Correction annotation in SweLL

http://hdl.handle.net/2077/69434

Lisa Rudebeck
Dept of Swedish Language and Multilingualism, Stockholm University

Correction annotation

Correction = a difference between the original learner text and the normalized version of the text

Correction annotation = a categorization of such differences

→Indicates a relationship between the original version and the normalized version (only indirectly properties of the source text)

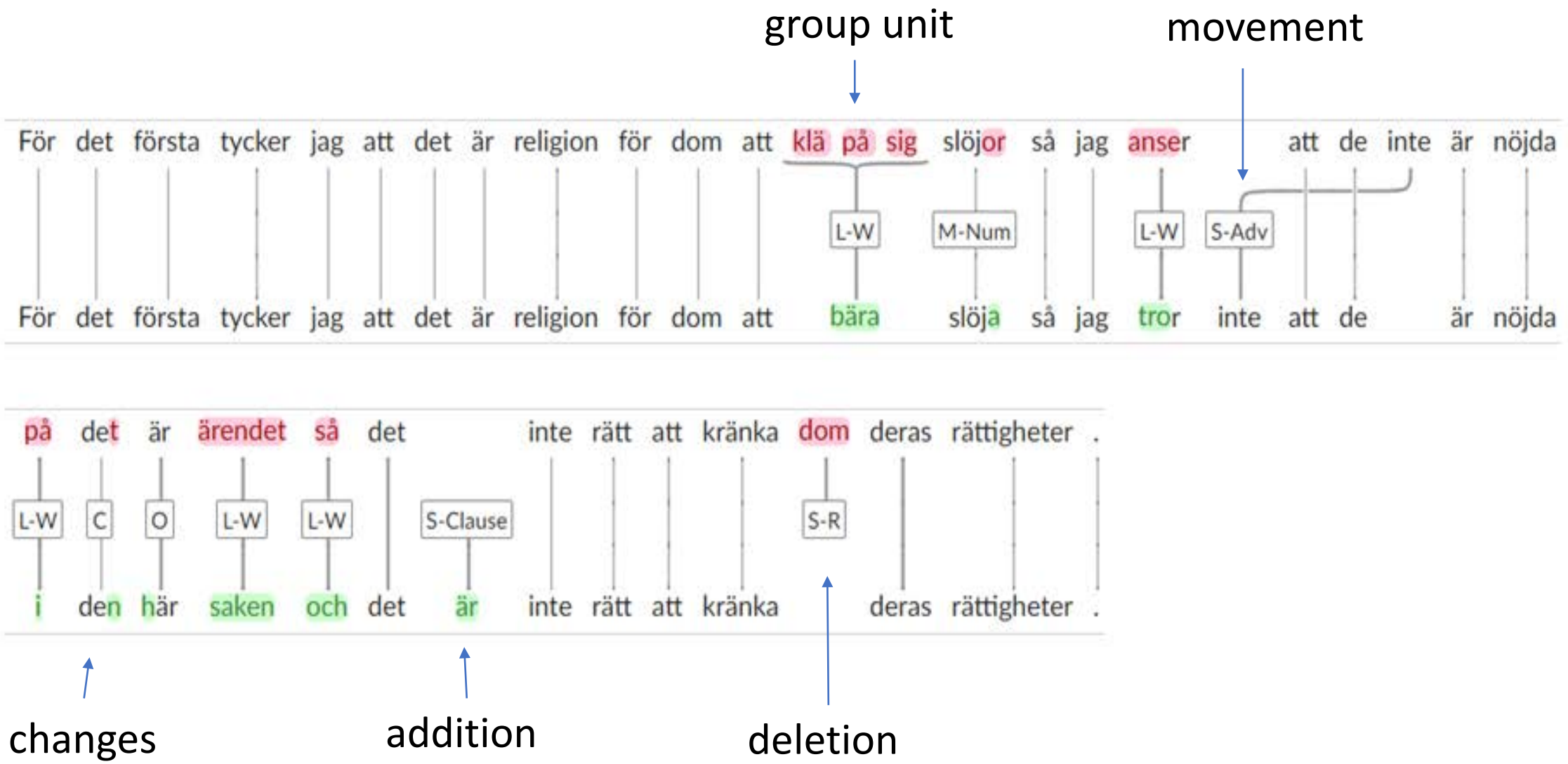
→Highly dependent on the preceding normalization

The purpose of the correction annotation

make the learner corpus searchable for different types of deviations from a standard language norm

Quicker path to finding answers and examples involving specific kinds of norm deviations, such as:

- Do students with a specific mother tongue have more or less difficulties than others with their adverbial placements?
- Examples of problems with choosing the right verb in the semantic field of cognition (*tycka, tänka, tro*).


group unit
movement
För det första tycker jag att det är religion för dom att klä på sig slöjor så jag anser att de inte är nöjda
L-W
M-Num
L-W
S-Adv
För det första tycker jag att det är religion för dom att bära slöja så jag tror inte att de är nöjda
på det är ärendet så det inte rätt att kränka dom deras rättigheter .
L-W
C
O
L-W
L-W
S-Clause
S-R
i den här saken och det är inte rätt att kränka deras rättigheter .
changes
addition
deletion

Some errors are not annotated

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*.

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

  (*'You can do many something* (misspelled)' → *'You can do many things')*

Orthographic

O: spelling

O-Cap: upper/lower case

O-Comp: spaces and hyphens between words

Lexical

L-Der: word formation

L-FL: non-Swedish word, corrected

L-Ref: anaphoric expression

L-W: wrong word

Morphological

M-Adj/adv: adjective corrected to adverb form

M-Case

M-Def: definiteness

M-F: grammatical category kept, form changed

M-Gend: gender

M-Num: number

M-Other

M-Verb: verb forms and tense constructions

Syntactical

S-Adv: adverbial placement

S-Comp: compound vs multi-word exp.

S-Clause: synt. function or hierarchical clause structure

S-Ext: extensive and complex

S-FinV: finite verb placement

S-M: word missing

S-Msubj: subject missing

S-Other

S-R: word redundant

S-Type: change of phrase type/POS

S-WO: word order

Punctuation

P-M: missing

P-R: redundant

P-Sent: sentence segmentation

P-W: wrong

Other

C: consistency

Tags available in Svala, not correction categories

- X – unintelligible string (unchanged or provided with a ”guess” about the intended meaning in the normalization)
- Cit-FL – non-Swedish string **kept** in normalization, i.e. **not** corrected (marked citations, appropriate code switching, etc.)

- OBS! – notes for the annotators
- Com! – comments for the corpus user

- Unid – unidentified correction

Orthographic

O: spelling

O-Cap: upper/lower case

O-Comp: spaces and hyphens between words

Lexical

L-Der: word formation

L-FL: non-Swedish word, corrected

L-Ref: anaphoric expression

L-W: wrong word

Morphological

M-Adj/adv: adjective corrected to adverb form

M-Case

M-Def: definiteness

M-F: grammatical category kept, form changed

M-Gend: gender

M-Num: number

M-Other

M-Verb: verb forms and tense constructions

Syntactical

S-Adv: adverbial placement

S-Comp: compound vs multi-word exp.

S-Clause: synt. function or hierarchical clause structure

S-Ext: extensive and complex

S-FinV: finite verb placement

S-M: word missing

S-Msubj: subject missing

S-Other

S-R: word redundant

S-Type: change of phrase type/POS

S-WO: word order

Punctuation

P-M: missing

P-R: redundant

P-Sent: sentence segmentation

P-W: wrong

Other

C: consistency

Right category, wrong form (M-F)

Nouns:

- **huvudproblemer → huvudproblem, musar → möss**

Verbs:

- jag **lager** mat → jag **lagar** mat
- **böjade → böjde**, **sjungde → sjöng**
- **kunnit → kunnat**, **drickit → druckit**

Adjectives:

- **högare → högre**, **långare → längre**
- **högast → högst**, **ungast → yngst**
- I Sverige är **mer kallt** än i xx → I Sverige är det **kallare** än i xx

Wording (L-W) vs word formation (L-Der)

Wording (L-W): *The root morpheme(s) of the category-defining part of the stem is exchanged*	**Word-formation (L-Der)**: *The root morpheme(s) of the category-defining part of the stem is kept*
1: *transformerade → förändrade*	2: *stressiga → stressade, ändring → förändring, förstöra → störa*
	3 a: *tvångsvenska → tvångssvenska* 3 b: *dagsskolan → förskolan*
4 a: *vårsemestern → vårterminen* 4b: *maktfull → maktfullkomlig*	4 c: *nybyggnad → nybyggd*
5 a: *ställning → inställning*	5 b: *historian → historieskrivningen*
6 a: *basketbollag → basketlag*	6 b: *semesterdag → semester*

Choice of anaphoric expression (L-Ref)

anaphoric expressions corrected in order to have the grammatical form, semantic content and specificity which suits its correlate and its textual position

1. 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

   his_reflexive → his_non-reflexive (’Before the tree was gone they got father’s apples and his wife made juice and sauce of the apples’)
2. Oavsett hur än pappa anstränger honom… → Oavsett hur mycket pappan än anstränger sig …

   him_non-reflexive → himself_reflexive (’regardless of how hard the father tries [lit. strains himself]’)
3. Jag ska resa där ikväll → Jag ska resa dit ikväll

   there_locational → there_directional (’I will go there tonight’)
4. Han blir bortgift mot viljan med sin kusin i pappas land … → Killen blir bortgift mot sin vilja med sin kusin i pappans land …

   he → the guy

Non-Swedish words (L-FL)

Tagging of non-Swedish words which have been corrected during normalization:

1. The form used may be interpreted as a misspelled Swedish word→ O: *kaffee* → *kaffe*; *can* → *kan*
2. The form used may be interpreted as a Swedish word with an incorrect usage of derivational affixes etc. → L-Der: *national helgdag* → *nationell helgdag*

3. Neither a nor b → L-FL: *balcony* → *balkong*

A word in the original text which is identifiable as a Swedish word is **not** tagged with L-FL, even though its usage is clearly influenced by another language:

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

   (*busiga* 'naughty' probably influenced by *busy*)

Compound/multi-word phrase etc. (O-Comp, S-Comp)

O-Comp: *mere orthography*	**S-Comp:** *choice between a compound and a multi-word expression, etc.*
kultur hus → kulturhus *engång → en gång* *tv program → tv-program* *fantasy och äventyrsroman → fantasy- och äventyrsroman*	*det vardagliga livet → vardagslivet* *livsskillnaden → skillnaden i liv* *svenska undervisningar → svenskundervisning* *inblandar → blandar in*
	sociala medier användning → användning av sociala medier *Sveriges bostad → bostäder i Sverige* *dit och hit → hit och dit*

Missing subject (S-Msubj)

1. *Tycker om min plats → Jag tycker om platsen där jag bor*
   'like my place' → 'I like the place where I live'
2. *Regnar ute → Det regnar ute*
   → 'it rains outside'
3. *Det är inte bara arbetet och arbetslivet kan ge stress → Det är inte bara arbetet och arbetslivet som kan ge stress*
   → 'It is not just the_work and the_working_life which can give stress'

The S-Msubj is placed on a *det* which has been inserted as a subject, even when the original clause contains a subject:

4. *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*
   → 'on the other hand are det *people who have so much money that they can buy the whole world'*

Word-order

Three word-order tags, with the following general priority

1. S-Adv: *adverbial moved*
2. S-FinV: *finite verb moved*
3. S-WO: *other word order changes*

Adverbial moved (S-Adv)

Both clause adverbials and VP-modifying adverbials.

1. *Jag **ofta** är vaken länge på kvällarna → Jag är **ofta** vaken länge på kvällarna*

   *→ 'I am **often** awake …'*

2. *Jag är jättetrött eftersom jag sover **inte** på nätterna → Jag är jättetrött eftersom jag **inte** sover på nätterna*

   *→ '…since I **not** sleep …'*

3. *om hur **under 600 år** Finland var en del av Sverige → om hur Finland var en del av Sverige **under 600 år***

   *→ 'about how Finland was a part of Sweden **during 600 years**'*

Finite verb moved (S-FinV)

Almost all instances concern the ordering between the subject and the finite verb (since S-Adv has higher priority than S-FinV):

1. Vor du **bor**? → Var **bor** du?

   → ’Where **live** you?’

2. I morgon jag **åker** → I morgon **åker** jag

   → ’Tomorrow **leave** I’

3. Jag har lärt mig hur **ska** man behandla en människa → Jag har lärt mig hur man **ska** behandla en människa

   → ’I have learnt how one **shall** treat a human_being’

Other word order changes (S-WO)

1. Bostaden min → Min bostad

   ‘the_house my’ → ‘my house’

2. Tidigare Finlands president → Finlands tidigare president

   ’Former Finland’s president’ → ’Finland’s former president’

3. Om man är otrogen, skulle påverkas deras kärleksrelation → Om man är otrogen, skulle deras kärleksrelation påverkas

   ’if one is unfaithful, would be_influenced their relationship’ → ’if one is unfaithful, would their relationship be_influenced’

Change of *fundament* ’topicalized element in pre-finite position’ (**even if it involves an adverbial**), both elements involved in the ”switch” are tagged:

4. Med min familj bor jag → Jag bor med min familj

   ’With my family live I’ → ’I live with my family’

Clause structure (S-Clause)

1. The basic clause type is changed or specified:

   *en fördet är att det finns biblitoteket och det finns många skolor → en fördel är att det finns ett bibliotek och **att** det finns många skolor*

   (addition of subordinating conjunction)

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

   *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***

   (*'how Swedish is situated' → 'the situation for Swedish'*)

3. The structure of a clause is changed in a way which involves changing the primary syntactic function of one or more of the words/phrases involved:

   *I texten **sägs det av henne** att … → I texten **säger hon** att ..*

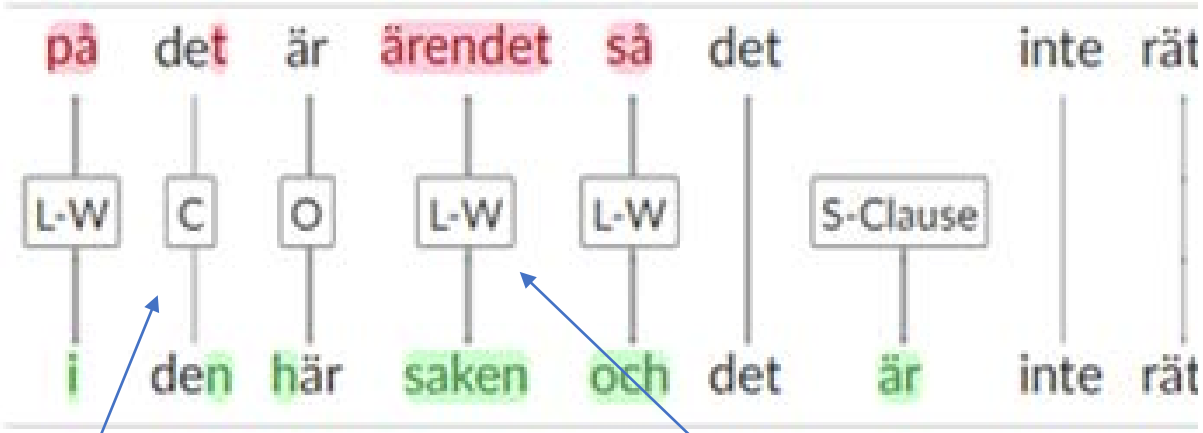
   (*'it is said by her' → 'she says'*)

Extensive syntactic revision (S-Ext)

Extensive, often complex, correction. Structure **created** rather than **corrected**. The range of possible normalizations is especially wide, and the normalization therefore involves a greater than usual element of subjectivity.

1. **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
2. 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
3. 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**
4. Men **har det hänt något problem med mig** → Men **jag har fått ett problem**

Consistency corrections (C)



Article changed to agree with the non-neuter word in the target text.

Neuter word changed to non-neuter word.

Overall tag statistics

1: M (all)	**8006**
M-Def	2968
M-Verb	2277
M-Num	1031
M-Gend	953
M-F	300
M-Case	257
M-Other	116
M-Adj/adv	104

2: S (all)	**7696**
S-M	1905
S-R	1423
S-Clause	1128
S-Adv	785
S-FinV	701
S-Type	595
S-Msubj	434
S-Ext	310
S-WO	199
S-Comp	136
S-Other	80

3: L (all)	**4875**
L-W	3442
L-Der	720
L-Ref	624
L-FL	89

4: O (all)	**4381**
O	3269
O-Cap	566
O-Comp	546

5: P (all)	**2754**
P-M	1834
P-R	444
P-W	437
P-Sent	39

6: C	**1208**