SWE-CLARIN

UPPSALA UNIVERSITET

SWEGRAM

Annotering och analys av svenska texter

Beáta Megyesi[1], Anne Palmér[2], Jesper Näsman[1]

[1]Institutionen för lingvistik och filologi

[2]Institutionen för nordiska språk

Uppsala universitet

Förord

Dokumentet syftar till att beskriva verktyget SWEGRAM med vars hjälp du kan genomföra automatisk annotering och lingvistisk analys av svenska och engelska texter eller skapa din egen, lingvistiskt annoterade textsamling, en så kallad korpus. Vi presenterar verktygets beståndsdelar och ger förslag på hur man kan genomföra storskalig, empirisk språklig analys med hjälp av verktyget.

SWEGRAM har utvecklats i samarbete mellan Institutionen för lingvistik och filologi och Institutionen för nordiska språk vid Uppsala universitet inom ramen av SWE-CLARIN-projektet. Det långsiktiga målet är att göra språkbaserade material tillgängliga som primära forskningsdata för humanistisk och samhällsvetenskaplig forskning med hjälp av avancerade bearbetningsverktyg för text och tal.

Vi vill härmed tacka alla som bidragit till verktygets utveckling och har gett synpunkter på denna manual.

Innehållsförteckning

Förord **3**

Innehållsförteckning **4**

1 Introduktion **6**

2 Annotering **7**

2.1 Format **8**

2.2 Annoteringskedja **11**

2.2.1 Formatering 11

2.2.2 Tokenisering och meningssegmentering 12

2.2.3 Normalisering: rättstavning och särskrivningar 13

2.2.4 Lemmatisering 13

2.2.5 Ordklassanalys 14

2.2.6 Syntaktisk analys 17

2.3 Olika modeller för automatisk analys **20**

2.4 Rättning av automatisk analys **20**

2.5 Spara din annoterade fil lokalt och exportera till Microsoft Excel **22**

3 Analys **22**

3.1 Metadata och filtrering **23**

3.2 Statistik och kvantitativa mått **24**

3.2.1 Allmän statistik 24

3.2.2 Ordklasstatistik 26

3.2.3 Läsbarhetsmått 27

3.2.3.1 Lix 28

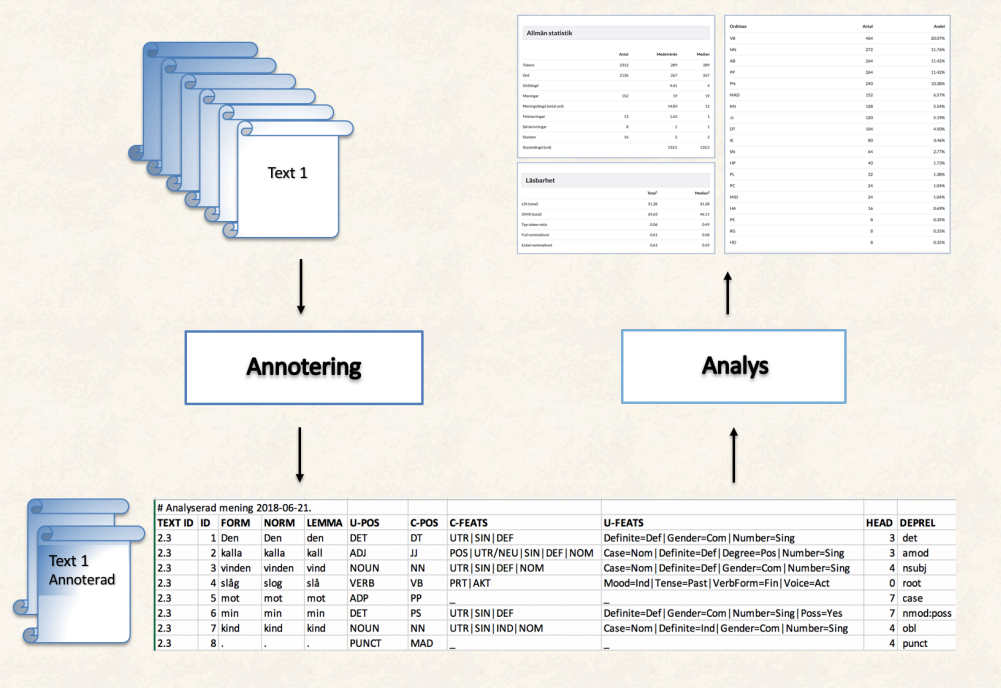
3.2.3.2 Typ-token ratio 28

3.2.3.3 Ovix 29

3.2.3.4 Enkel nominalkvot....30
3.2.3.5 Full nominalkvot....30
3.2.4 Frekvenser 31
3.3 Visualisera text och analys 32
3.4 Exportera statistik 34
4 Skapa din egen korpus (textsamling) med egen metadata 36
4.1 Exempel på förarbete med metadata 37
5 Om verktyget 39

1 Introduktion

SWEGRAM är ett webbaserat verktyg som ger dig möjlighet att annotera och analysera svenska eller engelska texter. Du kan ladda upp en eller flera texter för att få dessa annoterade lingvistiskt med morfologisk och syntaktisk information. De lingvistiskt annoterade texterna kan sedan användas för att genomföra kvantitativ analys för att få fram statistik om texterna med avseende på t.ex. meningslängd, antal ord, olika läsbarhetsmått, ordklasstatistik, frekvenser på ordformer, basform och felstavade ord. Figur 1 illustrerar verktygets två olika delar som beskrivs mer ingående i denna manual.



TEXT ID	ID	FORM	NORM	LEMMA	U-POS	C-POS	C-FEATS	U-FEATS	HEAD	DEPREL
2.3	1	Den	Den	den	DET	DT	UTR|SIN|DEF	Definite=Def|Gender=Com|Number=Sing	3	det
2.3	2	kalla	kalla	kall	ADJ	JJ	POS|UTR/NEU|SIN|DEF|NOM	Case=Nom|Definite=Def|Degree=Pos|Number=Sing	3	amod
2.3	3	vinden	vinden	vind	NOUN	NN	UTR|SIN|DEF|NOM	Case=Nom|Definite=Def|Gender=Com|Number=Sing	4	nsubj
2.3	4	slåg	slog	slå	VERB	VB	PRT|AKT	Mood=Ind|Tense=Past|VerbForm=Fin|Voice=Act	0	root
2.3	5	mot	mot	mot	ADP	PP	_	_	7	case
2.3	6	min	min	min	DET	PS	UTR|SIN|DEF	Definite=Def|Gender=Com|Number=Sing|Poss=Yes	7	nmod:poss
2.3	7	kind	kind	kind	NOUN	NN	UTR|SIN|IND|NOM	Case=Nom|Definite=Ind|Gender=Com|Number=Sing	4	obl
2.3	8	.	.	.	PUNCT	MAD	_	_	4	punct

Figur 1. SWEGRAM – komponenter.

Med hjälp av verktyget kan du enkelt skapa din egen lingvistiskt annoterade textsamling, en så kallad korpus, och få fram statistik om texternas språkliga egenskaper. Verktyget kräver inte någon kunskap om automatisk textbearbetning och inte heller någon programmeringskunskap. SWEGRAM är fritt tillgängligt för alla att använda och kan nås via SWEGRAMs webbsida: http://cl.lingfil.uu.se/swegram. Uppladdade texter sparas inte och tas bort från servern efter avslutad analys. Detta innebär att du kan ladda upp vilka texter du önskar utan att vi sparar information om vem du är eller vilka texter som du arbetat med. Detta innebär dock att du inte kan lämna programmet och sedan återgå till samma fil utan att ladda upp den på nytt.

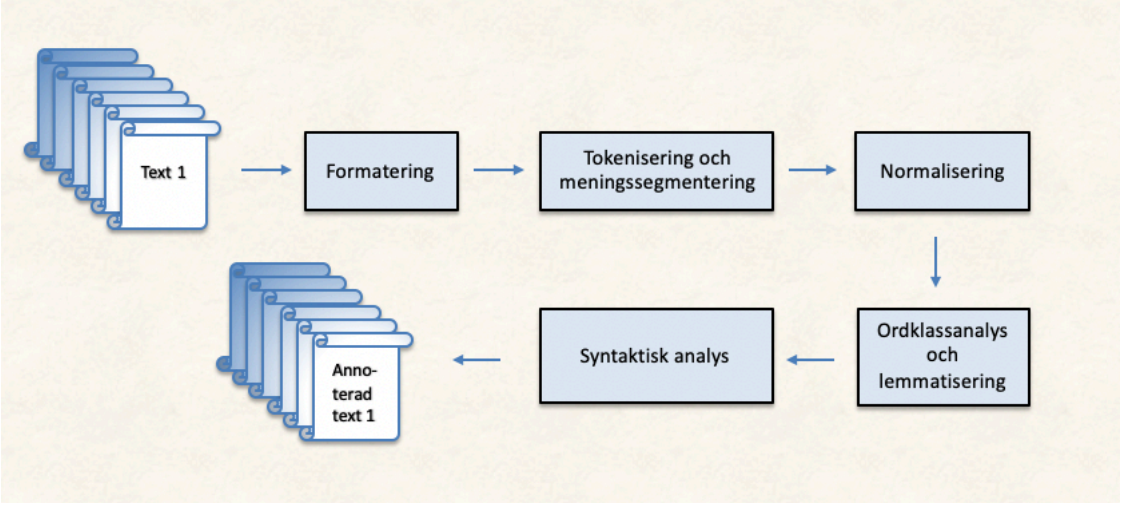
I manualen beskriver vi hur du kan använda verktyget och dess olika funktioner för svenska texter. Det finns också hjälpavsnitt i SWEGRAM som kortfattat beskriver verktygets olika delar

och funktioner. Om du saknar information i denna manual eller har frågor, kontakta oss: swegram@stp.lingfil.uu.se.

2 Annotering

För att en text ska kunna annoteras och analyseras kan den antingen klistras in i webbläsaren eller laddas upp enskilt eller tillsammans med andra texter. Varje text bearbetas lingvistiskt av verktygets olika komponenter. För att automatiskt bearbeta och annotera texter använder vi avancerade (s.k. state-of-the-art) verktyg med dokumenterad hög prestanda, vilka har utvecklats inom datorlingvistik/språkteknologi.

Annoteringen bygger på en verktygskedja för automatisk analys av svenska. Verktygskedjan illustreras i detalj i Figur 2.



Figur 2. Annoteringsprocessen och komponenter i SWEGRAM.

Textfiler kan vara skrivna med olika teckenkodningar. För att filen ska kunna bearbetas av SWEGRAM krävs det att texten är skriven i en universell teckenkodningUnicode UTF-8. Det första steget i annoteringsprocessen är därför att formatera om den uppladdade filen till detta format. Orden separeras sedan från skiljetecken genom en process som kallas för tokenisering och texten segmenteras i meningar så att varje ord står på en egen rad och varje mening avslutas med en tom rad. Felstavade ord rättas med avseende på stavfel, och särskrivna ord identifieras och skrivs ihop av normaliseringsverktyget. Den rättade, normaliserade versionen av texten analyseras sedan med hjälp av en ordklasstaggare vars uppgift är att annotera varje ord och skiljetecken med dess korrekta ordklass och morfologisk information. Lemmat (basformen) för varje ord identifieras sedan av en komponent som kallas för lemmatiserare. Slutligen analyseras meningarna syntaktiskt med hjälp av en parser, ett verktyg som

identifierar ordens relationer till varandra och ordets funktion i meningen. De komponenter som ingår i annoteringen beskrivs mer i detalj i avsnitt 2.2.

Om du har egna kommentarer i en fil som du inte vill ska annoteras och analyseras av SWEGRAM ska du inleda dessa med hashtagg (#) på varje rad för kommentaren. Kommentarfält kan också användas till att ange metadatainformation för påföljande text, som illustreras i Figur 3.

Resultatet av den analyserade texten kan du se direkt på SWEGRAMS webbsida under *Visualisera text* där du kan se den uppladdade textens olika egenskaper med avseende på ord och dess ordklass. Du kan också välja att ladda ner den analyserade texten i textformat (.txt) för att spara på din dator. I följande avsnitt ges en beskrivning av annoteringen och analysen.

2.1 Format

Texterna som annoteras med hjälp av verktygskedjan representeras i ett unikt format, och annoteringen på olika lingvistiska nivåer sker enligt en viss standard för svenska.

För teckenkodning använder vi Unicode (UTF-8). För att representera annoteringen på olika lingvistiska nivåer har vi valt ett enkelt format som följer internationell standard, det så kallade CoNLL-U formatet. Varje token, d.v.s. ord och skiljetecken separerade, står på egen rad med dess analys och varje ny mening inleds med en tom rad.

Varje token får en analys på olika lingvistiska nivåer och dessa representeras på samma rad som själva tokenet. De olika lingvistiska analyserna representeras kolumnvis och kolumnerna är separerade med tabb. Kolumnerna anger dels ID-nummer för stycke, mening och token, dels den lingvistiska analysen. Den lingvistiska analysen representeras dels på ordnivå genom ordklassannotering och morfologisk information, dels på meningsnivå genom syntaktisk analys. Ordklassannotering består av två typer: universella ordklasser (Nivre et al., 2016) samt ordklasser i Stockholm-Umeå Corpus (Ejerhed et al., 1992; Gustafson-Capková & Hartmann, 2006). Dessa beskrivs mer detaljerat i avsnitt 2.2.5 om Ordklassanalys och i avsnitt 2.2.6 om Syntaktisk analys.

TEXT ID	Stycke-Mening index, heltal fr.o.m. 1 för varje nytt stycke och mening.
ID	Token-index, heltal fr.o.m. 1 för varje ny mening; kan vara ett intervall för ursprungligen särskrivna ord som har rättats.
FORM	Ordform eller skiljetecken, s.k. token.
NORM	Rättad/normaliserad token t.ex. vid felstavning.
LEMMA	Lemma, d.v.s. ordets basform.
U-POS	Ordklasstagg baserat på Universella ordklasser.
C-POS	Ordklasstagg baserat på Stockholm-Umeå Corpus ordklasser.
C-FEATS	Lista av morfologiska särdrag enligt Stockholm-Umeå Corpus; ”_” om särdrag saknas.
U-FEATS	Lista av morfologiska särdrag enligt Universella ordklasser; ”_” om särdrag saknas.
HEAD	Ordets huvud, som är ett värde av ID eller noll (0) om ordet är meningens rot (root).

DEPREL	Dependensrelation till ordets huvudord HEAD baserad på Universella dependensrelationer. Om ordet inte har något huvudord anges rot (root) som dependensrelation (root iff HEAD = 0).
MISC	Övrig annotering.

Tabell 1 ger en kortfattad sammanfattning av annoteringsrepresentationen.

TEXT ID	Stycke-Mening index, heltal fr.o.m. 1 för varje nytt stycke och mening.
ID	Token-index, heltal fr.o.m. 1 för varje ny mening; kan vara ett intervall för ursprungligen särskrivna ord som har rättats.
FORM	Ordform eller skiljetecken, s.k. token.
NORM	Rättad/normaliserad token t.ex. vid felstavning.
LEMMA	Lemma, d.v.s. ordets basform.
U-POS	Ordklasstagg baserat på Universella ordklasser.
C-POS	Ordklasstagg baserat på Stockholm-Umeå Corpus ordklasser.
C-FEATS	Lista av morfologiska särdrag enligt Stockholm-Umeå Corpus; ”_” om särdrag saknas.
U-FEATS	Lista av morfologiska särdrag enligt Universella ordklasser; ”_” om särdrag saknas.
HEAD	Ordets huvud, som är ett värde av ID eller noll (0) om ordet är meningens rot (root).
DEPREL	Dependensrelation till ordets huvudord HEAD baserad på Universella dependensrelationer. Om ordet inte har något huvudord anges rot (root) som dependensrelation (root iff HEAD = 0).
MISC	Övrig annotering.

Tabell 1. Representation av annotering.

Output från den lingvistiska annoteringen illustreras i Figur 3 för meningen ”*Den kalla vinden slåg mot min kind.*” som också inkluderar ett felstavat ord *slåg* som SWEGRAMs normaliseringsmodul rättar till den korrekta stavningen *slog*. Kommentaren om texten som inleds med hashtaggen (# Analyserad mening 2018-06-21.) analyseras inte av verktyget. Raden efter kommentarsfältet markerad i fet stil lades till i detta exempel för att illustrera vilken typ av annotering respektive kolumn avser.

# Analyserad mening 2018-06-21.										
TEXT ID	**ID**	**FORM**	**NORM**	**LEMMA**	**U-POS**	**C-POS**	**C-FEATS**	**U-FEATS**	**HEAD**	**DEPREL**
2.3	1	Den	Den	den	DET	DT	UTR|SIN|DEF	Definite=Def|Gender=Com|Number=Sing	3	det
2.3	2	kalla	kalla	kall	ADJ	JJ	POS|UTR/NEU|SIN|DEF|NOM	Case=Nom|Definite=Def|Degree=Pos|Number=Sing	3	amod
2.3	3	vinden	vinden	vind	NOUN	NN	UTR|SIN|DEF|NOM	Case=Nom|Definite=Def|Gender=Com|Number=Sing	4	nsubj
2.3	4	slåg	slog	slå	VERB	VB	PRT|AKT	Mood=Ind|Tense=Past|VerbForm=Fin|Voice=Act	0	root
2.3	5	mot	mot	mot	ADP	PP	_	_	7	case
2.3	6	min	min	min	DET	PS	UTR|SIN|DEF	Definite=Def|Gender=Com|Number=Sing|Poss=Yes	7	nmod:poss
2.3	7	kind	kind	kind	NOUN	NN	UTR|SIN|IND|NOM	Case=Nom|Definite=Ind|Gender=Com|Number=Sing	4	obl
2.3	8	.	.	.	PUNCT	MAD	_	_	4	punct

Figur 3. Output – Lingvistisk annotering.

Kolumn 1 anger ID-nummer för stycke följt av mening (TEXT ID). I vårt exempel ovan förekommer den aktuella meningen i stycke nummer två och är den tredje meningen i stycket (2.3). Den andra kolumnen (ID) visar positionen för varje token i meningen. I tredje kolumnen (FORM) återges ursprungstexten skriven av skribenten token för token, såväl ord som skiljetecken. Den fjärde kolumnen (NORM) återger den normaliserade texten, med eventuellt felstavade ord rättade. I exemplet ovan rättas ordet *slåg* till *slog* av SWEGRAMs normaliseringsmodul. I femte kolumnen (LEMMA) anges basformen för varje token. Därefter följer den lingvistiska analysen med avseende på ordklass och syntaktisk information. Kolumnen U-POS återger ordklassen för varje token enligt den universella ordklassuppsättningen som definieras i Universal dependencies (UD). Kolumnen C-POS anger också ordklassannotering, men baserat på Stockholm-Umeå Corpus (SUC) 3.0 istället där

ordklasstaggar utgör de två första tecknen i SUC:s tagguppsättning (Gustafson-Capková & Hartmann, 2006). De efterföljande två kolumnerna anger morfologiska särdrag, kolumn C-FEATS enligt SUC-taggar och kolumn U-FEATS enligt UD-taggar. Kolumnerna HEAD och DEPREL återger den syntaktiska analysen baserad på universella dependenser. I HEAD anges ordets huvudord i meningen och i DEPREL dess syntaktiska funktion.

För att illustrera annoteringen går vi igenom meningen i detalj.

- Meningen förekommer i andra stycket i texten och det är den tredje meningen i stycket (2.3) som anges i kolumnen TEXT ID för varje token i meningen.
- Ordet *Den* är det första ordet i meningen med i ID nummer 1. Ursprungsformen är *Den* som anges i kolumnen FORM och eftersom ordet är rättstavat anges samma form i kolumnen NORM. Basordformen för ordet *Den* är *den* som återges i kolumnen för LEMMA. Ordet är en determinerare (artikel) och annoteras som *DET* enligt den universella tagguppsättningen U-POS och som *DT* enligt SUC:s tagguppsättning C-POS. Den morfologiska analysen för ordet är utrum avseende genus (UTR resp. Gender=Com), singular avseende numerus (SIN resp. Number=Sing) och bestämd form (”definit” på engelska) avseende species (DEF resp. Definite=Def), vilket anges enligt SUC-taggar C-FEATS som UTR|SIN|DEF och enligt UD-taggar U-FEATS som Definite=Def|Gender=Com|Number=Sing. Ordet *Den* har ord nummer 3 i meningen som sitt huvudord, d.v.s. ordet *vinden,* och *Den* har determinerare (*det)* som dependensrelation (DEPREL) i satsen.
- Nästa ord i meningen är ordet *kalla* som är ord nummer 2. Dess ursprungsform FORM är *kalla*. Ordet är rättstavat av skribenten och får samma form i kolumnen NORM. Basordformen LEMMA är *kall*. Ordet är ett adjektiv som anges som *ADJ* enligt universella dependenser (U-POS) och *JJ* enligt SUC-taggar (C-POS). Ordet är böjt morfologiskt som positiv, utrum eller neutrum, singular, bestämd form i nominativ som motsvarar SUC-taggen POS|UTR/NEU|SIN|DEF |NOM och som UD-taggen Case=Nom|Definite=Def|Degree=Pos|Number=Sing. Ordet *kalla* har ord nummer 3 (*vinden*) som huvudord och är adjektivisk modifierare (*amod*) till sin funktion.
- Ord nummer 3 är *vinden* som är korrekt stavat och dess lemma är *vind*. Det är ett substantiv (NOUN enligt SUC-tagg och NN enligt U-POS) i utrum, singularis, bestämd form i nominativ UTR|SIN|DEF |NOM enligt SUC-uppmärkning och Case=Nom|Definite=Def|Gender=Com|Number=Sing enligt UD:s annoteringssätt. Ordet *vinden* har det efterföljande ordet, ord nummer 4 (*slog)* som sitt huvudord och utgör subjektet NSUBJ i satsen.
- Ord nummer 4 *slåg* (FORM) har felstavats av skribenten och rättats till *slog* av SWEGRAMS normaliseringsverktyg, vilket anges i kolumnen NORM. Ordets lemma är *slå* och är ett verb (VERB resp. VB) i indikativ, preteritum (*past* på engelska), finit och aktiv form (PRT|AKT enligt SUC resp. Mood=Ind|Tense=Past|VerbForm=Fin|Voice=Act enligt UDs annotering). Eftersom

det är det finita verbet i satsen saknar ordet huvudord (0) och utgör roten (ROOT) för hela meningen.

- Ord nummer 5 är *mot* med samma basform och är preposition (ADP resp. PP) utan morfologiska särdrag (”_”). Dess huvudord är ord nummer 7 i satsen (*kind*) och fungerar som kasus (CASE).
- Ord nummer 6 är *min* med basformen *min* som är determinerare (DET resp. PS) i utrum, singularis i bestämd och possessiv form (UTR|SIN|DEF resp. Definite=Def|Gender=Com|Number=Sing|Poss=Yes). Dess huvudord är ord nummer 7 i satsen (*kind)* och utgör possessiv modifierare (NMOD:POSS) till efterföljande substantiv.
- Ord nummer 7 är *kind* med basformen *kind* som är ett substantiv (NOUN resp. NN) i utrum, singularis, obestämd form i nominativ (UTR|SIN|IND |NOM resp. Case=Nom|Definite=Ind|Gender=Com |Number=Sing). Dess huvudord är det finita verbet *slog* som är ord nummer 4 och fungerar som oblikt objekt (OBL).
- Meningen avslutas med tokenet ”.” som är det sista elementet i satsen med ID nummer 8 och är skiljetecken (PUNCT resp. MAD) utan morfologiska särdrag (-). Dess huvudord är det finita verbet i satsen, ord nummer 4 med skiljetecken (PUNCT) som funktion.

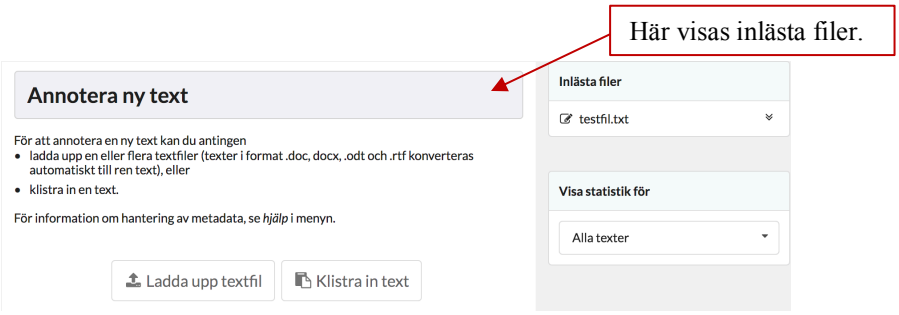
2.2 Annoteringskedja

Annoteringen genomförs automatiskt med hjälp av en verktygskedja som syftar till uppmärkning av svenska texter på ord- och meningsnivå. När du önskar att annotera en eller flera texter är grundinställningen i verktyget att alla verktyg tillämpas. Du kan också välja att utesluta vissa moduler i verktygskedjan om du vill låta verktyget arbeta med enskilda analysdelar. Hur du kan göra det beskrivs i avsnitt 2.4.

2.2.1 Formatering

Du kan välja att klistra in eller ladda upp en eller flera texter för annotering och analys. Texten kan vara i många olika format, till exempel txt, doc, docx eller rtf. Filformat som för tillfället stöds av SWEGRAM är doc, docx, odt, rtf och txt. Textfiler skapade med t.ex. Microsoft Word, LibreOffice eller OpenOffice konverteras automatiskt till ren text med programmet unoconv (2017) så att du inte behöver göra någon konvertering av filen på förhand utan SWEGRAM gör det automatiskt.

Skriv in eller klistra in en text direkt i fältet alternativt välj fil och tryck på *Ladda upp textfil*. Om du önskar ladda upp flera filer kan du göra det genom att upprepade gånger klicka på *välj fil* och sedan *ladda upp textfil* tills alla önskade filer är uppladdade. De uppladdade, inlästa filerna visas i höger kolumn vilket illustreras i Figur 4.



Figur 4. Att ladda upp filer för annotering.

Efter att texten har lästs in annoteras den med hjälp av verktygskedjan som beskrivs nedan.

2.2.2 Tokenisering och meningssegmentering

Efter att texten har formaterats tokeniseras den. Tokeniseringen är en förutsättning för vidare annotering och sker genom att tokeniseraren delar upp texten i meningar och ord, samt skiljer skiljetecken från orden. På så sätt delas texten upp i token (d.v.s. i ord och skiljetecken), vilka sparas i kolumn 3 i output. Varje token står sedan på en egen rad och ny mening inleds med en tom rad. Förkortningar bevaras som ett ord och särskrivna ord rättas inte i detta steg. Se exemplet för meningarna *”Den här texten t.ex. tokeniseras. Den menings segmenteras också.*” med förkortningen *t.ex*. på en rad som ett token och med det särskrivna ordet *menings segmenteras* som två token, som visas nedan i Figur 5.

```
Den
här
texten

t.ex.
tokeniseras
.

Den
menings
```

Figur 5. Exempel på tokeniserad och meningssegmenterad text.

Tokeniseraren och meningssegmenteraren (Cap et al., 2016) bygger på den tokeniserare som använts för den svenska trädbanken och som nu ingår som komponent i efselab (Östling, 2016).

2.2.3 Normalisering: rättstavning och särskrivningar

Texten kan rättas med avseende på felstavade ord och särskrivningar vilket sker med hjälp av verktyget för normalisering. I output anges den ursprungliga formen som ordet stod i kolumnen FORM och de rättade orden hamnar i kolumnen NORM. Sålunda representeras ursprungstexten sekventiellt ord för ord i FORM-kolumnen medan den korrekta meningen representeras ord för ord i NORM-kolumnen.

De ord som är särskrivna i ursprungsfilen slås samman på en ny rad och anges med de indexnummer som de ingående orden i den ursprungliga särskrivningen har separerat med bindestreck. Exempel på uppmärkning och annotering av särskrivningen ”*inspirations källa*” visas i Figur 6. Ord nummer 2 *inspirations* och ord nummer 3 *källa* analyseras inte utan det korrekta ordet *inspirationskälla,* som nu fått ID-nummer 2-3 (vilket är en kombination av de ID-nummer som de ingående formorden har), återges i NORM-kolumnen och analyseras lingvistiskt. När den annoterade texten används för vidare statistisk analys utelämnas de särskrivna orden och den rättade formen används som underlag.

TEXT ID	ID	FORM	NORM	LEMMA	U-POS	C-POS	C-FEATS	U-FEATS	HEAD	DEPREL
1.1	1	Min	Min	min	DET	PS	UTR|SIN|DEF	Definite=Def|Gender=Com|Number=Sing|Poss=Yes	2-3	nmod:poss
1.1	2-3		inspirationskälla	inspirationskälla	NOUN	NN	UTR|SIN|IND|NOM	Case=Nom|Definite=Ind|Gender=Com|Number=Sing	4	nsubj
1.1	2	inspirations								
1.1	3	källa				NN				
1.1	4	kommer	kommer	komma	VERB	VB	PRS|AKT	Mood=Ind|Tense=Pres|VerbForm=Fin|Voice=Act	0	root
1.1	5	från	från	från	ADP	PP	_	_	6	case
1.1	6	texten	texten	text	NOUN	NN	UTR|SIN|DEF|NOM	Case=Nom|Definite=Def|Gender=Com|Number=Sing	4	obl
1.1	7	.	.	.	PUNCT	MAD	_	_	4	punct

Figur 6. Annotering av särskrivningar.

Rättstavningen baseras på en modifierad version av verktyget Hist-Norm (Pettersson et al., 2013) för stavningskorrigering. HistNorm har utvecklats huvudsakligen för normalisering av ord i historiska texter till modern stavning. För att känna igen och rätta till särskrivningar används ett regelbaserat system. Verktyget är under utveckling och du kan räkna med att verktyget inte alltid rättar till alla felstavade ord eller särskrivningar. Om det är viktigt att en helt korrekt version av texten analyseras vidare rekommenderar vi att du rättar texterna innan du kör den lingvistiska analysen.

2.2.4 Lemmatisering

Ordets basform eller grundform, också kallad lemma, är vanligtvis den minst böjda formen av ett ord och oftast det ord man slår på i lexikon. Lemmat identifieras och återges i kolumn 5 i output. Lemmatiseringen sker i samband med ordklassannoteringen med hjälp av *efselab* (Östling, 2016).

2.2.5 Ordklassanalys

Ordklasstaggaren tilldelar texten ordklasser och morfologiska särdrag. Ordklasserna sparas i output i kolumn 6 och 7, och morfologiska särdrag anges i kolumn 8 och 9. Ordklassanalysen ges i form av universell tagguppsättning (kolumn 6 och 9) och SUC-tagguppsättning (kolumn 7 och 8).

För de universella ordklasserna (U-POS) representeras tagguppsättningen enligt följande:

Tagg	Förklaring	Exempel
ADJ	Adjektiv	fin
ADP	Adposition (preposition)	på
ADV	Adverb	snabbt, mycket
AUX	Hjälpverb	har
CCONJ	Konjunktion	och
DET	Artikel/Determinerare	ett
INTJ	Interjektion	ja
NOUN	Substantiv	bil
NUM	Räkneord	två
PART	Partikel	ut
PRON	Pronomen	han
PROPN	Egennamn	Jenny
PUNCT	Skiljetecken	, .
SCONJ	Subjunktion	att
SYM	Symbol	☺
VERB	Verb	in
X	Annat	xbbe

Tabell 2. Universella ordklasser (U-POS).

Tagguppsättningen för Stockholm-Umeå Corpus anges nedan:

Tagg	Förklaring	Exempel
AB	Adverb	inte
DT	Artikel/Determinerare	ett
HA	Frågande/relativt adverb	när
HD	Frågande/relativ determinerare	vilken
HP	Frågande/relativt pronomen	som
HS	Frågande/relativt possessivt pronomen	vars
IE	Infinitivmärke	att
IN	Interjektion	ja
JJ	Adjektiv	fin
KN	Konjunktion	och
MAD	Stora skiljetecken: punkt, frågetecken, utropstecken, kolon	.
MID	Mindre skiljetecken: kommatecken, tankstreck, semikolon, slash, *	,
NN	Substantiv	bil
PAD	Parvisa skiljetecken: parentes, citattecken	(
PC	Particip	dansande
PL	Partikel	in
PM	Egennamn	Jenny
PN	Pronomen	han

PP	Preposition	på
PS	Possessivt pronomen	hennes
RG	Grundtal	två
RO	Ordningstal	andra
SN	Subjunktion	att
UO	Utländskt ord	the
VB	Verb	fira

Tabell 3. Ordklasser i SUC 2.0.

Den morfologiska analysen återger en uppsättning särdrag som beskriver ordens lexikala och grammatiska egenskaper. Bland lexikala särdrag finns underkategorier av ordklasser, såsom egennamn för substantiv eller olika typer av pronomen. Grammatiska särdrag beskriver kategoriseringen av ordformer för en given ordklass, till exempel genus, numerus, kasus och bestämdhet för nominala ordklasser (substantiv och pronomen) eller kasus, bestämdhet, komparationsgrad och numerus för adjektiv. Den morfologiska analysen skiljer sig mellan UD och SUC.

Universella morfologiska särdrag anges i form av *Särdrag=Värde*-par där särdraget anger den morfologiska kategorin (som kan vara i förkortad form) och värdet beskriver det aktuella särdraget för ordet. Till exempel beskrivs adjektivet *kall* som: Case=Nom|Definite=Indef|Degree=Pos|Number=Sing vilket innebär att ordet *kall* är i nominativ kasus (Case=Nom), obestämd form (Definite=Indef), komparationsgraden är positiv (Degree=Pos) och i singularis (Number=Sing). Några av de viktigaste morfologiska särdragen som förekommer i UD för svenska finns listade i Tabell 4.

Särdrag	**Värde**	**Förklaring**	**Ordklass**
Case (Kasus)	Nom Acc Gen	Nominativ Ackusativ Genitiv	ADJ, NOUN, PRON, PROPN
Definite (Bestämdhet)	Ind Def	Obestämd Bestämd	ADJ, DET, NOUN, PRON, PROPN, ADJ
Gender (Genus)	Utr Neut Com	 Neutrum	ADJ, DET, NOUN, PRON, PROPN
Number (Numerus)	Sing Plur	Singularis Pluralis	ADJ, DET, NOUN, PRON, PROPN
Poss (Possessiv)	Yes	Possessiv	DET
Degree (Komparationsgrad)	Pos	Positiv Komparativ Superlativ	ADJ, ADV
Mood (Modus)	Ind	Indikativ	AUX, VERB
Tense (Tempus)	Pres Past	Presens Preteritum	AUX, VERB
VerbForm (Finithet)	Fin Inf	Finit Infinit	AUX, VERB
Voice	Act	Aktiv	AUX, VERB

(Diates)	Pass	Passiv	
Abbr (Förkortning)	Yes	Abbreviation	ADV

Tabell 4. Morfologiska särdrag i UD för svenska.

Till skillnad från UD representerar inte SUC själva särdragen utan enbart värdet för morfologiska särdraget. SUC:s morfologiska särdrag beskrivs i Tabell 5.

Särdrag	Värde		Ordklass
Genus	UTR	Utrum	DT, HD, HP, JJ, NN, PC, PN, PS, (RG, RO)
	NEU	Neutrum	
	MAS	Maskulinum	
Numerus	SIN	Singular	DT, HD, HP, JJ, NN, PC, PN, PS, (RG, RO)
	PLU	Plural	
Bestämdhet	IND	Indefinit	DT, (HD, HP, HS), JJ, NN, PC, PN, (PS, RG, RO)
	DEF	Definit	
Kasus	NOM	Nominative	JJ, NN, PC, PM, (RG, RO)
	GEN	Genitiv	
Tempus	PRS	Presens	VB
	PRT	Preteritum	
	SUP	Supinum	
	INF	Infinit	
Diates	AKT	Aktiv	VB
	SFO	Passiv eller deponens	
Modus	KON	Konjunktiv	VB
Particip	PRS	Presens	PC
	PRF	Perfekt	
Grad	POS	Positiv	(AB), JJ
	KOM	Komparativ	
	SUV	Superlativ	
Pronomen	SUB	Subjektform	PN
	OBJ	Objektform	
	SMS	Sammansättning	Alla ordklasser

Tabell 5. Morfologiska särdrag i SUC 2.0.

Den intresserade läsaren hänvisas till UDs beskrivning samt till SUC:s manual (Capková & Hartmann, 2006) för mer detaljer.

Annotering av ordklass med morfologiska särdrag sker med hjälp av *efselab* (Östling, 2016).

2.2.6 Syntaktisk analys

Den syntaktiska analysen bygger på dependensanalys, närmare bestämt den så kallade universella dependensformalismen[1] (Nivre et al., 2016). Orden är länkade till varandra parvis och representerar sålunda binära relationer där det ena ordet utgör huvudet och det andra är dess dependent. Verbet står i centrum i satsen och är dess huvudord. I UD version 2 (UD2), som vi använder oss av, utgör det verb som bär den *semantiska* informationen huvudordet till skillnad från andra dependensformalismer där det finita verbet (och sålunda även hjälpverbet) står som huvudord. Alla andra token (ord och skiljetecken) är antingen direkt eller indirekt kopplade till det (semantiska) verbet genom riktade bågar. I UD2 anges riktningen på bågarna från dependent till huvudord. Strukturen anger förhållandet mellan ett ord (ett huvud) och dess dependent(er) genom kopplade bågar med orden som noder. Syntaktiska eller grammatiska funktioner anges vanligen på länkarna mellan varje huvud-dependentrelation. De syntaktiska relationerna med länkar till definitioner återges i Tabell 6 nedan.

Namn	**Beskrivning**
acl	sats som modifierar substantiv (adjektivisk sats)
advcl	adverbialsats modifierare
advmod	adverbial modifierare
amod	adjektivisk modifierare
appos	appositionell modifierare
aux	hjälpverb
case	kasusmarkering
cc	konjunktion
ccomp	komplementsats
clf	*klassifierare*
compound	sammansättning
conj	konjunktion

[1] Universal Dependencies, version 2 (UD2)

cop	kopula
csubj	subjektsats
dep	ospecifierad dependensrelation
det	determinerare
discourse	diskurselement
dislocated	felplacerat element
expl	formellt subjekt i extraposition
fixed	fast flerordsuttryck
flat	platt flerordsuttryck
goeswith	särskrivning
iobj	indirekt objekt
list	lista
mark	infinitivmärke
nmod	nominal modifierare
nsubj	nominal subjekt
nummod	numerisk modifierare
obj	objekt
obl	oblik nominal
orphan	orphan
parataxis	paratax
punct	skiljetecken
reparandum	reparation

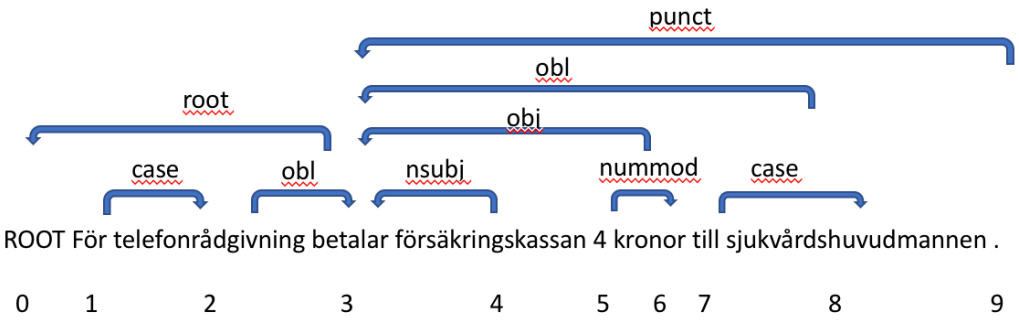
root	rot
vocative	vokativ
xcomp	öppet satskomplement

Tabell 6. Syntaktiska relationer i UD2.

Den syntaktiska analysen i SWEGRAMs output representeras i kolumn 10 (HEAD) och 11 (DEPREL). Vi illustrerar den syntaktiska analysen för meningen ”*För telefonrådgivning betalar försäkringskassan 4 kronor till sjukvårdshuvudmannen.*” dels i CoNLL-U-format som visas i Figur 7 där den syntaktiska analysen markeras i röd ruta, dels som dependensgraf i Figur 8.

TEXT ID	ID	FORM	NORM	LEMMA	U-POS	C-POS	C-FEATS	U-FEATS	HEAD	DEPREL
1.1	1	För	för	för	ADP	PP	_	_	2	case
1.1	2	telefonrådgivning	telefonrådgivning	telefonrådgivning	NOUN	NN	UTR|SIN|IND|NOM	Case=Nom|Definite=Ind|Gender=Com|Number=Sing	3	obl
1.1	3	betalar	betalar	betala	VERB	VB	PRS|AKT	Mood=Ind|Tense=Pres|VerbForm=Fin|Voice=Act	0	root
1.1	4	försäkringskassan	försäkringskassan	försäkringskassan	NOUN	NN	UTR|SIN|DEF|NOM	Case=Nom|Definite=Def|Gender=Com|Number=Sing	3	nsubj
1.1	5	4	4	4	NUM	RG	NOM	Case=Nom	6	nummod
1.1	6	kronor	kronor	krona	NOUN	NN	UTR|PLU|IND|NOM	Case=Nom|Definite=Ind|Gender=Com|Number=Plur	3	obj
1.1	7	till	till	till	ADP	PP	_	_	8	case
1.1	8	sjukvårdshuvudmannen	sjukvårdshuvudmannen	sjukvårdshuvudman	NOUN	NN	UTR|SIN|DEF|NOM	Case=Nom|Definite=Def|Gender=Com|Number=Sing	3	obl
1.1	9	.	.	.	PUNCT	MAD	_	_	3	punct

Figur 7. Annotering i CoNLL-U format.



Figur 8. Den syntaktiska analysen med dependensrelationer.

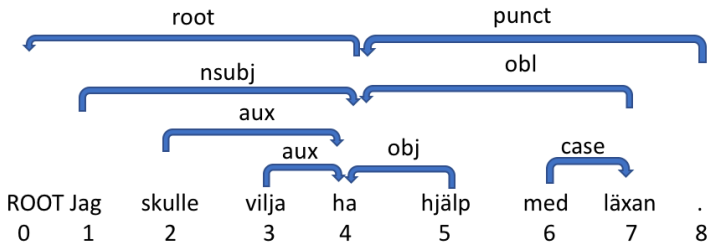
I hela meningen i exemplet ovan binds orden parvis ihop där ett ord i paret är huvudet och det andra är dess dependent med en viss syntaktisk funktion. Dependenter pekar på sitt huvudord med dess syntaktiska relation angiven. Varje mening utgår ifrån en tänkt nod ROOT som får nummer 0 vars syfte är att koppla ihop orden i meningen. Alla orden numreras sekventiellt enligt angiven ID (1–9 i vårt exempel). Huvudordet i meningen är det semantiska verbet *betalar* som har ROOTen som sitt huvudord. Verbet *betalar* är ord nummer 3 och har ett antal direkta dependenter, nämligen satsens subjekt *försäkringskassan* (NSUBJ), det direkta objektet *kronor* (OBJ), de två oblika objekten (OBL), *telefonrådgivning* och *sjukvårdshuvudmannen* samt skiljetecknet ”.” (PUNCT). Dessa ord har i sin tur dependenter. Ordet *4* har ord nummer 6, alltså *kronor,* som sitt huvudord och fungerar som numerisk modifierare (NUMMOD). Ordet *För* har *telefonrådgivning* (ord nummer 2) som sitt huvudord och fungerar som kasus (CASE). Ordet *till* är dependent till ord nummer 8 *sjukvårdshuvudmannen*, och har kasus (CASE) som

syntaktisk funktion. Prepositionerna *för* och *till* i de två prepositionsfraserna är alltså dependenter och utgör inte huvudord i UD2-formalismen.

Traditionellt betraktas hjälpverb som huvudord i många syntaktiska analyser. I UD2 är det däremot det verb som bär den semantiska informationen som utgör huvudet, och hjälpverb fungerar som dependent. Vi illustrerar detta för meningen ”*Jag skulle vilja ha hjälp med läxan.*” i Figur 9 och Figur 10. I verbkedjan *skulle vilja ha* betraktas *ha* som huvudord och verben *skulle* och *vilja* är dess dependenter med hjälpverb (AUX) som syntaktisk funktion.

TEXT ID	ID	FORM	NORM	LEMMA	U-POS	C-POS	C-FEATS	U-FEATS	HEAD	DEPREL
1.1	1	Jag	Jag	jag	PRON	PN	UTR|SIN|DEF|SUB	Case=Nom|Definite=Def|Gender=Com|Number=Sing	4	nsubj
1.1	2	skulle	skulle	skola	AUX	VB	PRT|AKT	Mood=Ind|Tense=Past|VerbForm=Fin|Voice=Act	4	aux
1.1	3	vilja	vilja	vilja	AUX	VB	INF|AKT	VerbForm=Inf|Voice=Act	4	aux
1.1	4	ha	ha	ha	VERB	VB	INF|AKT	VerbForm=Inf|Voice=Act	0	root
1.1	5	hjälp	hjälp	hjälp	NOUN	NN	UTR|SIN|IND|NOM	Case=Nom|Definite=Ind|Gender=Com|Number=Sing	4	obj
1.1	6	med	med	med	ADP	PP	_	_	7	case
1.1	7	läxan	läxan	läxa	NOUN	NN	UTR|SIN|DEF|NOM	Case=Nom|Definite=Def|Gender=Com|Number=Sing	4	obl
1.1	8	.	.	.	PUNCT	MAD	_	_	4	punct

Figur 9. Annotering av verbfras i CoNLL-U format.



Figur 10. Den syntaktiska annoteringen för hjälpverb-verb.

Den syntaktiska analysen genomförs med hjälp av MaltParser (Nivre et al., 2007) som finns implementerad i *efselab*, och är tränad på UD2 för svenska (Östling, 2016).

2.3 Olika modeller för automatisk analys

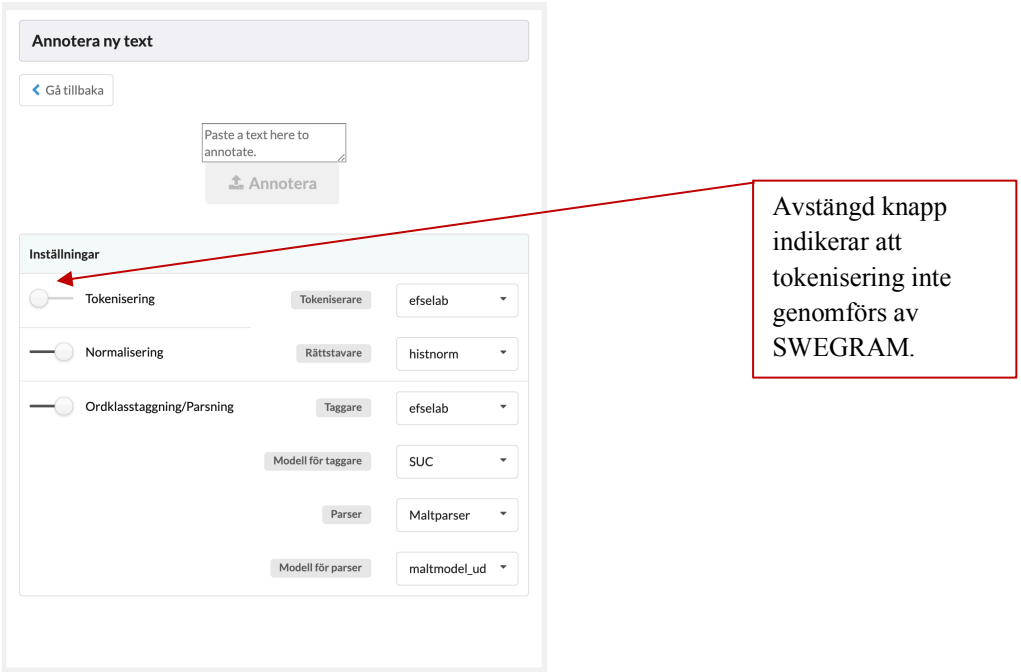
Den automatiska analysen för de olika modulerna (tokenisering, normalisering, ordklasstaggning och syntaktisk analys) kan baseras på olika modeller som utvecklats inom språkteknologin. För tillfället tillhandahålls enbart en modell för respektive modul och dessa utgör också standardinställningen. Modellerna är de bästa som finns att tillgå idag, men i framtiden kommer vi att inkludera andra, bättre modeller också allt eftersom dessa utvecklas.

2.4 Rättning av automatisk analys

Modularitet har varit en viktig faktor i utvecklingen av annoteringsverktyget. Du kan inaktivera vilken modul som helst vilket gör att du kan utesluta vissa delar av den automatiska

annoteringen. Vi rekommenderar detta om du vill utforma analysen på ditt eget sätt av texten eller om du önskar rätta den föreslagna automatiska analysen för senare bearbetning.

I annoteringsverktyget listas alla ingående komponenter som genomför den automatiska lingvistiska analysen: tokenisering, normalisering, ordklasstaggning med syntaktisk analys (parsning). I Figur 11 i vänster kolumn visas dessa komponenter. Alla delar i verktygskedjan är aktiverade från början. Om du vill avaktivera en modul, skjut den aktuella knappen för modulen till vänster så att modulen inte används. I Figur 11 nedan visas att tokeniseringen är avaktiverad, med knappen skjuten till vänster, medan övriga moduler är aktiverade. Detta kan vara användbart om du har egen tokeniserad och meningssegmenterad text som du vill ladda upp för att annotera den med ordklass och syntaktisk information.



Figur 11. Att välja analyskomponenter.

Du kan även välja att använda dig av modulerna i SWEGRAM tillsammans med dina egna modifieringar. Automatisk normalisering med stavningskontroll och sammanslagning av felaktigt särskrivna ord kan du rätta manuellt i efterhand för att sedan ladda upp din version för vidare analys. Du kan också rätta ordklassannoteringen och sedan köra den syntaktiska analysatorn på den rättade ordklasstaggade filen.

Filen med den rättade annoteringen kan du därefter ladda upp för att låta den analyseras vidare automatiskt av efterföljande komponenter. Du kan således bestämma vilka särskilda verktyg du behöver, och få en mer korrekt språklig analys baserad på rättad annotering, vilket kan bidra till mer korrekt annoterade texter.

Notera dock att du måste se till att avmarkera de verktyg som du vill utesluta och samtidigt se till att din analys finns tillgänglig i det format som systemet använder sig av. Du kan *inte*

införa egna annotationer utan måste använda dig av samma taggar som systemet kräver, annars kan inte påföljande analyskomponenter användas.

2.5 Spara din annoterade fil lokalt och exportera till Microsoft Excel

Den analyserade filen kan du spara på din egen dator.

Om du är van vid Microsoft Excel och skulle vilja analysera den lingvistiskt annoterade filen på egen hand kan du exportera filen till Microsoft Excel. I Figur 12 beskrivs hur du ska gå till väga (om du använder Office 2010) när du har laddat ner och sparat din annoterade fil.

1) Under rubriken *Inlästa filer* finns en lista över de filer som har lästs in. Klicka på den fil du vill ha och välj sedan *Ladda ned* för att spara den på din dator.
2) Gå till den mapp som den nya csv-filen har sparats i.
3) Högerklicka på filen och välj *Byt namn* i menyn som kommer upp. Ändra filändelsen från .csv till .txt och tryck på *Enter*.
4) Starta Microsoft Excel och välj *Arkiv* och sedan *Öppna*. Leta reda på .txt-filen som du nyss skapade i menyfönstret som kom upp och öppna den. Du kan behöva välja *Alla filer* på en knapp längst ner till höger i fönstret, eftersom filändelsen inte är densamma som för vanliga Excel-filer.
5) Nu kommer en dialogruta från textimportguiden upp. *Avgränsade fält* är den ursprungliga datatypen (eftersom textfilen skiljer på kolumner endast med hjälp av tab), så välj det alternativet om det inte redan är förvalt. Under rubriken *Filursprung* i mitten av dialogrutan, byt till *65001 : Unicode (UTF-8)*. Klicka sedan på *Nästa*.
6) Ändra *Textavgränsare* till *{ingen}*. Förhandsgranskningen i nedre delen av fönstret borde visa en uppdelning i kolumner nu. Klicka på *Nästa*.
7) Eftersom du antagligen inte vill att numeriska värden automatiskt ska avrundas eller förvandlas till datum i din fil, byt *Kolumndataformat* till *Text*. Klicka sedan på *Slutför*.
8) Spara filen som Microsoft Excel-fil. Klart!

Figur 12. Läsa in annoterad fil i Microsoft Excel.

3 Analys

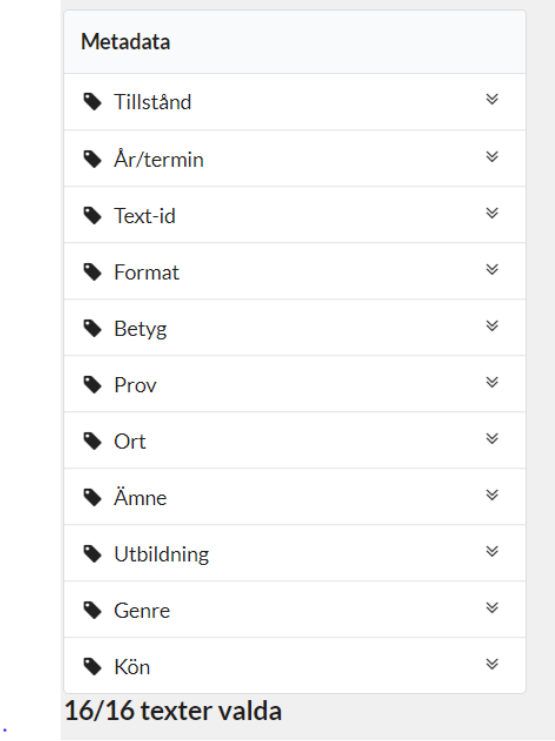
Det andra steget i SWEGRAMs funktioner är den fortsatta analysen av annoterade texter. SWEGRAM beräknar statistik på olika nivåer. Beräkningen kan ske för en uppladdad text, flera texter i en fil eller flera uppladdade filer. Du kan få fram allmän statistik, statistik om ordklassfördelning, stavfel, särskrivningsfel, läsbarhetsmått och ordfrekvenser. Vidare kan du studera enskilda texter i visualiseringsverktyget. Slutligen kan du filtrera din sökning så att

statistik visas för ett urval av de texter du laddat upp. Läs mer om filtrering och statistik nedan.

3.1 Metadata och filtrering

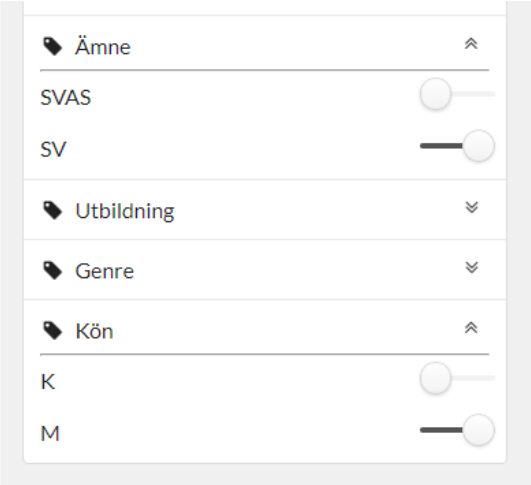
Om dina texter är kodade med metadata bestående av strukturerad information om texterna, läser SWEGRAM in dessa data. Under rubriken *Inlästa filer* visas information om den inlästa filen och här anges också om texten har metadata. (SWEGRAM kan bara läsa metadata om den uppladdade filen inleds med en definition av metadata; se vidare avsnitt 4.)

Under rubriken *Metadata* visas en lista över de variabler av metadata som lästs in. Variablerna är desamma som ingår i definitionen av metadata. Figur 13 visar listan över de metadata som har kodats i Uppsala elevtextkorpus, en korpus som har byggts upp med hjälp av SWEGRAM (Megyesi, Näsman & Palmér, 2016).



Figur 13. Lista över metadata från uppladdade filer hämtade från Uppsala elevtextkorpus.

Klicka på pilarna bredvid varje rubrik för att se vilka värden som finns för de olika variablerna i det inlästa materialet. Varje variabelvärde är försett med en knapp. Använd dessa knappar för att filtrera en sökning genom att välja bort/välja till varianter av metadata. Statistiken ändras varje gång du väljer bort/väljer till ett värde av metadatan. Under listan anges hur många av de uppladdade texterna som har filtrerats fram.



Figur 14. Filtrering av metadata efter Ämne och Kön.

Figur 14 visar hur filtrering kan se ut när SWEGRAM arbetar med Uppsala elevtextkorpus. Under variabeln *Ämne* har värdet SVAS (svenska som andraspråk) valts bort, så att endast texter skrivna i ämnet svenska (värde=SV) ingår i filtreringen. Under variabeln *Kön* har värdet K (kvinnor) valts bort. Därmed ingår endast texter skrivna av män (värde=M) i filtreringen.

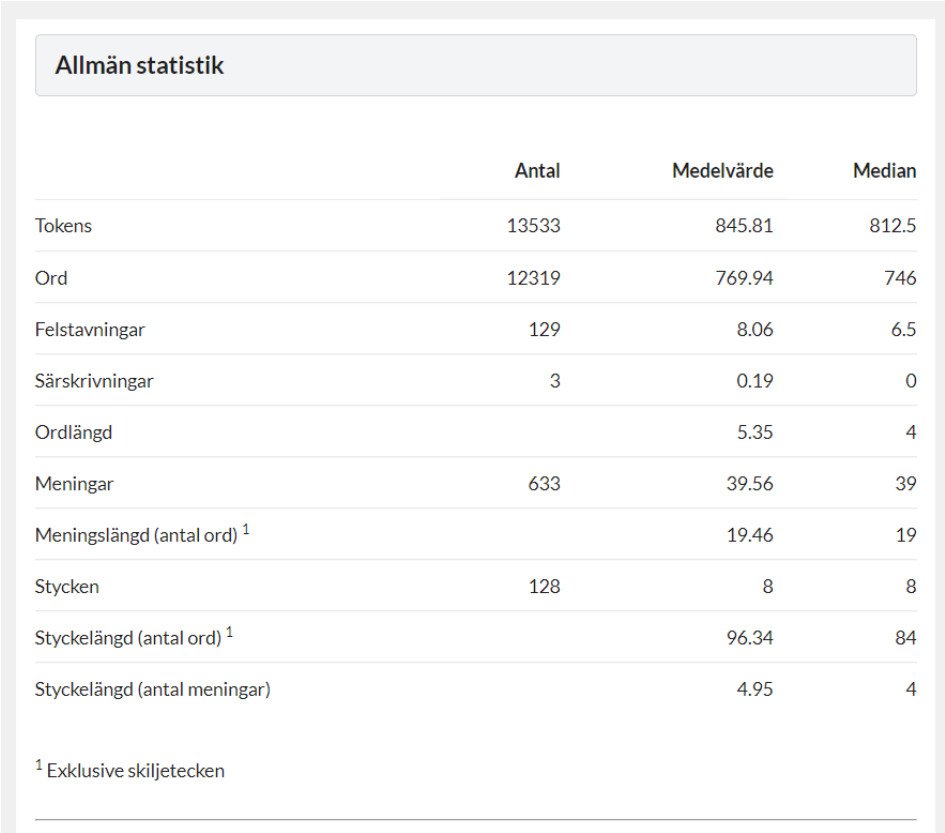
3.2 Statistik och kvantitativa mått

Under rubriken Statistik kan du via rullgardinsmenyn välja mellan rubrikerna Allmän statistik, Läsbarhet, Ordklasser och Frekvenser. Det finns även en flik för Exportera statistik. De olika funktionerna förklaras närmare nedan.

3.2.1 Allmän statistik

Den allmänna statistiken ger information om token, ord, ordlängd, meningar, meningslängd (antal ord), felstavningar, särskrivningar, ordlängd, stycken och styckelängd (antal ord och antal meningar). För varje kategori ges antal, medelvärde och median för alla valda texter.

Figur 15 visar hur den allmänna statistiken presenteras där information finns om token, ord, felstavningar, särskrivningar, ordlängd, meningar, meningslängd, stycken och styckelängd.



Allmän statistik

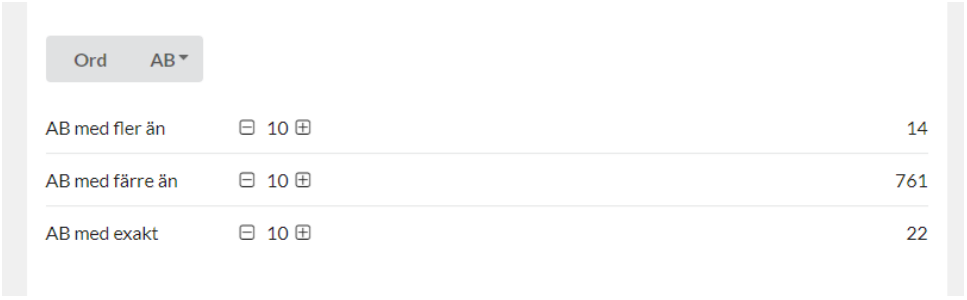
	Antal	Medelvärde	Median
Tokens	13533	845.81	812.5
Ord	12319	769.94	746
Felstavningar	129	8.06	6.5
Särskrivningar	3	0.19	0
Ordlängd		5.35	4
Meningar	633	39.56	39
Meningslängd (antal ord) [1]		19.46	19
Stycken	128	8	8
Styckelängd (antal ord) [1]		96.34	84
Styckelängd (antal meningar)		4.95	4

[1] Exklusive skiljetecken

Figur 15. Presentation av Allmän statistik i SWEGRAM.

Dessutom innehåller den allmänna statistiken uppgifter om antal ord med visst antal bokstäver. Utgångsläget är antal ord med fler än, färre än och exakt 3 bokstäver. Använd knapparna + och – om du vill söka på ord utifrån en annan ordlängd.

Du kan också söka på ord inom en viss ordklass. I Figur 16 gäller sökningen antal adverb (AB) med fler än, färre än och exakt 10 bokstäver.



Figur 16. Presentation av antal ord *från ordklassen adverb (AB) med visst antal bokstäver.*

3.2.2 Ordklasstatistik

Ordklasstatistiken utgår från SUC som presenterades i avsnitt 2.2.5, Tabell 3. I standardläget visas antal och andel av samtliga ordklasser från SUC. I Figur 17 har tre av ordklasserna valts bort, MAD, PAD och MID, som alla utgör skiljetecken. I statistiken över ordklassfördelningen som visas i Figur 18 ingår därmed inte dessa.

Figur 17. Filtrering av ordklasser inför ordklasstatistik.

Om du vill begränsa sökningen till ett fåtal ordklasser är det enklast att först välja bort alla ordklasser genom att klicka på översta rutan ”Välj bort alla ordklasser” och därefter välja till de ordklasser du vill att sökningen ska omfatta.

I den statistiska analysen visar SWEGRAM alla ordklasser i storleksordning efter dels antal, dels andel i de uppladdade texterna. Figur 18 visar en del av resultatet av en sökning där de åtta vanligaste ordklasserna i sökningen presenteras i figuren.

Ordklass	Antal	Andel
NN	2693	21.87%
VB	2345	19.04%
PP	1242	10.09%
AB	1018	8.27%
PN	873	7.09%
JJ	843	6.85%
KN	679	5.51%
DT	665	5.40%

Figur 18. Presentation av de åtta vanligaste ordklasserna i en sökning.

3.2.3 Läsbarhetsmått

I SWEGRAM beräknas också olika mått för läsbarhet, ordvariation och nominalitet. Dessa samlas under rubriken Läsbarhetsmått, som omfattar Lix, Ovix, Typ-token ratio (TTR), Full nominalkvot och Enkel nominalkvot, som illustreras i Figur 19.

Läsbarhet

	Total[1]	Median[2]
LIX (total)	48.6	48.91
OVIX (total)	57.93	59.62
Typ-token ratio	0.2	0.47
Full nominalkvot	1.27	1.21
Enkel nominalkvot	1.35	1.35

[1] Behandlar alla texter som en text och beräknar läsbarhet utifrån det
[2] Medianvärde för den valda gruppen enskilda texter

Figur 19. De mått som ingår i SWEGRAMS läsbarhetsstatistik.

I figurens första kolumn listas de aktuella måtten. I den andra kolumnen visas läsbarhetsmåttsvärdet för varje läsbarhetsmått i samtliga texter som sökningen behandlar. Rubriken TOTAL anger måttet för alla texter i urvalet. Den tredje kolumnen visar medianvärdet för enskilda texter i urvalet. Nedan beskrivs och diskuteras varje mått i ordningsföljden Lix, TTR, Ovix, Enkel nominalkvot och Full nominalkvot.

3.2.3.1 Lix

Lix står för läsbarhetsindex, beräknat enligt en formel som togs fram av Carl-Hugo Björnsson (1968). Utifrån antalet ord, antalet meningar och antalet långa ord över 6 bokstäver beräknas textens Lix-värde.

Formel för beräkning av Lix:

LIX: (antal ord/antal meningar) + ((antal långa_ord*100) / totalt antal ord)

Formel 1. Lix.

Lix-värdet anses vara ett enkelt mått på hur ordens och meningarnas längd kan antas påverka läsbarheten i en text. Inför tolkning av resultatet kan följande lista, hämtad från en lärobok i Stilistik (Melin & Lange, 2006), användas:

Texttyp	Lix	Ord/mening	Långord	Tolkning
Barn- och ungdomsböcker	27	12	15	Mycket lätt
Skönlitteratur	33	15	18	Lätt
Dags- och veckopress	39	14	25	Medelsvår
Saklitteratur	47	18	29	Svår
Facklitteratur	56	20	35	Mycket svår

Tabell 7. Lix-värden, olika genrer (Melin & Lange, 2006)

3.2.3.2 Typ-token ratio

Typ-token ratio, förkortat TTR, är ett enklare mått på ordvariation. På svenska används ibland uttrycket lex-/löp-kvot.

Formel för att beräkna TTR:

TTR: antal unika token / totalt antal token

Formel 2. TTR: Typ-token ratio.

TTR diskuteras av Johansson (2009:142 f.) och slutsatsen är att måttet bör användas framför allt för jämförelser av ordvariation i lika långa texter, alternativt i kombination med andra mått.

3.2.3.3 Ovix

Ovix står för ordvariationsindex. Formeln togs ursprungligen fram av Tor G. Hultman i samband med forskningsprojektet Skrivsyntax stora undersökning av bruksprosa och elevtexter (Hultman & Westman, 1977). Måttet mäter antalet lexem (lexord) i förhållande till antalet löpord i en text, och det är konstruerat så att textlängd inte ska påverka resultatet, vilket blir fallet vid enklare mätningar av lex-/löpkvot (se ovan om TTR).[2]

Några olika formler för att beräkna Ovix förekommer i forskningslitteraturen. Ovix enligt Hultman & Westman (1977:264) beräknas $V=N\ (2\text{-}N^k)$. V står där för antalet lexikonord, N står för antalet löpord, K är en konstant och Ovix = 1/k.

Enligt Hultman (1994:62) beräknas Ovix som det anges i Formel 3.

```
OVIX: 1 / (log (2- (log(typord) / log(token)))  / log(token))
```

Formel 3. Ovix (Hultman, 1994:62)

Lix.se anger beräkning av Ovix enligt Formel 4, som är ekvivalent med Formel 3 ovan. Det är också denna formel som används i Swegram.

```
OVIX: log(token) / log(2-(log(types)/log(token))
```

Formel 4. Ovix enligt Lix.se.

Ett högre Ovix-värde innebär större ordvariation i förhållande till textens längd (Hultman & Westman, 1977:56). Ovix ska, enligt Hultman & Westman (1977:60), inte förstås som ett mått på hur många synonymer av ett och samma ord som används i en text, utan snarare som ett mått på synpunktsrikedom: ”Troligen är det snarare variationer i informationsrikedom än i synonymvariation som vårt ordvariationsmått mäter”. Det finns dock inget enkelt samband mellan Ovix-värdet och textkvalitet (Hultman, 1994:63).

I jämförelser mellan grupper av texter bör medianvärden användas (jfr Hultman & Westman, 1977:56).

Många svenska studier har använt sig av Ovix, till exempel Nyström (2000), Östlund-Stjärnegårdh (2002), Magnusson & Johansson Kokkinakis (2011) och Nordenfors (2011). Nyström (2000:192) problematiserar måttet och drar slutsatsen att "[o]vix som statistiskt

[2] Geijerstam (2006:107) anger att Ovix tas fram genom att dividera "summan av olika ord i texten med summan av alla ord i texten" men detta beskriver snarare TTR.

värde troligtvis har ett slags optimal nivå. Ett alltför högt värde tyder på en alltför snabb ämnesprogression, medan ett alltför lågt kan tyda på innehållslig tunnhet".

3.2.3.4 Enkel nominalkvot

Enkel nominalkvot ger indikationer om textens nominalitet och informationstäthet samt visar förhållandet mellan substantiv (nn) och verb (vb) i en text. Måttet har använts bland annat av af Geijerstam (2006:108 ff.) för att visa densiteten, den lexikala tätheten, i en text. En högre andel substantiv jämfört med verb ger en mer informationstät text, eftersom substantiv ofta är centrala vid förmedling av fakta.

Beräkning av enkel nominalkvot:

$$\frac{nn}{vb}$$

Formel 5. Enkel nominalkvot.

3.2.3.5 Full nominalkvot

Full nominalkvot mäter grad av nominalitet på ett mer avancerat sätt. Denna kvot visar förhållandet mellan å ena sidan substantiv och ordklasser som relaterar till substantiv och å andra sidan verb, adverb och pronomen. Substantiv (nn), prepositioner (pp) och particip[3] (pc) anses bidra till hög nominalitet och informationstäthet, medan verb (vb), adverb (ab) och pronomen (pn) i stället bidrar till att informationen i en text blir glesare, mindre informationstät.

Hultman & Westman (1977) visar att skriftspråk har betydligt högre nominalkvot än talspråk. Grad av nominalitet blir därmed ett stildrag genom att hög andel substantiv, prepositioner och particip gör en text skriftspråklig medan en lägre nominalkvot innebär att stilen i en text ligger närmare talspråket. Nominalkvot ska inte ses som ett mått på nominalfrasen, men måttet ger en indikation på grad av lexikalitet i nominalfraserna. I skrivutvecklingsforskning har det visat sig att äldre skribenter i högre utsträckning väljer substantiv som huvudord i sina nominalfraser, medan yngre oftare väljer pronomen (t.ex. Scott 1988).

Formel för full nominalkvot:

[3] Till participen räknas både presens och perfekt particip.

(nn+pp+pc) / (pn+ab+vb)

Formel 6. Full nominalkvot.

3.2.4 Frekvenser

Under rubriken Frekvenser visar SWEGRAM vanligt förekommande ord och skiljetecken. Resultatet av en sökning visar orden och skiljetecknen (token) sorterade efter frekvens. Vidare anges ordklass, antal förkomster och ordets andel av texturvalet som helhet. I Figur 20 visas ett exempel på hur början av en frekvenslista kan se ut.

Index	Token	Ordklass	Förekomster	Andel
1	.	MAD	576	4.26%
2	och	KN	438	3.24%
3	,	MID	387	2.86%
4	att	IE	256	1.89%
5	det	PN	245	1.81%
6	i	PP	242	1.79%
7	att	SN	230	1.70%

Figur 20. Resultatet av sökning på frekvenser.

Frekvenssökningens utgångsläge är *Norm*, d.v.s. sökningen görs på token som är normaliserade. Välj *Lemma* om du vill söka på ordens grundformer och *Form* om du vill söka på de vanligaste orden utan SWEGRAMs rättning av stavfel.

Välj bort ordklasser om du vill begränsa sökningen (jfr Figur 17). Om du bara vill studera frekvensen av vissa ordklasser kan du först välja bort samtliga och sedan välja till.

Exempel: Figur 21 visar en frekvenslista på normaliserade ord där enbart substantiv och verb har valts.

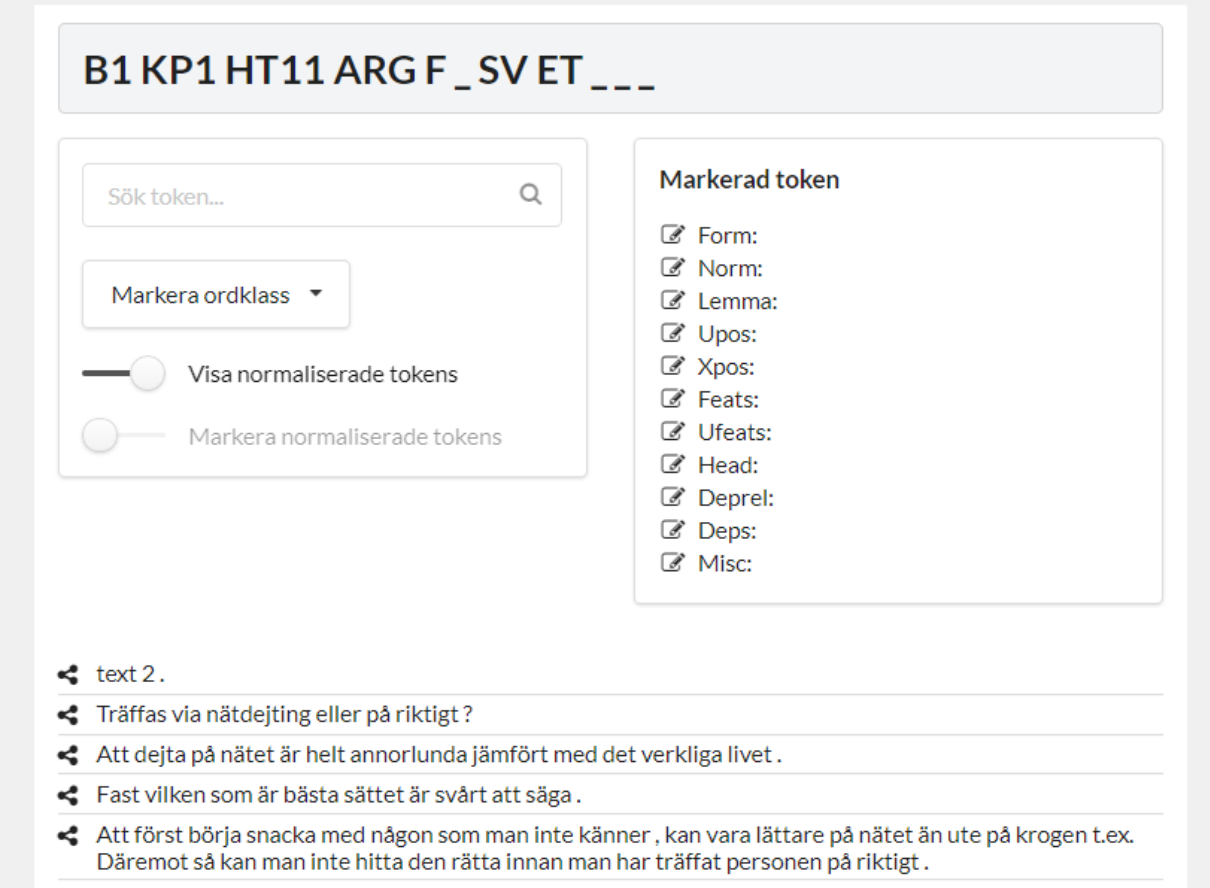
Index	Token	Ordklass	Förekomster	Andel
1	är	VB	216	4.28%
2	har	VB	138	2.74%
3	kan	VB	103	2.04%
4	kommer	VB	83	1.65%
5	böcker	NN	59	1.17%
6	var	VB	56	1.11%
7	sätt	NN	45	0.89%
8	vara	VB	41	0.81%

Figur 21. Frekvenslista utifrån Norm där enbart verb (VB) och substantiv (NN) ingår.

3.3 Visualisera text och analys

Under Visualisera text kan du få en tydlig, visuell bild av SWEGRAMs analys av enskilda texter. Du kan få fram bilder av enskilda token och av en eller flera ordklasser i texten.

Figur 22 visar meny för visualisering och början av en elevtext. Första raden visar textens metadata: B1 KP1 HT11 ARG F _ SV ET _ _ _. I menyn kan du inför visualiseringen välja enskilda token som du vill se i texten genom att skriva in ordet eller skiljetecknet i rutan. SWEGRAM fetmarkerar de ord eller skiljetecken som du har valt. Du kan också visualisera en eller flera ordklasser genom att gå till *Markera ordklasser* i samma meny under sökrutan och slå på knapparna för önskvärd(a) ordklass(er). De valda ordklasserna markeras på respektive tillhörande token i texten med specifik färg för varje ordklass.

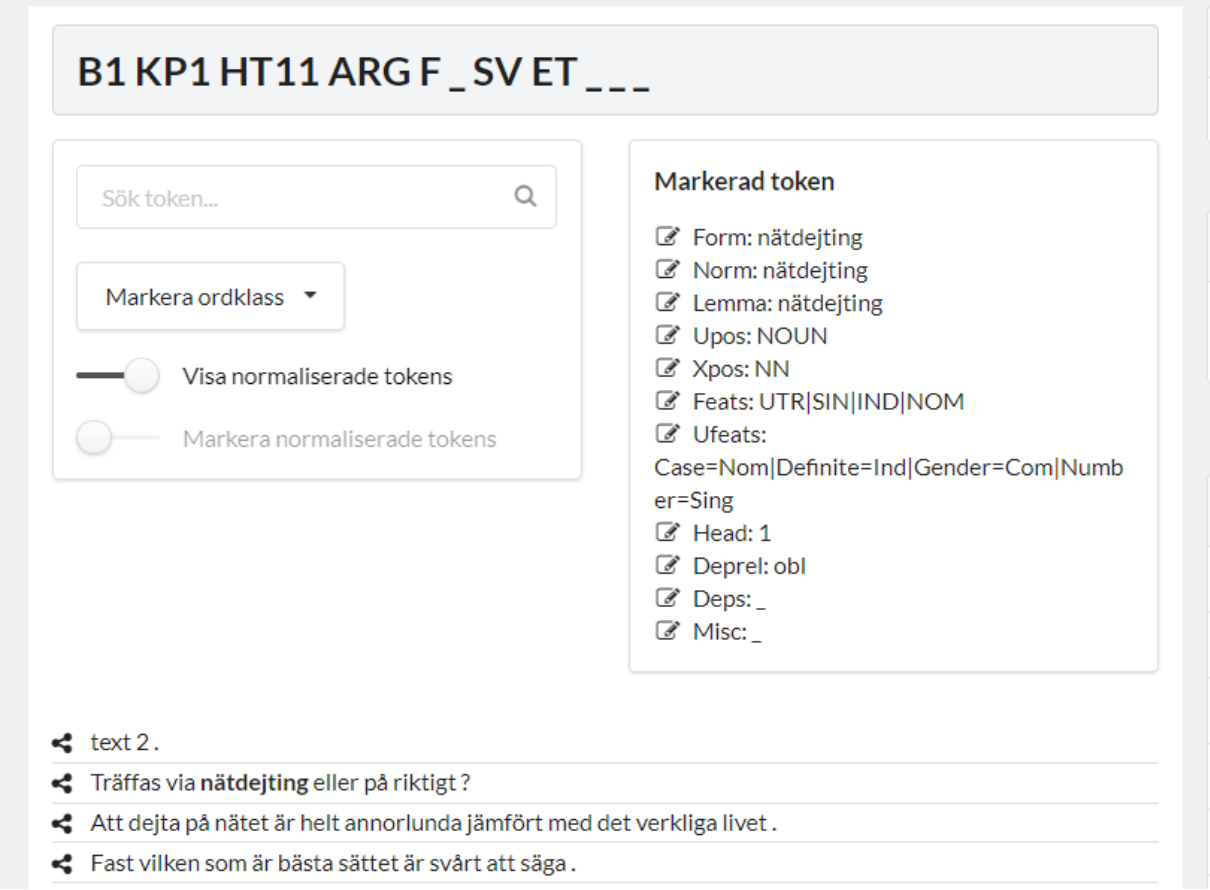


Figur 22. Menyn för visualisering när en text har valts.

Om du vill titta på ursprungstexten istället för den normaliserade texten som är förvald kan du välja att inte visa normaliserade token. Slutligen kan du, med knappen under, välja att markera normaliserade token.

Den nedre delen av Figur 22 visar textens fem första meningar, mening för mening, rad för rad, enligt verktygets annotering.

I Figur 23 har ett token markerats: nätdejting. I rutan ”Markerad token” visas den fullständiga analysen av detta token.



Figur 23. Presentation av fullständig analys av ett enskilt token: nätdejting.

3.4 Exportera statistik

Om du vill spara resultatet av dina sökningar kan du exportera statistik till csv-filer och sedan öppna dessa i Microsoft Excel. Figur 24 visar hur menyn för att exportera statistik ser ut.

Figur 24. Att exportera statistik.

Under fliken Exportera statistik har SWEGRAM förinställt att exporten ska inkludera Allmän statistik, Läsbarhetsstatistik, Ordklasstatistik och Frekvensordlista. Du kan välja bort delar av statistiken genom att bocka av dem i rutorna. Välj hur många token som ska ingå i exporten av frekvensordlistan genom att skriva in antalet i den tomma rutan. I exemplet ska 25 token ingå, vilket är standardinställningen.

Välj mellan att exportera sammanfattande statistik eller statistik för varje enskild text. Statistik för varje enskild text kan vara intressant om du vill göra ytterligare beräkningar utifrån dina data, till exempel beräkningar av korrelationer, spridning eller statistisk signifikans.

Välj också om den exporterade statistiken ska använda punkt eller komma som decimalavgränsare. Standardinställningen är punkt. Även denna funktion är användbar om du vill använda Microsoft Excel för att göra ytterligare beräkningar av dina data. Ditt Excel-program kan vara inställt på punkt eller komma som decimalavgränsare, och dina exporterade data bör anpassas efter det. Om inget aktivt val görs används punkt som decimalavgränsare.

I

1) Välj *Exportera statistik*.
2) Välj ev. vilken decimalavgränsare du vill ha i ditt kommande Excel-dokument, punkt eller komma.
3) Ladda ner statistiken genom att klicka på *Exportera sammanfattande statistik* eller *Exportera statistik för varje enskild text* beroende på vad du vill ha för statistik.
4) Gå till den mapp som den nya csv-filen har sparats i.
5) Högerklicka på filen och välj *Byt namn* i menyn som kommer upp. Ändra filändelsen från .csv till .txt och tryck på *Enter*.
6) Starta Excel och välj *Arkiv* och *Öppna*. Leta reda på .txt-filen som du nyss skapade i menyfönstret som kom upp och öppna den. Du kan behöva välja *Alla filer* på en knapp längst ner till höger i fönstret, eftersom filändelsen inte är densamma som för vanliga Excel-filer.
7) Nu kommer en dialogruta från Textimportguiden upp. *Avgränsade fält* är den ursprungliga datatypen (eftersom textfilen skiljer på kolumner med hjälp av tabb), så välj det alternativet om det inte redan är förvalt. Under rubriken *Filursprung* i mitten av dialogrutan, byt till *65001: Unicode (UTF-8)*. Klicka sedan på *Nästa*.
8) Ändra *Textavgränsare* till {ingen}. Förhandsgranskningen i nedre delen av fönstret borde visa en uppdelning i kolumner nu. Klicka på *Nästa*.
9) Välj *Allmänt* som *Kolumndataformat* om detta inte redan är förvalt. Klicka sedan på

Figur 25 nedan sammanfattas de nödvändiga stegen vid exportering av statistiska data till Microsoft Excel.

Figur 25. Exportera statistik till Microsoft Excel.

4 Skapa din egen korpus (textsamling) med egen metadata

Med hjälp av SWEGRAM kan du ladda upp en eller flera textfiler och verktyget skapar en lingvistiskt annoterad fil, en korpus, som du sedan kan analysera med hjälp av analysverktyget. Du kan bygga upp ett korpusmaterial av ett antal filer med en eller flera texter i varje. Texterna annoteras automatiskt med hjälp av verktyget: texterna segmenteras i meningar och ord, orden märks upp med dess ordklass och morfologi, ordens basform anges, felstavade eller särskrivna ord rättas och meningen analyseras syntaktiskt. Antalet texter som kan ingå i korpusen är obegränsat.

Om du vill utnyttja möjligheten att filtrera materialet vid den statistiska analysen behöver du också koda dina texter med metadata (se också avsnitt 3.1). Metadatan kan omfatta all möjlig information du har om texterna, t.ex. information om författarna (identitet, kön, ålder, geografisk hemvist) och texterna i sig själva (textidentitet, tillkomstår, genre, publikationsställe). Du väljer själv vilka och hur många olika metadata din korpus ska omfatta.

4.1 Exempel på förarbete med metadata

Här visar vi hur du kan arbeta om du vill skapa en korpus med metadata. I exemplet utgår vi från att du vill bygga upp en korpus av 30 elevtexter skrivna av elever från årskurs 6. Eleverna är både pojkar och flickor och texterna kan vara utformad enligt berättande eller argumenterande genre. Eftersom de är hämtade från ett prov har de också ett betyg: F, E, D, C, B eller A.

Du har alltså metadatainformation om 30 texter, om författarens kön, textens genre och textens betyg. Då behöver du välja ut koder för denna information. En bra metod kan vara att ställa upp din information i en tabell och lista tillgänglig information som du vill använda i dina sökningar, den typ av metadata som informationen representerar och de koder du vill använda.

Tillgänglig information	**Typ av metadata**	**Förslag på koder**	**Förklaring**
Antalet elevtexter är 30.	TEXT-ID	T1, T2, T3, T4 etc.	Text 1 etc.
För varje elevtext finns elevens kön angivet.	KÖN	FLI, POJ	Flicka, Pojke
Elevtexterna är antingen berättande eller argumenterande texter.	GENRE	BER, ARG	Berättande, Argumenterande
Elevtexterna är försedda med betygen F, E, D, C, B och A.	BETYG	F, E, D, C, B, A	Betyg F etc.

Tabell 8. Att skapa metadata för en korpus.

I Tabell 8 anges tillgänglig information, typ av metadata, förslag på koder och förklaring av koderna för vårt exempel. Se för tydlighetens skull till att varje kod blir unik, om det är möjligt. I exemplet väljer vi t.ex. inte att beskriva genre enbart med initialerna B i berättande och A i argumenterande, eftersom dessa koder skulle kunna förväxlas med koderna i betygen.

Nästa steg är att skapa en eller flera textfiler med texterna och metadatainformationen. I exemplet arbetar vi med texter i Microsoft Word, men även txt-filer kan användas. Eftersom korpusen bara omfattar 30 texter är det lämpligt att samla alla en dokumentfil. Det finns ingen teknisk begränsning av hur många texter som kan samlas i en fil, men vid större korpusar kan

det passa att dela upp texterna på flera filer, för att göra dessa mer överskådliga vid manuell hantering.

Allra först i dokumentet skriver du in din metadatadefinition. Denna visar vilka typer av metadata din korpus innehåller och i vilken ordning typerna anges. Varje typ av metadata får en post. Mellan varje post är det mellanslag. I exemplet (jfr Tabell 8) ingår fyra typer av metadata. Metadatan har alltså fyra poster, och metadatadefinitionen blir <Text-id Kön Genre Betyg> (jfr Figur 26). Definitionen inleds med < och avslutas med >.

Lägg in texterna i dokumentet. Börja med definitionen av metadatan (i vårt exempel <Text-id Kön Genre Betyg>), som gäller alla texter. Fortsätt sedan med att tilldela unika kodrader för varje enskild text. Dessa kodrader har samma struktur som metadatadefinitionen, men datatypnamnen byts ut till de kodvariabler som gäller för respektive text, se exempel i Figur 26. Börja med kodraden för text 1. Fortsätt därefter med kodrad för text 2, text 2, kodrad för text 3, text 3 etc. Både definitionen av metadatan och kodraderna ska inledas och avslutas med < och >. I Figur 26 visas början av ett dokument som innehåller en kortare text.

<TEXT-ID KÖN GENRE BETYG>

<T1 POJ BER E>

Bästa vintem
De var en kall vinter i år och isen var stark. Så jag och mina kompisar bästämde oss. För gå ner till sjön och åka skickskor. När vi kom fram var de många som var där och åkte. Jag och mina tre kompisar Alvin, Emrah och Yasar gick på isen. Mest var de unga som åkte. Lite längre bort var de nåra killar som spelade is hockey.

Figur 26. Början av ett worddokument med korpus av elevtexter, åk 6.

I Figur 26 syns först metadatadefinitionen: ”<TEXT-ID KÖN GENRE BETYG>. Därefter följer kodraden: <T1 POJ BER E>. Det betyder att texten har text-id 1, är skriven av en pojke, är berättande och har fått betyget E. Därefter följer början av elevtexten T1.

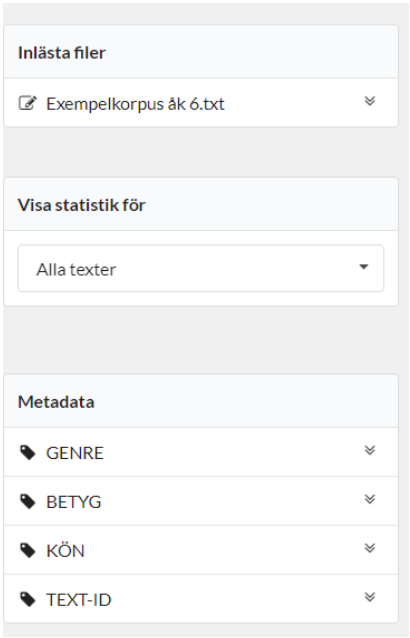
Efter den första texten läggs övriga texter in i dokumentet. Före varje text placeras dess unika kodrad. Det är viktigt att koderna alltid följer den ordningsföljd som anges i metadatadefinitionen, att inga mellanslag placeras före kodraden eller mellan en klammer och första koden.

Om du för någon text skulle sakna viss metadatainformation, kan detta i kodraden markeras med understreck, "_". Om du t.ex. saknar information om elevens kön för en text som är nummer 5 i samlingen, argumenterande och har fått betyget C, kan kodraden se ut så här:

<T5 _ ARG C>.

När dina texter är samlade och kodade i dokumentet har du en korpus, en textsamling att annotera. I exemplet får textsamlingen heta Exempelkorpus åk 6.

Nu kan du annotera din korpus och använda SWEGRAMs analysverktyg för statistik och visualisering av information i texterna. Figur 27 visar hur menyerna med information om metadata ser ut i SWEGRAM när exempelkorpusen har blivit uppladdad i verktyget.



Figur 27. Meny med metadata när korpusen Exempelkorpus åk 6 har laddats upp i SWEGRAM.

Du kan utöka din korpus med fler texter och fler dokumentfiler, som alla kan laddas upp samtidigt för analys av SWEGRAM.

Observera att om koderna kommer i fel ordning eller ett extra mellanslag finns i koden kan verktyget inte genomföra analysen eller analysen blir fel. I sådant fall får man ett felmeddelande av typen ”Server error” och du måste själv manuellt rätta felet i filen.

5 Om verktyget

SWEGRAM har utvecklats i samarbete mellan Institutionen för Lingvistik och filologi och Institutionen för nordiska språk vid Uppsala universitet inom ramen för SWE-CLARIN projektet, finansierat av Vetenskapsrådet. SWE-CLARIN syftar till att språkbaserade och

lingvistiskt analyserade material med hjälp av språkteknologiska verktyg ska kunna tillgängliggöras som forskningsdata för humanistisk och samhällsvetenskaplig forskning.

SWEGRAMs komponenter är skrivna i Python och användargränssnittet är skrivet i HTML, CSS och JavaScript. Webbgränssnittet utvecklades med hjälp av Django. Verktyget är öppen källkod och är tillgängligt för alla intresserade. För att säkerställa att filer som laddas upp i SWEGRAM för annotering och/eller analys inte sparas på vår server tas dina uppladdade filer bort från servern 5 minuter efter att analysen med SWEGRAM är genomförd, eller när webbläsaren stängs. Det medför att SWEGRAM inte kan läsa in dina filer på nytt och du måste därför ladda upp filerna igen om du önskar arbeta vidare med verktyget.

- Megyesi, B., Palmér, A. & Näsman J. (2019) *SWEGRAM: Annotering och analys av svenska texter.* Institutionen för lingvistik och filologi och Institutionen för nordiska språk, Uppsala universitet.

- Näsman, J., Megyesi, B., & Palmér, A. (2017) SWEGRAM – A WebBased Tool for Automatic Annotation and Analysis of Swedish Texts. In *Proceedings of 21st Nordic Conference on Computational Linguistics,* Nodalida 2017.

Om du har synpunkter eller får problem med verktyget, kontakta oss: SWEGRAM@stp.lingfil.uu.se.

The SWEGRAM project

Projektledare: Beáta Megyesi, Institutionen för lingvistik och filologi, UU
Projektdeltagare: Anne Palmér och Catrin Isaksson, Institutionen för nordiska språk, UU
Webbutveckling: Jesper Näsman, Institutionen för lingvistik och filologi, UU

Referenser

Åsa af Geijerstam. (2006). *Att skriva i naturorienterande ämnen i skolan*. (Studia Linguistica Upsaliensia 3.) Uppsala: Uppsala universitet.

C. H. Björnsson. (1968). *Läsbarhet*. Stockholm: Liber.

Fabienne Cap, Yvonne Adesam, Lars Ahrenberg, Lars Borin, Gerloff Bouma, Markus Forsberg, Viggo Kann, Robert Östling, Aaron Smith, Mats Wirén & Joakim Nivre. (2016). SWORD: Towards Cutting-Edge Swedish Word Processing. Abstract of a conference presentation at the Swedish Language Technology Conference (SLTC 2016), in Umeå, October 16–7, 2016.

Eva Ejerhed, Gunnel Källgren, Ola Wärnstedt & Magnus Åström. (1992). *The Linguistic Annotation System of the Stockholm-Umeå Corpus Project*. Report NO 33. University of Umeå: Department of Linguistics.

Sofia Gustafson-Capková & Britt Hartmann. (2006). *Manual of the Stockholm-Umeå Corpus version 2.0*. Stockholm. Stockholms universitet.

Tor G. Hultman & Margareta Westman. (1977). *Gymnasistsvenska*. Lund. LiberLäromedel.

Tor G. Hultman (1994). Hur gick det med OVIX? Jörgensson, N, Platzack, C & Svensson, J (red.): *Språkbruk, grammatik och språkförändring*. Lund: Lunds universitet, s. 55–64.

Victoria Johansson. (2009). *Developmental Aspects of Text Production on Writing and Speech*. Travaux de l'institut de linguistique de Lund 48. Lund: Lund University.

Ulrika Magnusson & Sofie Johansson Kokkinakis. (2011) Computer based quantitative methods applied to first and second language student writing. I: *Young Urban Swedish. Variation and change in multilingual settings*. Inger Källström & Inger Lindberg (eds.). Göteborgsstudier i nordisk språkvetenskap 14. University of Gothenburg. S. 105–124.

Beáta Megyesi, Jesper Näsman & Anne Palmér. (2016). The Uppsala Corpus of Student Writings: Corpus creation, Annotation, and Analysis. I: *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC 2016). Paris: European Language Resources Association. S. 3192–3199.

Lars Melin & Sven Lange. (2006). *Att analysera text. Stilanalys med exempel*. Lund: Studentlitteratur. 3:e uppl.

Joakim Nivre, Johan Hall, Jens Nilsson, Atanas Chanev, Gülşen Eryigit, Sandra Kübler, Svetoslav Marinov & Erwin Marsi. (2007). *MaltParser: A language-independent system for data-driven dependency parsing.* Natural Language Engineering, 13(2):95–135.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan D Hajič, Christopher Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty & Daniel Zeman. (2016). Universal Dependencies v1: A Multilingual Treebank Collection. In Proceedings of LREC.

Mikael Nordenfors. (2011). *Skriftspråksutveckling under högstadiet*. (Göteborgsstudier i nordisk språkvetenskap 16.) Göteborg: Göteborgs universitet.

Catharina Nyström. (2000). *Gymnasisters skrivande. En studie av genre, textstruktur och sammanhang*. (Skrifter utgivna av Institutionen för nordiska språk vid Uppsala universitet 51.) Uppsala: Uppsala universitet.

Eva Pettersson, Beáta Megyesi & Joakim Nivre. (2013). Normalisation of Historical Text using Context-Sensitive Weighted Levenshtein Distance and Compound Splitting. In *Proceedings of the 19th Nordic Conference of Computational Linguistics*, NODALIDA.

Cheryl, M. Scott. (1988). Spoken and Written Syntax. In: *Later Language Development. Ages Nine through Nineteen*. Ed. Nippold, M. A. pp. 49–95. Pro ed Austin, Texas.

Unoconv (2017) Universal Office Converter: unoconv. https://github.com/dagwieers/unoconv (2018-02-20)

Robert Östling. (2016). Efficient Sequence Labeling: efselab. Stockholms universitet. https://github.com/robertostling/efselab (2018-02-20).

Eva Östlund-Stjärnegårdh. (2002). *Godkänd i svenska? Bedömning och analys av gymnasieelevers texter*. (Skrifter utgivna av Institutionen för nordiska språk vid Uppsala universitet 57.) Uppsala: Uppsala universitet.