UPPSALA
UNIVERSITET

Annotating Errors in Student Texts: First Experiences and Experiments

Sara Stymne, Eva Pettersson, Beáta Megyesi & Anne Palmér

Department of Linguistics and Philology
Department of Scandinavian Languages
Uppsala University

2017-05-22

UPPSALA
UNIVERSITET

Goals

- The creation of an annotation layer for word-based writing errors for a corpus of student writings in Swedish
- Human annotation of word-based errors focusing on spelling, split compounds, merged words, and simple grammatical errors
- Two initial experiments, exemplifying how this subset of the corpus can be used

UPPSALA
UNIVERSITET

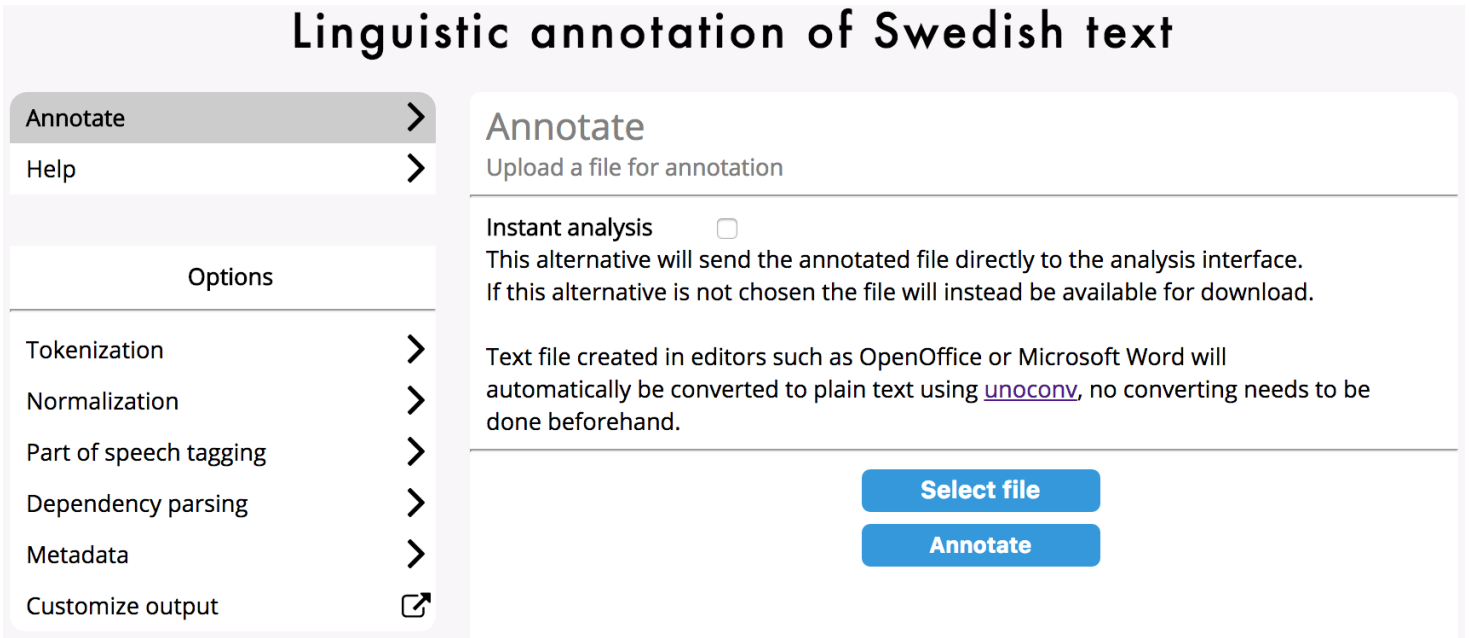
Corpus

- The Uppsala Corpus of Student Writings (Megyesi et al., 2016)
- 2,500 essays written as part of Swedish national tests in the subjects Swedish and Swedish as a second language
- 1.5 million tokens
- Written by students in different grades,
  - Compulsory school: year 3, (5), 6, 9 (age 9, (11), 12, 15)
  - Upper secondary school; year 1 and 3 (age 16, 19).

UPPSALA
UNIVERSITET

Automatic annotation

- Automatically in a pipeline using SweGram (Näsman et al., Nodalida 2017)
- Online tool for automatic analysis of Swedish texts



UPPSALA
UNIVERSITET

Automatic annotation

Preprocessing → Tokenization → Normalization → PoS tagging → Parsing

TEXT ID	ID	FORM	NORM	LEMMA	U-POS	CPOS	C-FEATS	U-FEATS	HEAD	DEPREL	Translation
2.2	1	Den		den	DET	DT	UTR|SIN|DEF	Definite=Def|Gender=Com|Number=Sing	3	det	The
2.2	2	kalla		kall	ADJ	JJ	POS|UTR/NEU|SIN|DEF|NOM	Case=Nom|Definite=Def|Degree=Pos|Number=Sing	3	amod	cold
2.2	3	vinden		vind	NOUN	NN	UTR|SIN|DEF|NOM	Case=Nom|Definite=Def|Gender=Com|Number=Sing	4	nsubj	wind
2.2	4	slåg	slog	slå	VERB	VB	PRT|AKT	Mood=Ind|Tense=Past|VerbForm=Fin|Voice=Act	0	root	hit
2.2	5	mot		mot	ADP	PP	_	_	7	case	[against]
2.2	6	mina		min	DET	PS	UTR/NEU|PLU|DEF	Definite=Def|Number=Plur|Poss=Yes	7	nmod:poss	my
2.2	7	kinder		kind	NOUN	NN	UTR|PLU|IND|NOM	Case=Nom|Definite=Ind|Gender=Com|Number=Plur	4	nmod	cheeks
2.2	8	.		.	PUNCT	MAD	_	_	4	punct	.

[Extended version of the CoNLL-U format]

UPPSALA
UNIVERSITET

Error annotation

- Human annotation of word-based errors
- Training and test sets
- Goals
  - Enable the training and evaluation of NLP tools
  - Enable the study of errors in student writing for different groups

UPPSALA
UNIVERSITET

Texts

		Training essays			Test essays		
Level	Age	Sw	SwSL	Tokens	Sw	SwSL	Tokens
C-3	9	50	50	13,624	36	19	4,831
C-5	11	–	–	–	29	12	6,962
C-6	12	50	49	37,718	17	7	8,554
C-9	15	49	52	54,970	30	10	17,143
US-1	16	0	50	25,087	15	4	7,719
US-3	18	–	–	–	12	4	13,493
Total		149	201	131,399	139	56	58,702

UPPSALA
UNIVERSITET

Error Categories

- Spelling
  - Casing
- Split compounds
- Merged words
- Simple grammatical errors

UPPSALA
UNIVERSITET

Spelling

- Mainly misspelled words
- No differentiation due to typos/slip of pen and spelling competence errors
- Spelling norms based on SAOL
- Informal spelling variants are allowed (based on SAOL)
- Both real word errors and errors resulting in incorrect words
- Punctuation and upper/lower case is also corrected in this category
- Foreign words corrected based on the language, and marked (if possible, mainly English)

UPPSALA
UNIVERSITET

Spelling errors – Examples

kännislor	känslor	('feeling')
ända	enda	('end'/'only')
dåm	dom	('they')
carolina	Carolina	
tex	t.ex. ('e.g.')	
sand-låda	sandlåda	('sand box')
back flipp	back flip	Foreign

UPPSALA
UNIVERSITET

Spelling errors – Annotation

		Original	Auto	Human	Comment	Gloss
5.6	1	När	När	När		*When*
5.6	2	Bläckfisken	Bläckfisken	bläckfisken		*octopus*
5.6	3	Mar	Mar	mår		*feels*

- Only annotated by changing the misspelt word to correct spelling/casing

UPPSALA
UNIVERSITET

Casing

- In our analysis we show errors with only differences in casing separately from other spelling errors
- They are not marked any differently in the annotation, but easily identifiable

UPPSALA
UNIVERSITET

Split compounds

- Words that should have been written as a closed compound, but are written as separate words
- Also other types of split words

jätte bra	(’very good’)
jete god	(’vary good’)
för svar	(’def ence’)
spela de	(’play ed’)
schim- pans	(’chimp- anzee’)

UPPSALA
UNIVERSITET

Split Compounds – Annotation

2.21	7-8		favoritkläder	
2.21	7	favorit		*favorite*
2.21	8	kläder		*clothes*
5.5	14-15		bläckfärg	*ink color*
5.5	14	bläck		*ink*
5.5	15	fäj	färg	*color*

- Annotated by adding new lines, showing indeces
- Spelling errors and grammar errors are annotated for parts if needed

UPPSALA
UNIVERSITET

Merged Words

- Words that are written as one word, but should have been several words
- In some sense the opposite of split compounds

ihela	i hela	('in+whole')
påväg	på väg	('on+road')
hovar	hon var	('she+was')
justet	just det	('just+so')
kroppen.Taggarna	kroppen . Taggarna	('body . Pegs')

UPPSALA
UNIVERSITET

Merged Words – Annotation

2.1	8.1		i	*in*
2.1	8.2		hela	*whole*
2.1	8	ihela		*in+whole*
2.20	4.1	ho	hon	*she*
2.20	4.2		var	*was*
2.20	4	hovar		**hoofs*

- Annotated by adding new lines, with sub-indeces
- Spelling errors and grammar errors are annotated for parts if needed

UPPSALA
UNIVERSITET

Simple grammatical errors

- Diverse category consisting of different types of errors
  - Morphological errors
  - Wrong word (switch of words)
  - Words marked for effect (*sååååå*)
  - Extra words
  - Invented words (*retig*, marked)
- No further sub-classification of errors (nearly)
- Work in progress!

UPPSALA
UNIVERSITET

Simple Grammatical Errors – Annotation

5.6	2	Bläckfisken	Bläckfisken	bläckfisken		*octopus*
5.6	3	Mar	Mar	mår		*feels*
5.6	4	dolig	dålig	dåligt	x	*bad*

- Corrected, and x in comment field
- No specific marking of combinations of spelling and grammar errors

UPPSALA
UNIVERSITET

Errors that are not annotated

- Missing words
- Word order errors
- Discourse-level errors
- ...

UPPSALA
UNIVERSITET

Problematic Cases

- Borderline spelling/grammar
  - *hon to*
  - *igår svara jag*

UPPSALA
UNIVERSITET

Problematic Cases

- Borderline spelling/grammar
  - *hon to* → Spelling
  - *igår svara jag* → Grammar

UPPSALA
UNIVERSITET

Problematic Cases

- Borderline spelling/grammar
  - *hon to* → Spelling
  - *igår svara jag* → Grammar
- Long distance errors
  - *Bläckfisken är blå och* ***de*** *blir . . .*
  - *. . . ofta* ***skrämda***

UPPSALA
UNIVERSITET

Problematic Cases

- Borderline spelling/grammar
  - *hon to* → Spelling
  - *igår svara jag* → Grammar
- Long distance errors
  - *Bläckfisken är blå och* ***de*** *blir . . .* → *den*
  - *. . . ofta* ***skrämda***

UPPSALA
UNIVERSITET

Problematic Cases

- Borderline spelling/grammar
  - *hon to* → Spelling
  - *igår svara jag* → Grammar
- Long distance errors
  - *Bläckfisken är blå och* ***de*** *blir . . .* → *den*
  - *. . . ofta* ***skrämda*** → *skrämd* FÖLJD

UPPSALA
UNIVERSITET

Annotation Process

- Guidelines created to describe the annotation
- CoNLL format in text editor or Excel
- 3 annotators in final phase, native speakers
  - Computational linguist
  - Student of Swedish
  - Research Assistant in Swedish

UPPSALA
UNIVERSITET

Inter-annotator agreement

	All		−correct	
	Agree	Kappa	Agree	Kappa
A1/A2	.97	.96	.72	.65
A1/A3	.97	.96	.70	.62
A2/A3	.97	.97	.72	.66

UPPSALA
UNIVERSITET

Inter-annotator agreement

	Co	Spe	Gr	Spl	Me	Ca
Correct (Co)	2,138		23			2
Spelling (Spe)	2	73	3			
Grammar (Gr)	15	4	65			
Split (Spl)	3			13		
Merged (Me)	1				7	
Casing (Ca)	17					23

UPPSALA
UNIVERSITET

Error Statistics

	Training	Test
Total tokens	132,348	59,297
Total	7,189	2,074
Spelling	2,826	1,205
Grammar	2,465	336
Split	548	218
Merged	192	73
Casing	1,158	242
Split+spelling	123	35
Split+grammar	29	1
Merged+spelling	46	24
Merged+grammar	9	2

UPPSALA
UNIVERSITET

Pilot experiments

- Spelling correction
  - Levenshtein distance (LD) + dictionary mapping
  - **Baseline**: unweighted LD + standard dictionary
  - **Refined**: weighted LD + standard dictionary + human annotations

UPPSALA
UNIVERSITET

Pilot experiments

- Spelling correction
  - Levenshtein distance (LD) + dictionary mapping
  - **Baseline**: unweighted LD + standard dictionary
  - **Refined**: weighted LD + standard dictionary + human annotations
- Comparison of tagging and parsing labels of raw and corrected student texts

UPPSALA
UNIVERSITET

Spelling results – Overall

	Precision	Recall	Accuracy
Baseline	84.9	57.6	70.9
Refined	80.9	63.5	78.2

UPPSALA
UNIVERSITET

Spelling results – Groups

	Prec	Recall	Acc
Swedish			
in-domain data	82.1	62.0	75.4
all data	77.9	64.2	80.7
Swedish as a second language			
in-domain data	86.2	61.9	72.0
all data	86.3	62.6	73.3
Younger students			
in-domain data	91.0	64.9	76.0
all data	87.4	65.2	76.9
Older students			
in-domain data	72.8	59.5	82.8
all data	70.0	60.2	84.6
All texts	80.9	63.5	78.2

UPPSALA
UNIVERSITET

Spelling analysis

- Dictionary mapping too crude
- Real word errors not handled
- Errors which needed many edits (we only allowed 1)
- Some inconsistencies between training and test data
- Errors do indeed vary for different ages and Sw/SwSL

UPPSALA
UNIVERSITET

Tagging and Parsing Consistency

	POS		Labels		Heads	
Correct	447	(.4)	2,989	(2)	9,551	(8)
Spelling	942	(34)	994	(36)	887	(32)
Grammar	434	(16)	726	(26)	749	(27)
Split	109	(20)	247	(45)	316	(58)
Merged	108	(57)	144	(75)	139	(73)
Casing	96	(9)	138	(12)	209	(19)

UPPSALA
UNIVERSITET

Tagging and Parsing Confusions

POS-tag		Dependency label	
VERB-NOUN	170	nsubj-dobj	130
ADV-NOUN	152	dobj-nsubj	108
PRON-DET	90	nmod-dobj	105
ADV-ADJ	90	dobj-nmod	91
AUX-VERB	88	nsubj-nmod	72
PROPN-NOUN	85	root-advcl	70
ADJ-NOUN	81	root-nsubj	66
VERB-ADJ	81	nsubj-det	61

UPPSALA
UNIVERSITET

Conclusion and Future Work

- We have presented a useful resource of human annotations of errors
- Annotated data important both for NLP and writing studies
- Many issues left for future work
  - Annotation consistency
  - Error typology, especially for grammar errors
  - Alignment and visualization of original and corrected versions
  - Develop/adapt tools for identification and correction of errors
  - Annotation of more complex errors