

LINKÖPINGS UNIVERSITET  
Institutionen för datavetenskap  
Inlämningsuppgift

17 februari 2005

Inlämningsuppgift  
**TDDA94 Lingvistik**

Stavnings- och grammatikkontroll

**Namn** David Hall  
**E-post** davpe813@student.liu.se

**Examinator**  
Magnus Merkel

# 1 Bakgrund

Datoriserade språkgranskningsverktyg kan åtminstone delas in i två huvudgrupper: stavningskontroll och grammatikkontroll.

## 1.1 Stavningskontroll

Kontroll av felstavningar är den enklaste formen av språkkontroll och har funnits relativt länge för ordbehandlingssystem. Ord i dokumentet kontrolleras mot ett lexikon på det aktuella språket. Lexikonet kan i sin enklaste form bestå av enbart en upprädningslista av ord i alla möjliga böjningsformer. En något mer sofistikerad lösning är att ta med ordet i sin grundform tillsammans med ett paradigm<sup>1</sup> (böjningsschema) för att därigenom spara visst lagringsutrymme, vilket inte fungerar lika bra om stammen ändras vid böjning<sup>2</sup>. Eftersom svenskan tillåter ett närmast fritt bildande av ordsammansättningar är det inte hållbart att ta med alla kombinationer i ett lexikon. Detta gör det mer komplicerat än exempelvis engelskan där nya begrepp ofta bildas av flera fristående ord som alla redan finns i lexikonet. Ett sätt är då att stavningskontrollen ser om ordet i dokumentet går att bilda utifrån korrekta orddelar. Risken finns dock att stavningskontrollen godkänner ord som är ospråkliga ("köpenlampa") eller som inte finns (än) i sinnevärlden ("båttransistor").

Särskrivningar kan vara svåra att upptäcka när man kontrollerar mot ett lexikon eftersom de olika orddelarna oftast är helt korrekta ord även fristående. En möjlig väg är att kontrollera texten mot en lista över kända särskrivningar men man riskerar då missa mindre vanliga särskrivningar. Vissa ord särskrivna kan fortfarande vara korrekta men få en helt ny mening, t.ex. *råttkött* jämfört med *rätt kött*. I dessa fall går det inte att avgöra vad som är korrekt utan kunskap om semantiken i texten.

I vissa fall förekommer även ord hopskrivna. Dessa kan vara svåra att upptäcka särskilt som denna kontroll kommer i konflikt med bildandet av sammansättningar.

En del språkgranskningsverktyg har ordlistor över ord, eller stavningar av ord, som enbart förekommer i talspråk, som *dej* och *mej* istället för *dig* respektive *mig*. Det finns även kontroll över ord som nästan enbart förekommer i så kallad kanslisvenska, dvs. ett byråkratiskt och för många oförståeligt språk. I båda dessa fall brukar verktygen påtala användaren om detta och föreslå alternativa skrivningar.

## 1.2 Grammatikkontroll

Grammatikkontroll är mer avancerat än stavningskontroll. Till skillnad mot stavningskontrollen som arbetar med enskilda ord arbetar grammatikkontrollen med hela meningar. En grammatikkontroll kan bl.a. hitta fel i genus-, numerus- och specieskongruens, dubbelt supinum och avsaknad av subjekt eller finit verb. Grunden är en taggning av texten, där ordklass bestäms för varje ord<sup>3</sup>. På denna taggning appliceras sedan grammatiska regler.

<sup>1</sup>paradigm, NE, [http://www.ne.se/jsp/search/article.jsp?i\\_art\\_id=279751](http://www.ne.se/jsp/search/article.jsp?i_art_id=279751), 2005-02-16

<sup>2</sup>böjning, NE, [http://www.ne.se/jsp/search/article.jsp?i\\_art\\_id=139722](http://www.ne.se/jsp/search/article.jsp?i_art_id=139722), 2005-02-16

<sup>3</sup>Inget grammatikprogram klarar alla fel, Språkkonsulten nr 4 - 1999

### 1.3 Undersökta produkter

De verktyg jag testat är Microsoft Word, Granska, Grim och ”Stavningskontroll på svenska”. Stavningskontrollen och grammatikkontrollen i Microsoft Word är utvecklad av det finska företaget Lingsoft Ab.

Granska är resultatet av projektet Svensk grammatikgranskning<sup>4</sup> vid Institutionen för numerisk analys och datalogi vid KTH.

Grim<sup>5</sup> är ett program för inläring av svenska som utvecklas inom projektet ”The use of language tools for writers in the context of learning Swedish as a second language” som genomförs av KTH och Stockholms universitet. Programmet finns i olika versioner. Jag har valt att testa Java-versionen.

Stavningskontroll på svenska<sup>6</sup> är ett program för stavningskontroll över webben skrivet av Thomas Padron-McCarthy.

## 2 Exempeldokument

Exempeldokumentet består av ett utdrag ur de första två kapitlen i Maj Sjöwalls och Per Wahlöös deckare ”Polis, polis, potatismos” från 1970. Vissa ord, främst substantiv, kan idag verka föråldrade. En stor del av texten utgörs av dialog med talspråk och fragmentariska meningar, jag har valt att inte ta hänsyn till Microsoft Words utmärkning av talspråkliga ord.

### 2.1 Felstavningar

Utöver vanliga felstavningar har jag även sorterat in fall där versal och gemen förväxlats, dessa är sammansättningarna ”Köpenhamnsbåt” och ”Malmöpolisen” som i exempel- och boktext skrivs med inledande gemen<sup>7</sup>. Till felstavningar räknar jag även låneorden terylenebyxor (eg. terylenbyxor) och menthol (eg. mentol).

Följande felstavningar hittades vid en manuell kontroll av texten: ijusen, köpenhamnsbåt, analkande, á la, malmöpolisen, groggglas, toma, femtiårsåldern, terylenebyxor, menthol, skuten, signalment, altså, int, mirstroget (14 stycken)

### 2.2 Särskrivningar och hopskrivningar

Följande särskrivningar hittades vid en manuell kontroll: allt för, medel åldern, hamn inloppet, serverings bordet, ny komlingen, med det samma, för resten, Back lund, tand petaren, sur mulet (10 stycken)

Följande hopskrivningar hittades: förstekriminalassistent, medandraord, brun-kavaj, bättrevittnen (4 stycken)

<sup>4</sup><http://www.nada.kth.se/theory/projects/granska/index.html>

<sup>5</sup><http://skrutten.nada.kth.se/grim/>

<sup>6</sup><http://plutten.dnsalias.org/scripts/stava/stava.pike>

<sup>7</sup><http://www.spraknamnden.se/sprakladan/ShowSearch.aspx?id=id=31746;objekttyp=lan>

## 2.3 Kongruensfel

”Kongruens är då ett led böjs efter ett annat led.”<sup>8</sup> Det kan t.ex. röra sig om att substantivets genus eller numerus styr adjektivet. (Bilen är grön. Huset är grönt. Bilarna är gröna.)

Hittade fel: somrarna minst lik ofta, det stora hotell mitt emot, tystlåtna människor, ingen lyckades ... fastslår, det mörknande kvällshimlen, de verkade handfallen, en man har blivit nerskjutna, all uniformerade personal, någon direkt avgörande att säga (9 stycken)

## 2.4 Meningsbyggnad

På vissa ställen i texten har kommateringar fallit bort. Detta skapar ogrammatiska meningar som verktygen bör fånga. Exempel: ”...än till midnattssolen längs horisonten ser man ijusen...” samt ”...kastade en snabb likgiltig blick på den analkande mannen och började genast åter vända sig mot sina gäster utan att för en sekund avbryta den utläggning han höll på med, och i samma ögonblick...” (2 stycken)

Dialogerna i texten är skrivna på ett sätt som ligger talspråket nära. Vissa av dessa meningar saknar därför subjekt eller verb. Exempel: ”Går inte att få tag i.” och ”Tja, omkring halv nio.”.

# 3 Resultat

## 3.1 Stavning

Alla testade verktyg missade felstavningen ”á la” (ska vara à la), vilket antagligen beror på att varje ord kontrollerats var för sig. Andra ord som slapp igenom tre av fyra verktyg var: malmöpolisen, skuten och signalment. De verkar ha slunkit igenom för att verktygen antingen gjort felaktiga sammansättningar (malmö-polisen, skut-en, signal-ment) eller för att verktygen (Granska och Grim) inte tar hänsyn till att egennamn som Malmö ska ha stor bokstav. Bäst var Microsoft Word som hittade 13 av 15 felstavningar (femtiårsåldern markerades som talspråkligt), näst bäst var Granska och Stavningskontroll på svenska som båda hittade 10 av 15.

Alla verktyg märkte ut *gripenbergare* som felstavning, ett ord som bara verkar förekomma i Sjöwall-Wahlöös böcker och det får därför anses vara acceptabelt att alla verktyg märkt den som felstavad. Tre av fyra anmärker på *grapetonic* (det kan diskuteras om huruvida det ska skrivas ihop eller inte). Både Granska och Grim misstänker stavningsfel på ”grapetonic” och särskrivning för ”grape tonic”. Microsoft Word märker ut egennamn som Regementsgatan och Davidshallstorg som felstavningar men även ett ord som rumsbokningar. Microsoft Word är den som felmärker mest antal ord som felstavningar, Granska är den som felmärker minst antal ord. Anledningen till att Word hittar flest felstavningar och samtidigt märker ut flest ord som felaktiga bör bero på att den är mer restriktiv vad gäller att sätta samman ord.

<sup>8</sup>Bolander, Funktionell svensk grammatik, 2005, s. 227

### 3.2 Särskrivningar

De särskrivningar som alla verktyg missar är ”medel åldern”, ”för resten”, ”Backlund” och ”sur mulet”. Bäst på att fånga särskrivningar är Grim (6 av 10), sämst är Stavningskontroll på svenska (1 av 10) som till skillnad mot de andra verktygen inte kontrollerar särskrivningar särskilt. Microsoft Word hittar bara två särskrivningar (2 av 10).

Microsoft Word är det enda verktyg som lyckas identifiera några hopskrivningar (medandraord och bättrevittnen, 2 av 4). Ord som slinker igenom alla verktyg är således förstekriminalassistent och brunkavaj, båda ord som teoretiskt skulle kunna ha en betydelse. Även här utmärker sig Microsoft Word genom att hitta få särskrivningar men att hitta hopskrivningarna. Återigen skulle jag tro det beror på att den är mer försiktig på att bilda sammansättningar.

### 3.3 Kongruensfel

Alla tre verktygen med grammatikkontroll hittar två av nio kongruensfel. Det enda fel alla tre hittar är ”det stora hotell”. Därutöver hittar Microsoft Word ”en man har blivit nerskjutna” och Granska/Grim (som antagligen har samma kod för just denna funktion) ”tystlätne människor”. Detta är alltså ett område där ingen av verktygen lyckas särskilt bra.

### 3.4 Meningsbyggnad

Inget av verktygen hittade de missade kommateringarna.

Granska är onekligen bäst på att hitta meningar utan verb eller subjekt (4 utan subjekt, 12 utan verb, 1 utan finit verb). Microsoft Word lyckas bra med att hitta meningar utan verb (12 stycken) men inga utan subjekt, vilket grammatikkontrollen inte verkar ha stöd för<sup>9</sup>. Grim kommer inte i närheten (3 utan verb, 0 utan subjekt).

### 3.5 Felaktiga anmärkningar

Verktygen Granska och Grim har märkt ut en del meningar som felaktiga där jag själv inte gjort det. I vissa fall kan det vara en smaksak vilket man väljer, i andra har verktyget uppenbart inte förstått grammatiken trots att den är korrekt.

Ett exempel på det förstnämnda är ”rockvaktmästaren läste ostörd en klassiker i djupet av sin garderob” som markeras som felplacerat adjektiv (Granska och Grim). Däremot går ”ostörd läste rockvaktmästaren en klassiker i djupet av sin garderob” igenom.

Saker som Grim inte förstår och därför felmarkerar är ”...som han skam till sägandes...” och ”en välklädd och solenn samling”. I det sistnämnda exemplet går det felaktiga ”en välklädd och solent samling” igenom.

---

<sup>9</sup><http://www.lingsoft.fi/doc/swegc/errtypes.html>