Blir studenternas språk sämre?

Viggo Kann

KTH Skolan för datavetenskap och kommunikation
viggo@kth.se

Abstract

In Sweden, there has since 2013 been a debate in public media, where university professors, mostly from departments of history, have argued that today's students entering university are much less accomplished than earlier students when it comes to basic Swedish language skills. According to the debate, both the spelling and grammar of Swedish students are weak. The first signs of this are said to have been observed in 2010. In order to objectively study the language skills of Swedish first-year university students, we have constructed a tool, based on the Swedish spell checker Stava and the Swedish grammar checker Granska, that measures the language skills that, according to the critics, have been deteriorating during the last 3–4 years. We have collected two corpora of authentic first-year student essays from two Swedish universities. Each corpus contains essays from six different years. The results show, surprisingly, that the language skills of the studied groups of students have not deteriorated during the period. If anything, the skills have slightly improved regarding the level of difficulty of the language.

1. Bakgrund och frågeställning

Året 2013 inleddes av att nio historiker i kraftfulla ordalag, ”ett veritabelt nödrop”, ondgjorde sig över svenska studenters dåliga språk i en debattartikel i Upsala Nya Tidning. Klagomålen var bland annat ”oerhört begränsat ordförråd”, ”stavningen över lag eländig” och ”grammatiska förmågan är ytterst begränsad”. Svenska Dagbladets ledare hängde snabbt på: ”Om läget är så här bekymmersamt framstår det till exempel som fullständigt vettlöst att öka antalet högskoleplatser.” Diskussionen fortsatte med 163 kommentarer på SvD-webben och i mängder av bloggar.

Den 5 februari 2013 medverkade universitetslektorn Ebba Lisberg Jensen i Obs i P1 i Sveriges Radio. Hon beskrev sina och kollegernas erfarenheter av studenters undermåliga språk- och skrivkunskaper. I en uppföljande artikel i tidningen Universitetsläraren (5/2013) säger hon att hon såg de första tecknen för tre-fyra år sedan, på att de brister som varit sällsynta undantag hos enskilda studenter blev alltmer frekventa. Hon kallar det ”ett kognitivt glapp”:

> . . . studenter med obefintlig känsla för hur skriftspråk ska se ut. Det finns de som, för att ta två av många exempel, inte vet att en mening börjar med versal och avslutas med punkt. Läraren måste avgöra var meningarna börjar och slutar.

Dagens Nyheter, Metro och andra medier har också tagit upp debatten.

Är det så allvarligt som dessa lärare påstår? Är det sant att förstaårsstudenternas språk blir allt sämre? Hur kan man på ett objektivt sätt mäta studenternas brist på skriftlig språkförmåga? Dessa frågor har undersökts i detta projekt.

Skolelevers språkförmåga har tidigare studerats, genom manuell genomgång av uppsatser från nationella prov från olika år (Tordebring, 2007), men större studier på svenska förstaårsstudenter tycks saknas, kanske för att det är svårt att få tag på material från olika år som är jämförbart. Manuella genomgångar är också mycket arbetskrävande och potentiellt subjektiva.

2. Metod

För att på ett objektivt och effektivt sätt mäta texters kvalitet på en basal nivå kan språkteknologi användas. I detta projekt har vi utvecklat ett automatiskt verktyg som mäter basal språkkvalitet med avseende på de språkliga problem som nämnts i debatten. Verktyget använder de av KTH:s språkteknologigrupp byggda systemen (Kann, 2010) Stava för stavningskontroll (Domeij, Hollman & Kann, 1994), Granska Tagger för ordklasstaggning (Carlberger & Kann, 1999) och Granska[1] för grund syntaktisk analys (Knutsson, Bigert & Kann, 2003) och grammatikkontroll (Bigert, Kann, Knutsson & Sjöbergh, 2005), efter viss modifiering.

Bland annat har följande mått beräknats:

- *Antal stavfel per ord* har tagits fram genom att dokumentet stavningskontrollerats med Stava på både svenska och engelska (eftersom texterna ofta innehåller engelska facktermer och citat). Därefter har alla hittade stavfel gåtts igenom manuellt och korrekta ord (mest korrekt stavade namn) har plockats bort.
- *Antal grammatikfel per ord* har mätts genom att dokumentet grammatikkontrollerats med Granska (regelsamling version 8 utvidgad med ett par regler).
- *Vokabulärstorlek* har mätts på de första 550 respektive 1100 orden i dokumentet, vilket är mittpunkten på det specificerade ordantalet i respektive uppgiftslydelse. Anledningen att vokabulärstorleken mätts på detta sätt är för att den ska bli jämförbar mellan dokument av varierande storlek. Orden har lemmatiserats med hjälp av Granska Tagger och antalet olika lemmaformer i dokumentet har mätts.
- *Andel ovanliga ord*. Som ovanliga ord räknas de ord som inte är med i vår lista med de 8300 vanligaste svenska orden. Denna lista bygger på tidningstext. I tidningstext är andelen ord från denna lista omkring 80%.

[1] Källkoden för Stava och Granska kan laddas ner från http://www.csc.kth.se/tcs/humanlang/tools.html.

Enligt debatten sågs de första tecknen på försämring av skrivförmågan hos förstaårsstudenter omkring 2010. För att undersöka om en försämring ägt rum behövs alltså ett material med jämförbara uppsatser skrivna av studenter i början av en universitetsutbildning under en period från 2009/2010 och några år framåt. Dessutom måste materialet vara i maskinläsbar form för att det ska kunna studeras med det verktyg som vi utvecklat. Vi har efter åtskilligt letande hittat två uppsättningar studentuppsatser som uppfyller kraven:

1. 1300 uppsatser skrivna av förstaårsstudenter på civilingenjörsprogrammet i Datateknik på KTH 2010–2016. Studenterna har i den programsammanhållande kursen under tredje veckan på höstterminen fått i uppgift att skriva en reflekterande text om studieteknik och studiemotivation på 500–600 ord. Varje år sedan 2010 skriver cirka 200 nyantagna studenter denna uppsats.

   Dokumentformatet är PDF. För att konvertera dokumenten till HTML med bibehållna stycken har Linuxprogrammen pdftohtml och pdfreflow använts.

2. 300 uppsatser skrivna av förstaårsstudenter på kognitionsvetarkandidatprogrammet på Linköpings universitet mellan 2002 och 2013. Studenterna har fått i uppgift att skriva en uppsats på 1000–1200 ord som diskuterar frågan *kan maskiner tänka?*. Varje år sedan 2002 skriver cirka 40 nyantagna studenter denna uppsats. Fram till 2009 var längdkravet 5–7 sidor, vilket gör att uppsatserna från dessa år är lite längre än uppsatserna efter 2010.

   Dokumentformatet är DOC. För att konvertera dokumenten till HTML-format har Libreoffice använts med kommandot `soffice -headless -convert-to html:HTML` följt av en HTML-rensning med Linuxprogrammet HTML tidy.

Vi har skrivit ett program i Python 3.4 som parsar HTML-dokumenten med hjälp av HTMLParser och filtrerar bort sidhuvuden, innehållsförteckningar och källförteckningar.

För några få dokument i varje dokumentsamling har det inte varit möjligt att utvinna läsbar text. Dessa dokument har inte tagits med i analysen.

3. Resultat

Resultaten för dokumentsamling 1 framgår av tabell 1. Det är knappt 0,3 % stavfel per ord och omkring 0,9 % grammatikfel per ord. De vanligaste stavfelen är i tur och ordning olika former av *prokastinering*, *hittils*, *seminarie* och *programering*, se ordmolnet i figur 1. De vanligaste grammatikfelen är *kommer* utan *att* (som numera av många inte ens anses vara fel) och mening utan verb.

I tabell 2 visas resultaten för dokumentsamling 2. Här är de vanligaste stavfelen *intentionala*, *reaktionell*, *talsvarsystem* och *tillfredställande*. De vanligaste grammatikfelen är även här *kommer* utan *att* och mening utan verb.

Något som knappt förekommer alls är avsaknad av punkt i slutet av meningar och stor bokstav i början av meningar.

4. Diskussion och slutsatser

Resultaten visar ingen signifikant ändring av antalet stavfel eller grammatikfel under de studerade åren, 2010–2015 på KTH och 2002–2013 på Linköpings universitet. Möjligen kan en svag uppgång i stavfelsfrekvenserna skönjas mellan 2004 och 2009 i Linköpingsmaterialet. Antalet felstavningar per ord är dock mycket lågt, så uppgången är bara från strax under 0,2 % till strax under 0,3 %. Andelen stavfel för KTH- och Linköpingsstudenter är i stort sett lika.

Inte heller vokabulärstorleken har ändrats under åren. Det enda mått där vi kan se en förändring är andelen ovanliga ord. Studenterna använder mer och mer ovanliga ord i sina uppsatser. Detta mönster syns både på KTH och Linköpings universitet.

Denna undersökning finner alltså inte stöd för hypotesen att förstaårsstudenters språkanvändning har fallit under de senaste åren. Studenterna gör inte fler fel nu än för fem år sedan. Möjligen gjorde studenterna i början av 2000-talet något färre stavfel, men skillnaden är liten. Att studenterna använder mer svåra ord i sitt språk nu skulle kunna tyda på att deras språkbehärskning snarare är lite bättre nu än tidigare.

Har debattörerna alltså haft fel i sin kritik? Ja, åtminstone stämmer inte kritiken för all svensk universitetsutbildning. I detta projekt har vi studerat ingenjörsstudenter vid KTH och kognitionsvetarstudenter vid Linköpings universitet och har alltså vederlagt kritiken för grupper av studenter från olika områden och olika lärosäten. Det är allvarligt om universitetsvärlden går och tror att studenternas skrivförmåga blivit riktig dålig under senare tid. Risken är att gymnasieskolan beskylls för att inte ha förberett eleverna tillräckligt för högre studier.

Vi ska alltså inte sänka våra förväntningar på studenternas skrivförmåga. Det är dock fortfarande viktigt att de får tillfälle att utveckla skrivförmågan under sin utbildning, inte minst när det gäller det vetenskapliga språket.

Tack

Forskningen har finansierats av Erik Wellanders fond. Tack till Sara Stymne som tipsade författaren om existensen av dokumentsamling 2 från Linköpings universitet. Uppsatserna i den andra dokumentsamlingen har tillhandahållits av Nils Dahlbäck och Annika Silvervarg vid Linköpings universitet, vilket författaren är mycket tacksam för.

Bibliografi

J. Bigert, V. Kann, O. Knutsson, and J. Sjöbergh. 2005. Grammar checking for Swedish second language learners. *CALL for the Nordic Languages* **30**, 33–47.

J. Carlberger and V. Kann. 1999. Implementing an efficient part-of-speech tagger. *Software Practice and Experience* **29**, 815–832.

R. Domeij, J. Hollman, and V. Kann. 1994. Detection of spelling errors in Swedish not using a word list en clair. *J. of Quantitative Linguistics* **1**, 195–201.

V. Kann. 2010. KTHs morfologiska och lexikografiska verktyg och resurser. *LexicoNordica* **17**, 99–117.

O. Knutsson, J. Bigert, and V. Kann. 2003. A robust shallow parser for Swedish. *Proc. 14th Nordic Conf. on Computational Linguistics*.

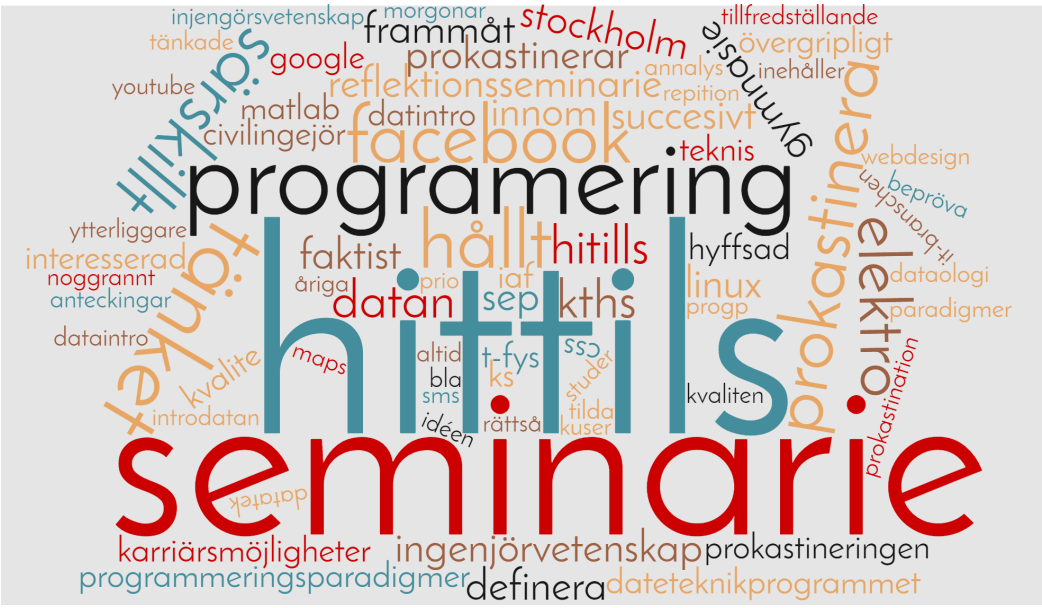
S. Tordebring. 2007. Var det bättre förr?: En kvantitativ jämförelse av elevtexter från 1985, 2005 och 2006. Examensarbete på lärarprogrammet. Karlstads universitet.

	2010	2011	2012	2013	2014	2015	2016
Antal dokument	179	153	197	217	171	206	194
Antal ord/dokument	619	529	568	563	595	625	644
Stavfel per ord	0,26%	0,22%	0,22%	0,22%	0,32%	0,24%	0,30%
Andel ovanliga ord	11,1%	12,6%	12,9%	12,0%	13,0%	13,2%	13,2%
Vokabulärstorlek	220	219	231	226	230	231	231
Grammatikfel per ord	1,08%	1,00%	0,99%	0,88%	0,93%	0,98%	1,03%

Tabell 1: Resultat för dokumentsamling 1 (KTH-uppsatser), medelvärden för olika mått.

	2002	2004	2009	2011	2012	2013
Antal dokument	57	51	39	47	44	45
Antal ord/dokument	2028	1874	1799	1116	1125	1136
Stavfel per ord	0,18%	0,15%	0,23%	0,23%	0,35%	0,28%
Andel ovanliga ord	12,2%	12,5%	11,9%	13,9%	15,5%	17,8%
Vokabulärstorlek	355	349	328	348	348	336
Grammatikfel per ord	0,62%	0,55%	0,60%	0,73%	0,69%	0,67%

Tabell 2: Resultat för dokumentsamling 2 (Linköpingsuppsatser), medelvärden för olika mått.



Figur 1: Stavfelsförekomster i KTH-uppsatserna.