X tag may be used both in cases when there is no reasonable interpretation of the string, and when there are several somewhat reasonable but diverging interpretations, and none of these interpretations may be considered as better than the other.

- **En argumentera på är viktigt hur man bor är bor på en bra hem**
- och vi har 3 rum it huset **dar rum** är meka bra liv med maen familija dar huset familjbostder och jag vill **sta** och jog kan ante skriv meka ord → och vi har 3 rum i huset . **dar rum** är mycket bra liv med min familj där i huset från Familjebostäder och jag vill **stanna** och jag kan inte skriva många ord

  The text from which this example is collected has some German or possibly Dutch traits, and a fairly resonable guess is that *dar rum* is meant to be *darum*. Another, less likely, possibility is that *dar rum* is intended to mean *där rum*, but that interpretation necessitates extensive changes in the rest of the text passage in order to create a syntactically functional string. By marking *dar rum* with X and keeping the rest of the sentence unchanged in the normalized text version, the normalization as a whole is better adjusted to the principle of *minimal change*.

  Stanna is a reasonable guess about the intention with *sta*, but not well founded enough for the correction to be marked as an orthographic correction. Many other interpretations are obviously also possible.

- Jag bo i ett lägenhet med min sambo och våras **yng** dotter → Jag bor i en lägenhet med min sambo och vår **lilla** dotter

  While the string *yng* suggests that the intended word might be *yngsta*, the rest of the text makes *lilla* appear more likely.

Back to the menu

6 Discriminating between specific correction categories

6.1 O vs punctuation tags

The removal, addition or change of token-internal punctuation is in some instances tagged as O, and in some instances tagged with the suitable punctuation tag. (Cf. also 7.4 *Spaces, hyphens and dashes (O-Comp, P-M, P-R, P-W)*.)

- Removal or addition of periods from abbreviations is tagged O:
  - **t.v. → tv**

    (*tv* is used as for *television*, not *tills vidare*)
- Additions of colons between abbreviations (etc.) and suffixes are tagged O:
  - **vdar → vd:ar**
- Change of punctuation in tokens mainly consisting of numbers is generally tagged with the suitable punctuation tag:
  - Forskning och framsteg **2009.2** → Forskning och framsteg **2009:2** (P-W)
  - **5.5** procent → **5,5** procent (P-W)

Back to the menu

6.2 O-Comp vs S-Comp

Corrections concerning the forming of an expression as a compound or a multi-word expression are divided into two categories in the SweLL correction taxonomy: Corrections which are judged to concern the mere orthographic rendering with or without a space between two words are marked with the O-Comp tag, while corrections which are judged to concern the actual choice between a compound and a multi-word expression are marked with the S-Comp tag. Borderline or unclear cases between these two categories obviously exist, but for the most part one of the options is clearly the better one.

The **O-Comp** tag is primarily used for corrections of standard cases of *särskrivning* (the faulty writing of a compound with a space in between the two compounded words):

- Det ligger ett **kultur hus** nära min bostad → **kulturhus**

It may also be used for corrections which involve changing the first word of the compound into a specific compound form, in addition to the removal of the space between the two words. In such cases, the correction is marked with both the O-Comp tag and the L-Der tag:

- **Kommunikation förändring → Kommunikationsförändring**

The O-Comp tag may also be applied when a multi-word expression in the original text has been corrected through the adding of a space between two of the words:

- **engång → en gång**

The **S-Comp** tag is used whenever the correction is more complex than the mere addition or removal of a space between two words (and possibly changing the form of the first word of a compound).

- **det vardagliga livet → vardagslivet**
- **livsskillnaden → skillnaden i liv**
- **svenska undervisningar → svenskundervisning**
- Enligt Hyltenstam så kan minoritetsspråk räddas om man **inblandar** dem äldre som kan språket → Enligt Hyltenstam så kan minoritetsspråk räddas om man **blandar in** de äldre som kan språket

However, in some cases the S-Comp tag is more suitable than the O-Comp tag, although the correction superficially merely involves the presence of a space. This is the case when two words may be correctly formed as a compound (without a space) *and* as a two-word expression (with a space), and the meaning of the compound version and the two-word version are either the same or else both plausible in the context. The correction made is thus not due to the chosen expression being unthinkable, but due to the other expressions happening to be the established lexical unit:

- Om det händer att det finns planhalvor så kan det bero på blyghet, osäkerhet, rädsla eller **socialfobi → social fobi**

In this case, *socialfobi* is a perfectly well formed compound, but it is corrected to the two-word expression *social fobi* because this is the established way to express the intended meaning. The mistake made by the writer is therefore not judged to be a case of having missed a space in between two words, but it is judged to be a case of the writer actually having chosen the compound expression instead of the established two-word expression. The correction is thus marked with the S-Comp tag.

Moreover, in some cases the S-Comp tag may be applied although neither the original nor the normalized string is a single orthographic word. This includes cases when the string in the original text may be interpreted as an instance of a compound, although it includes spaces:

- Hejdlös **sociala medier användning** orsakar ensamhet → Hejdlös **användning av sociala medier** orsakar ensamhet

In this particular case, the string in the original text may be interpreted as a compound between a two-word phrase (*sociala medier*) and the word *användning*. According to the norms of written standard Swedish, such compounds should be written as *sociala medier-användning* etc. While such a minimal correction is not an unthinkable solution for the normalization, the normalizer has here judged a restructuring of the NP as a better solution, and the correction should be tagged S-Comp.

Back to the menu

6.3 L-Der vs L-W

When the stem of a one-word unit is completely exchanged or morphologically restructured, the correction is tagged either with L-Der or with L-W. (Changes of non-Swedish words to Swedish words are excepted, these are tagged L-FL, see 7.2 *Non-Swedish words and sequences (O, Cit-FL, L-Der, L-FL, L-W)*.)

The following general rule determines which of the L-Der and L-W tags that should be chosen:

- If the root morpheme(s) of the category-defining part of the stem is kept, the L-Der tag should be chosen. If not, the L-W tag should be chosen.

This general rule implies the following:

1 The word is completely exchanged (no common morphemes) → **L-W**

- *transformerade →förändrade*

2 Only derivational morphemes are corrected (added, removed or exchanged) → **L-Der**

- *stressiga →stressade*
- *ändring →förändring*
- *förstöra →störa*

3 Corrections of the first word of a compound → **L-Der**

a The form of the first word of a compound is corrected → **L-Der**

- *tvångsvenska →tvångssvenska*
- *sagabok →sagobok*

b The first word of a compound is exchanged → **L-Der**

- *dagsskolan →förskolan*
- *människogrupper →folkgrupper*

4 Corrections of the second word of a compound

a The second word of a compound is completely exchanged (no common morphemes) → **L-W**

- *vårsemestern →vårterminen*

b The second word of a compound is changed in a way which also involves root morphemes (additions, removals, exchanges) → **L-W**

- *maktfull* → *maktfullkomlig*

c The second word of a compound is changed with regard to derivational morphemes (additions, removals, exchanges), but the root is kept → **L-Der**

- *nybyggnad* → *nybyggd*

5 An additional word is added to the word in the original text, forming a compound out of a simplex, or a longer compound out of a shorter one

a A word is added *before* the original word, forming a compound in which the second, i.e. the category-defining, word in the normalized version is identical to the word in the original version → **L-Der**

- *ställning* → *inställning*

b A word is added *after* the original word, forming a compound in which the second, i.e. the category-defining, word in the normalized version is *another* than the word in the original version → **L-W**

- *historian* → *historieskrivningen*

6 A part of a compound is removed, forming a simplex or a shorter compound

a The last, category-defining, part of the compound is kept → **L-Der**

- *orsaksföljden* → *följden*
- *basketbollag* → *basketlag*

b The last, category-defining, part of the compound is removed → **L-W**

- *fikatid* → *fikat*
- *semesterdag* → *semester*

Note: Changes from *båda* to *både*, or the other way around, are tagged as L-Der (and S-Type) rather than as L-W:

- något ha en " svart " avställning **båda** i jobbet och från bostad → Någon har en " svart " inkomst **både** från jobbet och från bostaden

Back to the menu

6.4 L-Der vs M-Verb

Participle forms of verbs are treated as derivations rather than as inflections. Corrections between participle forms and verb forms (e.g. supine forms) are thus tagged L-Der (and S-Type) rather than M-Verb:

- vi har precis **flyttad** till Norrby → vi har precis **flyttat** till Norrby

Accordingly, the L-Der (and S-Type) tag is also used for changes between s-passives and periphrastic passives formed with a copula use of the verb *vara* and a participle:

- I såna kurser skulle arbetsgivare förklara vad arbetsmoral betyder och vilka personliga ekenskaper **är sökta** → I såna kurser skulle arbetsgivare förklara vad arbetsmoral betyder och vilka personliga egenskaper **som söks**

  The link between *sökta* and *söks* is tagged L-Der and S-Type. A link between *är* and "nothing" is tagged C (since the removal of *är* is due to the change from a periphrastic passive to an s-passive). The added *som* is tagged S-Msubj.

Back to the menu

6.5 L-Der vs S-Comp

The L-Der tag is exclusively used for changes between *one-word units* (not necessarily one-token units, since a word may be mistakenly written as two tokens). Both the original unit and the normalized unit should thus consist of just one word for the L-Der tag to be applicable.

When the same lexical morphemes are arranged as a multi-word unit in one text version and as a single word unit in the other text version, the S-Comp tag should be applied rather than the L-Der tag, even when derivational morphemes are involved in the correction:

- Cecila Christner skriver om hur det svenska språket i skolor blir kallad tvångsvenska **i samma tid** svenska har blivit **icke populärt** i Finland → Cecilia Christner skriver om hur det svenska språket i skolorna blir kallat tvångssvenska **samtidigt** som svenska har blivit **impopulärt** i Finland

Back to the menu

6.6 L-W vs M-Case

When the form *dem* is changed to the form *de* used as a definite article, the correction is tagged as L-W and S-Type, not as M-Case:

- den obligatoriska svenskundervisningen i **dem** finska skolorna → den obligatoriska svenskundervisningen i **de** finska skolorna

Back to the menu

6.7 L-W vs M-Verb

The L-W tag is normally used for complete exchanges of words (no root morpheme kept, cf. *6.3 L-Der vs L-W*). This includes most exchanges of auxiliary verbs. Only when a *ha, skola* or *komma* auxiliary is exchanged for another *ha, skola* or *komma* auxiliary is the M-Verb tag used instead.

- *Ha*, *skola* or *komma* auxiliary exchanged for another *ha*, *skola* or *komma* auxiliary, tagged M-Verb:
  - Tyvärr **ska** jag inte komma på kursen → Tyvärr **kommer** jag inte **att** komma på kursen

- Change between a *ha*, *skola* or *komma* auxiliary and another verb, tagged L-W:
    - Jag tror att det **vill** bli bra → Jag tror att det **kommer att** bli bra
- Change between other auxiliary verbs, tagged L-W:
    - Det **måste** inte gå fel → Det **får** inte gå fel
- Change between auxiliary verbs, when at least one of the auxiliaries is not a *ha, skola* or *komma* auxiliary, which also involves a change of tense form or the like, tagged both L-W and M-Verb.
    - Det är så vi **ville** göra → Det är så vi **ska** göra

Back to the menu

6.8 L-W vs S-Clause

Word changes made to change a main clause to a subordinate clause or the other way around are tagged S-Clause rather than L-W:

- Jag är jättetrött **därför** sover jag inte på nätterna → Jag är jättetrött **eftersom** jag inte sover på nätterna

Back to the menu

6.9 L-W vs S-Comp

Both the L-W tag and the S-Comp tag may be used for changes between a one-word unit and a multi-word unit, and for changes between two multi-word units.

The L-W tag may only be used when at least one lexical morpheme is different between the two strings:

- Finns många nya lagenheterna i dyrare delar i huvudstaden , men detta är **lång distans från räker nummer** . → Det finns många nya lägenheter i dyrare delar i huvudstaden , men detta är **långt ifrån tillräckligt** .

If both the original string and the normalized string contain the same lexical morphemes but with another internal structure, the S-Comp tag should be applied rather than the L-W tag:

- **det vardagliga livet** → **vardagslivet**

Back to the menu

6.10 M-Def vs M-Num, S-R and S-M

When a singular indefinite article is removed from a plural indefinite NP, the removed article should be tagged M-Num rather than M-Def or S-R; such a correction concerns number rather than definiteness.

- kanske där kan du träffa **en** nya vänner → kanske kan du träffa nya vänner där

The same holds for additions of singular indefinite articles which are due to the change of an indefinite plural NP to an indefinite singular NP. In such cases, the added article should be tagged M-Num rather than M-Def or S-M:

- nya vänner → **en** ny vän
- större mängder → **en** större mängd

Back to the menu

6.11 S-Adv, S-FinV, S-WO: Exceptions to the default ranking of the word order tags

The three word order tags (S-Adv, S-FinV and S-WO) are in most cases ranked in the following way:

1. S-Adv
2. S-FinV
3. S-WO

There are however two exceptions to this default ranking:

Exception 1: Choose the word order tag which gives the least number of tags

When the choice of one word order tag means that one tag suffices, while the choice of another word order tag requires an additional word order tag, the single tag should always be chosen rather than the multiple tags, regardless of the ranking presented above.

Examples:

- Jag berättade vad **ville** jag helst göra → Jag berättade vad jag helst **ville** göra

  This correction involves the movement of the finite verb *ville* relative to both the subject *jag* and the adverbial *helst*. The word order tag should be placed on the finite verb, and the **S-FinV** tag should be chosen. (If the S-Adv tag had been used for the movement of *helst*, an additional S-WO tag would have had to be placed on the subject *jag*.)

- Jag berättade vad helst ville **jag** göra → Jag berättade vad **jag** helst ville göra

  This correction involves the movement of the subject *jag* relative to both the adverbial *helst* and the finite verb *ville*. The word order tag should be placed on the subject, and the **S-WO** tag should be chosen. (Otherwise both the S-Adv tag and the S-FinV tag would have had to be used.)

Exception 2: Word order changes which are due to a change of *fundament* ('pre-finite element') are tagged S-WO

In some (rare) cases the word order of a clause has been changed not for the sake of clause-internal correctness, but because the textual flow demands it. This is then generally done by changing the element in the *fundament* 'pre-finite position' of the clause. Such corrections are tagged with S-WO rather than with S-FinV or S-Adv, regardless of the function of the moved elements. The finite verb should not be marked with a word order tag as long as this is correctly positioned in the original clause.

Example:

- **Med min familj** bor **jag** → **Jag** bor **med min familj**

  Here, the subject *jag* has replaced the adverbial *med min familj* in the *fundament* 'pre-finite position' in the normalized version. Superficially, this means that *jag* and *med min familj* have switched places. Both these elements are tagged with S-WO. The finite verb *bor* should not be tagged with a word order tag since it is correctly placed in the original version.

Back to the menu

6.12 S-Clause vs word order tags

There are three cases when a word order change should be tagged S-Clause rather than with one of the word order tags (S-Adv, S-FinV or S-WO). These three cases are described below.

Case 1: A unit is moved because its syntactic function has been changed

When a phrase is moved because its syntactic function has changed, this phrase should be tagged S-Clause, instead of a word order tag being used. For instance, in the following example, the phrase *likheter och skillnader* is moved relative to the finite verb *finns* because its syntactic function has changed, which means that *likheter och skillnader* should be tagged S-Clause rather than the finite verb being tagged S-FinV:

- För att sammanfatta och svara på frågan om **likheter och skillnader** finns mellan nynorskans ställning i Norge och svenkans ställning i Finland → För att sammanfatta och svara på frågan om **det** finns **likheter och skillnader** mellan nynorskans ställning i Norge och svenskans ställning i Finland

  The inserted expletive subject *det* is tagged S-Msubj. The movement of the phrase *likheter och skillnader* indicates its change of function from subject to *egentligt subjekt* 'object-positioned subject', and the link between this unit in the original text and the same unit in the normalized text is thus tagged S-Clause.

Case 2: A unit has been moved as part of a change between a main clause structure and a subordinate clause structure *which is also indicated by other means*

When a clause is changed from a main clause to a subordinate clause, or the other way around, the correction is tagged S-Clause.

Many S-Adv and S-FinV corrections concern changes between the word order typical for main clauses (*fa-ordföljd* '*fa*-word order') and the word order typical for subordinate clauses (*af-ordföljd* '*af*-word order'). However, Swedish allows some main clauses with *af*-word order and some subordinate clauses with *fa*-word order. Therefore, a word order change of this type is not in itself considered to indicate a shift between a main clause structure and a subordinate clause structure. Such a change has to be indicated by the removal or the addition of a conjunction, the change from a co-ordinating to a subordinating conjunction, or the like.

For mere word order changes between an *af*-word order and a *fa*-word order, the S-Adv and S-FinV tags are used:

- Jag är jättetrött eftersom jag sover **inte** på nätterna. → Jag är jättetrött eftersom jag **inte** sover på nätterna. (**S-Adv**)
- Jag berättade vad **ville** jag helst göra → Jag berättade vad jag helst **ville** göra (**S-FinV**)

When a correction between a main clause structure and a subordinate clause structure is indicated by a conjunction change/addition or the like, but *also* involves a word order change, the word order change is treated as involved in or a consequence of the change of clause structure:

- Tråkigt att det är tyst och folk hälsar **knappt** på varandra i din område → Tråkigt att det är tyst och **att** folk **knappt** hälsar på varandra i ditt område

  The addition of *att* is tagged with S-Clause, while the movement of the adverbial *knappt* relative to the finite verb *hälsar* is tagged with C rather than with S-Adv, since this word order change is a consequence of the change from a main clause structure to a subordinate clause structure.

- där berättare Matts Lindqvist hatet som finns mot finlandssvenska eller svenska språket är **inte** resultat av partiet Sannfinländarna → där berättar Matts Lindqvist om **att** hatet som finns mot finlandssvenska eller svenska språket **inte** är ett resultat av partiet Sannfinländarna

  This example is a parallel to the example above. The addition of *att* is tagged S-Clause and the movement of the negation *inte* relative to the finite verb *är* is tagged C.

Case 3: Change between an "ordinary" main clause structure and a question structure

- Jag vill att vet vad är problemet som jag inte kan gå till kursen **det finns** ingen plats eller **ni har** stoppade den här kurs och ni ska inte har det igen → Jag vill veta vad det är för problem som gör att jag inte kan gå på kursen : **Finns det** ingen plats eller **har ni** stoppat den här kursen och ska inte ha den igen ?

  In this case, the changed placement of the finite verb relative to the subject is marked with an S-Clause tag on the finite verb instead of an S-FinV tag.

Back to the menu

6.13 S-Clause vs S-Comp

The S-Comp tag is only used for restructuring of the same lexical morphemes within the same phrase and without affecting the primary syntactic function of the words/phrases involved.

In the following example the object predicative *som obligatorisk* in the verb phrase is turned into the attribute *obligatorisk* in the object NP, and the S-Clause tag is thus applied rather than the S-Comp tag:

- Hon skriver också om att ha svenska **som obligatorisk** i Finland visar att alla är ju finnar och att " finlandssvenska " är bara påhittat . → Hon skriver också om att det att ha **obligatorisk** svenska i Finland visar att alla är ju finnar och att " finlandssvenska " är bara påhittat .

Back to the menu

6.14 S-Clause vs S-Ext

The distinction between S-Clause and S-Ext corrections is not clear cut, and borderline cases exist.

The following correction has been tagged with S-Clause on each of the added tokens (*där*, *man*, *kan*), but is to be considered a borderline case for an S-Ext classification:

- Du kan vila på skogen där och finns bad simma → Du kan vila i skogen där och det finns bad **där man kan** simma (G54GT3)

This corrections involves:

- a creation of a clause by adding both a subject and a finite verb
- a creation of a relationship between the created clause and the rest of the sentence, by adding the adverb *där*, thus making the created clause into a subordinate clause

The clause which is created in the normalized version is suggested in the original version only by the infinitive *simma*, and the relationship between the main clause and the created clause is completely implicit in the original version. This lack of support for a specific syntactic organization of the normalized interpretation speaks in favor of an S-Ext classification of the correction. On the other hand, the additions in the normalized version are not that semantically specific or syntactically complex, and both the clause creation (the addition of the subject and the finite verb) and the creation of a subordination relationship (the addition of the adverb *där*) are in themselves standard examples of S-Clause corrections, which speaks in favor of an S-Clause classification.

Back to the menu

6.15 S-Clause vs S-M

Additions of *subjunktioner* ('subordinating conjunctions') and other *bisatsinledare* ('subordinating connectors') are tagged S-Clause rather than S-M. The additions may indicate a change from a main clause structure to a subordinate clause structure, or specify an otherwise unspecified relationship between the clauses in the original version. (See also 7.1 *Clause corrected or created? (M-Verb, S-Clause, S-M, S-Msubj)*.)

- Sådana känslor gör användaren mår dåligt → Sådana känslor gör **att** användaren mår dåligt
- Man kan promonera lång tid finns det blåser → Man kan promenera länge **när** det blåser
- Konsekvenserna man skulle få ifall undervisning i svenska blev frivillig så skulle mer än hälften av finska befolkningen avskaffa svenskan som modersmålsundervisning och istället fokusera på finska då dem sällan använder svenskan → Konsekvenserna man skulle få ifall undervisning i svenska blev frivilligt är **att** då skulle mer än hälften av den finska befolkningen välja bort svenskan som modersmålsundervisning och istället fokusera på finska då de sällan använder svenskan (Cf. this example under *S-Clause*.)
- 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 → 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

Back to the menu

6.16 S-Clause vs S-Msubj

When *det* is added as subject to a clause, *det* should be tagged S-Msubj even in those cases when the original clause already contains a subject. This means that the S-Msubj should be used *instead of* or, in some instances, *in addition to* the S-Clause tag in these cases.

The S-Clause tag is otherwise used for changes of the primary syntactic function of a word in a clause, e.g. changing the function of the original subject. Instances of adding a *det* subject are thus treated as a special case, in order to have all additions of *det* subjects tagged with the same tag – S-Msubj. Additional examples are given under *5.5.7 S-Msubj (subject missing)*.

Examples:

- *du* blir bättre → **det** blir bättre *för dig*

  A link is created between *du* and *för dig*. This link is tagged S-Clause. The inserted *det* subject is tagged S-Msubj. (See the same example under *5.5.2 S-Clause (basic clause structure)*.)

- på andra sidan finns *människor som har så mycket pengar att de kan köpa halva världen* → Å andra sidan finns **det** *människor som har så mycket pengar att de kan köpa halva världen*

  Det is tagged with S-Msubj. In this particular example, no additional visible correction corresponding to the change of the original subject to an *egentligt subjekt* ('object-positioned subject') is made, and this correction is therefore not tagged at all (cf. *4.1.2 Only visible corrections are tagged*.).

- För att sammanfatta och svara på frågan om **likheter och skillnader** finns mellan nynorskans ställning i Norge och svenkans ställning i Finland → För att sammanfatta och svara på frågan om **det** finns **likheter och skillnader** mellan nynorskans ställning i Norge och svenskans ställning i Finland

  The inserted expletive subject *det* is tagged S-Msubj. The movement of the phrase *likheter och skillnader* indicates its change of function from subject to *egentligt subjekt* 'object-positioned subject', and the link between this unit in the original text and the same unit in the normalized text is thus tagged S-Clause. (Cf. 6.12 *S-Clause vs word order tags*.)

Note: The S-Msubj tag should only be applied in those cases when the clause is already present in the original text. When a subject has been added as part of a correction involving the creation of a clause which was not present in the original text, the subject should be tagged as S-Clause, just as the rest of the added or changed elements involved in the creation of the clause. See *7.1 Clause corrected or created?*

Back to the menu

6.17 S-Clause vs S-R

Restructuring of phrases and clauses which involve the removal of a clause should be tagged with S-Clause rather than S-R, for instance:

- The structure of a noun phrase is changed by changing an attribute in the form of a relative clause to an attribute in the form of a PP:
    - och han tycker att misstänksamheten mellan **grupperna som befinner sig i södra Finland** lyser med sin frånvaro → och han tycker att den misstänksamhet som finns mellan **grupperna i södra Finland** lyser med sin frånvaro

      The underlined string, which is removed in the normalized version, is tagged S-Clause (the link runs between the string in the original version and "nothing").
    - man kan betala **hela avgift kommer från en lagenhet** → man kan betala **hela avgiften för en lägenhet**

      The underlined strings are linked and the link is marked with S-Clause.

The removal of a *subjunktion* 'subordinate conjunction' or another *bisatsinledare* ('subordinating connector'), indicating the shift from a subordinate clause structure to a main clause structure, should also be tagged S-Clause rather than S-R:

- **Om** du skapar något för att inte känner dig ensam → Du kan göra något för att inte känna dig ensam

Back to the menu

6.18 S-Ext vs S-M

The distinction between S-Ext and S-M is not clear cut, and borderline cases exist. When the added words may be seen as a correction of a goal structure which is fairly clearly aimed at in the original text, the S-M tag should be used. But when the added words are rather a creation of a structure, the S-Ext tag should be used.

The following examples are borderline cases between S-Ext and S-M corrections, which have been tagged as S-Ext:

- att ta det lugnt och njuta av livet och allt som sker i de, inte bara pengarna → att ta det lugnt och njuta av livet och allt som sker i det, inte bara **tänka på** pengar

  In this case, the added verb becomes the main word in the infinitive phrase of the normalized text, which speaks in favor of considering the structure created rather than corrected.
- Riad skriver om ett positiv framtid för svenska i Finland på allt folk har gemensamt kulturell → Riad skriver om en positiv framtid för svenskan i Finland **med tanke** på allt folk har gemensamt kulturellt

The following example has however been tagged as S-M, since the added words are a repetition of the words in the preceding clause, and thus not subject to interpretation to the same extent as in the examples above:

- I Finland och i Norge ungdommar har känslan att de blir tvingat att lära ett språk som majoritet av dem ville inte för att språket inte längre har eller aldrig har haft ett stort betydelse i samhället → I Finland och i Norge har ungdomar känslan att de blir tvingade att lära sig ett språk som majoriteten av dem inte vill **lära sig** för att språket inte längre har eller aldrig har haft en stor betydelse i samhället

Back to the menu

6.19 S-Ext vs X

The S-Ext tag should only be applied in cases when a correction has actually been made, and when the original text gives fairly sound support for the interpretation presented in the normalized version. Text segments which are so difficult to interpret that they are either left unchanged or normalized on the basis of guesses rather than interpretations should be tagged with X.

Drawing the line between unintelligible text segments (X) and text segments with a very fuzzy but still interpretable structure (S-Ext) is not easy and in many cases a matter of subjective judgement. The following example represents a borderline case between an X case and an S-Ext case:

- 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**

The choice to categorize this correction as S-Ext – i.e. as an interpretation with sound support from the original string – rather than as an X case – i.e. as a mere guess – is not self-evident.

Back to the menu

6.20 S-Type vs S-M and S-R

Quite often, the addition or removal of a word also means that the syntactic category of a phrase is changed. For instance, the addition of a preposition may turn an NP to a PP, and the removal of a preposition may result in the opposite transformation. This calls for a clarification of when to use the S-M and S-R tags and when to use the S-Type tag.

Whenever the syntactic category of a phrase is changed *only* by the addition or removal of one or more words, the S-M and S-R tags should be used rather than the S-Type tag:

- jag bor in lägenhet plan ett → Jag bor i en lägenhet **på** plan ett (**S-M**)

- Bostad i D-hemland är litet het topik **för** att diskussera → Bostäder i D-hemland är ett lite hett ämne att diskutera (**S-R**)

However, when the syntactic category of a phrase involves an addition or a removal of a word *on top of another correction*, the correction as a whole (including the addition/removal) is tagged S-Type:

- jag behover pengar f$r **liv** och betalning av min hus → jag behöver pengar för **att leva** och betala för mitt hus

  att leva in the normalized text is grouped as one unit linked with *liv* in the original text, and the link is exclusively tagged with S-Type.

Back to the menu

6.21 C vs M-Def, M-Verb and M-Num

It is far from always self-evident whether a correction should be considered a consequence of another correction or a separate correction. Many of the borderline cases involve changes involving definiteness of NPs, changes of verb phrase constructions, and number changes involving more than just one word in an NP. The following guidelines are meant to facilitate a consistency in the treatment of such cases:

- **M-Def**: All words (articles, adjectival attributes, the main noun) involved in a change of definiteness of an NP are labeled with separate M-Def tags. The C tag is not used.
- **M-Verb**: All words involved in a change of a type which is covered by the M-Verb tag are labeled with separate M-Verb tags. The C tag is not used.
- **M-Num**: When several words in the same NP are changed with regard to number, including the main word, the tag M-Num should only be placed on the main word. The other words are tagged with C, since these corrections are a consequence of the number change of the main word.
  - därför kommer det att leda till **ett negativt konsekvens** → därför kommer det att leda till **negativa konsekvenser**

    The number change from *konsekvens* to *konsekvenser* is tagged M-Num. The change from *negativt* to *negativa* is tagged with C (as a consistency change caused by the adjective depending on the corrected noun) and with M-Gend (since the gender error of the original *negativt* is also eliminated through this change, cf. 4.3 *Several tags on the same link*). The removal of the article *ett* is, for the same reasons, also tagged with both C and M-Gend.

Back to the menu

7 Other categorization issues

7.1 Clause corrected or created? (M-Verb, S-Clause, S-M, S-Msubj)

In some cases, the choice of tag is dependent on whether a clause in the normalized version is present already in the original version or has been created as part of the normalization. In short, we consider an expression a clause if it 1) contains a finite verb (regardless of whether it contains a subject) or 2) contains a subject and a verb (regardless of whether the verb is finite).

When a clause has been *created in the normalization*, the S-Clause tag is used for the corrections involved, and the following applies:

- An added subject is tagged S-Clause rather than S-Msubj.
- An added finite verb is tagged S-Clause rather than S-M or M-Verb.
- A non-finite form of a verb which has been changed to a finite form is tagged S-Clause rather than M-Verb.

A clause is considered *created in the normalization* in the two following cases:

1 Both the subject and the finite verb of the normalized clause are lacking in the original version. (A non-finite verb may be present.)

   - Jag studerar på eftermiddag Sfi och förmiddag praktik → Jag studerar sfi på eftermiddagen och på förmiddagen **har jag** praktik

     Both *har* and *jag* are tagged as S-Clause, not as S-M and S-Msubj respectively.

   - **Att växa** upp som en flicka så var det väldigt många orättvisor man fick vänja sig vid. → **När man växte** upp som en flicka så var det väldigt många orättvisor man fick vänja sig vid.

     Att växa and *när man växte* are grouped together, and the link between these two group units is tagged S-Clause; no additional tagging is made. Thus, *man* is not tagged S-Msubj, and the change of the infinite verb form *växa* to the finite form *växte* is not tagged M-Verb.

   - **min plats → platsen där jag bor**

     min plats and *platsen där jag bor* are grouped together, and the link between these two group units is tagged S-Clause; no additional tagging is made. Thus, *jag* is not tagged S-Msubj and and *bor* is not tagged S-M.

   - Dem som tycker att avskaffa den obligatoriska svenskan i skolan i Finland är Tuija Nikko och Maria Tolppanen → De som tycker att **man ska** avskaffa den obligatoriska svenskan i skolan i Finland är Tuija Nikko och Maria Tolppanen

     Both *man* and *ska* are tagged as S-Clause, not as S-Msubj and S-M/M-Verb respectively.

2 The subject of the normalized clause exists in the original text, but *no verb at all* (finite or infinite) corresponding to or part of the VP of the normalized clause is present in the original version.

   - Mitt område jättefint → Mitt område **är** jättefint

     The added finite verb *är* is tagged as S-Clause rather than as S-M.

A clause is considered *corrected* rather than created when none of these two conditions hold, i.e. when the original expression either contains a finite verb (regardless of whether it also contains a subject) or contains a subject and a verb (regardless of whether the verb is finite). In these cases the tags **S-Msubj, S-M** and **M-Verb** are applied rather than S-Clause.

- Jag gillar inte tapeterna i min kusins lägenhet men han inte byta dem eftersom det är hyresrätt → ... han **kan** inte byta dem ...

  Here, the added finite verb *kan* is tagged as S-M rather than as S-Clause, since both the subject (*han*) and a non-finite verb (*byta*) is present already in the original version.

Back to the menu

7.2 Non-Swedish words and sequences (O, Cit-FL, L-Der, L-FL, L-W)

There are a number of ways to handle non-Swedish words during normalization and correction annotation. Many of the fundamental choices are made during the normalization process rather than during the correction annotation process.

The first judgement to be made when coming across a word stemming from another language is naturally whether the word may be recognized as having been incorporated into written standard Swedish; in such cases the word is left uncorrected and untagged. This judgement is made during normalization.

When a word (or sequence) in an original text is recognized as belonging to a foreign language – or as having traits from a foreign language – and when this word may *not* be recognized as part of written standard Swedish, a number of options are at hand:

1 The word/sequence is judged as a genre appropriate usage of cited foreign language (explicitly signaled citations, code switching etc.). → Not corrected, tagged **Cit-FL** during normalization.

2 The word is not judged as a genre appropriate usage of cited foreign language and is thus corrected to a Swedish word during normalization:

   a The form used may be interpreted as a misspelled Swedish word. → Corrected during normalization, tagged **O** during correction annotation: **kaffee → kaffe**; **can → kan**

   b The form used may be interpreted as a Swedish word with an incorrect usage of derivational affixes etc. → Corrected during normalization, tagged **L-Der** during correction annotation: **national helgdag → nationell helgdag**

   c Neither a nor b applies. → Corrected during normalization, tagged **L-FL** during correction annotation: **balkony → balkong**; **family → familj**; **gas bojler → gaskokare**

Note (1): A word in the original text which is identifiable as a Swedish word, but which is used with another meaning in a way which is likely to be due to influence from a similar non-Swedish word, should be corrected and marked as L-W (not as L-FL):

- Alla blir **busiga** med sina sociala medier. → Alla blir **upptagna** med sina sociala medier

In this example, it is likely that the incorrect usage of the correct Swedish word *busiga* is influenced by the word's similarity to the English word *busy* – and it is partly based on this assumption that the writer's intended meaning has been interpreted as 'upptagna'. But since *busiga* is a correct Swedish word, with a distinctly Swedish morphological structure, the correction is tagged as L-W rather than as L-FL.

Note (2): Phrases which are clearly influenced by a foreign language (and changed in the normalized text) but which consist of identifiable Swedish words, should not be tagged with L-FL, but the relationship between the original phrase and the normalized phrase should be analyzed as a Swedish-internal relationship:

- det finns ingenting **för fri** → det finns ingenting **som är gratis**

  Although *för fri* is clearly an *översättningslån* 'directly translated phrase' from English, it is not tagged L-FL. Instead, the words with the main semantic content, *fri* and *gratis*, are linked and tagged L-W, *för* is tagged S-R, and *som* and *är* are both tagged with S-Clause (since the correction involves an increased number of phrases).

Back to the menu

7.3 Verbal particles and reflexives (O-Comp, S-Comp, L-Der, L-W, S-M, S-R)

Several tags are used for corrections involving phrasal or compound verbs made up by a verb and verbal particle or a reflexive marker, primarily O-Comp, S-Comp, L-Der, L-W, S-M and S-R. This section provides an overview of the usage of these six tags for this category of corrections.

- **O-Comp**: A space is removed between a verbal particle and a following verb, making up a compound verb:
  - **upp mana** → **uppmana**
- **S-Comp**: A compound form of a particle verb is changed to a phrasal form, or the other way around:
  - Enligt Hyltenstam så kan minoritetsspråk räddas om man **inblandar** dem äldre som kan språket → Enligt Hyltenstam så kan minoritetsspråk räddas om man **blandar in** de äldre som kan språket
  - Tåget **går av** från spår 3 → Tåget **avgår** från spår 3
- **L-Der**: A particle of a compound particle verb is changed, removed or added. Both the original and the normalized string consist of one single token. The original string may or may not be an existing Swedish word.
  - Internet **uppmanar** vår förståelse → Internet **utmanar** vår förståelse
  - Han **inhämtade** väskorna → Han **hämtade** väskorna
  - Hon **minde** honom om mötet → Hon **påminde** honom om mötet
- **L-W, placed on the whole phrasal verb**: Either the original string or the normalized string (or both) is a phrasal verb, and the verb itself is changed, not just the particle/reflexive marker:

    - Traditioner ger människorna tid att **stå still** och fundera över livet → … att **stanna upp** och fundera över livet
    - **ställa fram → framföra**

- **L-W, placed on the particle**: A verbal particle is replaced by another verbal particle, but the verb is kept:
    - Han torkade **bort** bordet → Han torkade **av** bordet
- **S-M**: An independent (non-compound) verbal particle or a reflexive marker is added:
    - men känner bara fysiskt närvarande → men känner **sig** bara fysiskt närvarande
    - Jag slår ord i ordboken när jag inte vet → Jag slår **upp** ord i ordboken när jag inte vet...
- **S-R**: An independent (non-compound) verbal particle or a reflexive is removed.
    - De promenerade **sig** i parken → De promenerade i parken
    - Hon gav **bort** honom en blomma → Hon gav honom en blomma

Back to the menu

7.4 Spaces, hyphens and dashes (O-Comp, P-M, P-R, P-W)

Only some errors involving spaces, hyphens and dashes are corrected, and not all of the corrected errors are tagged.

- The O-Comp tag is used for corrections which involve the removal of a space between two words which have been interpreted as making up a compound in the normalized text version, and for the adding of a space between two words. It may also be used for corrections regarding the use of hyphens in compounds.
- The P-R tag is used for removals of hyphens used in other ways, and for removals of dashes.
- The P-M tag is used for added dashes.
- The P-W tag is used for changes between, on the one hand, a hyphen or a dash, and, on the other hand, another punctuation mark (not including a space).

However:

- Instances where a hyphen has been used in the original text where a dash would be more appropriate are left uncorrected and should thus not appear as corrections to be annotated.
- Possible errors involving the incorrect placement of *a space before a punctuation mark* will not be corrected in the normalization process, since spaces are always inserted before punctuation marks for the sake of tokenization. Consequently, such errors will not be tagged.

- Possible 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 are nevertheless left untagged.

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

      In this example, a space is inserted between the period and *tack*, and the t in *tack* is changed from lower to upper case. *Tack* will be tagged with O-Cap, but the insertion of the space will not be tagged.

Back to the menu

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