"Det er fort gjort og skrive feil."

En presentasjon av en automatisk grammatikkontroll for bokmål

Av Kristin Hagen og Pia Lane

Det siste året har Tekstlaboratoriet ved Universitetet i Oslo og det finske firmaet Lingsoft[1] utviklet en grammatikkontroll for bokmål. Resultatet finnes i Microsofts nye Office-pakke, Office XP. Den svenske grammatikkontrollen *Grammatifix* (Arpe 2000, Birn 2000) er brukt som arbeidsmodell, men alle reglene er nyskrevet for norsk.

I dette foredraget vil vi først gi en kort presentasjon av hvilke feil denne grammatikkontrollen forsøker å finne. Deretter vil vi ved hjelp av konkrete eksempler forklare hvordan den virker, og til slutt vil vi diskutere noen av svakhetene ved grammatikkontrollen.

1. Feiltyper

Figur 1 gir en oversikt over de viktigste feiltypene grammatikkontrollen forsøker å finne:

Figur 1: De viktigste feiltypene grammatikkontrollen forsøker å finne:

- Manglende samsvar i kjønn i substantivfraser: *Jeg kjøpte en nytt hus/et ny hus à Jeg kjøpte et nytt hus*
- Manglende samsvar i tall i substantivfraser : *Jeg spiste et epler à Jeg spiste et eple*
- Feil bruk av substantivets bestemthetsform: *Jeg solgte et huset à Jeg solgte et hus*
- Feil bruk av adjektivets bestemthetsform: *Jeg kjøpte et grønne hus à Jeg kjøpte et grønt hus*
- Flere adjektiv etter hverandre: *Han kjørte en rød rask bil à Han kjørte en rød, rask bil/ Han kjørte en rød og rask bil*
- Manglende samsvar mellom subjekt og predikativ: *Bilen er rødt à Bilen er rød*
- Bruk av *ingen* og *noen* ved negasjon: *Jeg kjøpte ikke ingen epler på butikken àJeg kjøpte ikke noen epler på butikken* eller *Jeg kjøper noen bok àJeg kjøper ei bok*
- og/å-feil: *De gikk å sang àDe gikk og sang, Hun skal å vise meg den nye kjolen à Hun skal vise meg den nye kjolen, Den lille gutten kan både snakke å synge à Den lille gutten kan både snakke og synge, Jeg trenger og sove àJeg trenger å sove, Han skal prøve og skrive korrekt àHan skal prøve å skrive korrekt* eller *Han ønsket sove àHan ønsket å sove*
- Finitt verb etter hjelpeverb: *De vil spaserer à De vil spasere*
- Perfektum partisipp uten *ha*: *Jeg skal spist innen klokka tre àJeg skal ha spist innen klokka tre*
- Feil bruk av s-passiv: *Duken behøver vaskes àDuken må vaskes/Duken behøver ikke vaskes*
- For mange finitte verb i setningen: *I Norge er var det slik àI Norge er det slik*
- Ingen finitte verb i setningen: *Den gamle mannen syk. àDen gamle mannen er syk.*
- Plassering av adverb i leddsetninger: *Jeg går ikke ut hvis det slutter ikke å regne àJeg går ikke ut hvis det ikke slutter å regne*
- Plassering av subjekt og finitt verb: *Nå gutten kommer àNå kommer gutten*
- Pronomener i akkusativ: *Han ser på jeg àHan ser på meg*

2. Slik virker grammatikkontrollen

En grammatikkontroll må ha opplysninger om ordenes bøying og ordklassetilhørighet for å finne feil. Grunnen til dette er at man for eksempel ikke kan si noe om at *en* ikke kan stå sammen med *sykler* dersom man ikke har opplysninger om at *sykler* er et substantiv i flertall og at *en* er en entallsartikkel. Mange ord er flertydige slik som *sykler,* og brukeren bør ikke belemres med feilmeldinger som sier at *en sykler* er feil i en setning som: *en sykler gjerne til jobben når det er fint vær* siden *sykler* her er et verb. For å få en god grammatikkontroll må derfor setningene analyseres med hensyn til ordenes ordklasse og bøying.

La oss si at brukeren skriver "Det er fort gjort og skrive feil" og kjører grammatikkontrollen. Det brukeren får opp på skjermen, ser slik ut:

Figur 2: Skjermbilde for setningen: *Det er fort gjort og skrive feil*

Det som ligger bak denne feilmeldingen er følgende: Grammatikkontrollen fungerer på periodenivå, og ordene i hver periode blir først tildelt tagger med morfologisk informasjon som blant annet viser ordenes ordklasse og bøying. Informasjonen kommer fra en elektronisk ordliste. Legg merke til at ordene på dette stadiet får alle tagger som er mulige, det vil si at for eksempel *fort* får både verb-, adjektiv- og substantivlesninger:

Figur 3: Multitagget setning: *Det er fort gjort og skrive feil*

```
"<Det>"
    "det" pron pers 3 noeyt ent
    "det" det dem noeyt ent
```

```
"<er>"
    "være" verb pres a5 pr1 pr2 <aux1/perf_part>
"<fort>"
    "fore" verb perf-part
    "fore" adj <perf-part> m/f ub ent
    "fore" adj <perf-part> noeyt ub ent
    "fort" adj pos <adv>
    "fort" subst noeyt appell ub ent
    "fort" subst noeyt appell ub fl
    "forte" verb imp rl4
"<gjort>"
    "gjøre" verb perf-part tr1 rl9 pr3
    "gjøre" adj <perf-part> m/f ub ent tr1 rl9 pr3
    "gjøre" adj <perf-part> noeyt ub ent tr1 rl9 pr3
"<og>"
    "og" konj
    "og" konj clb
    "og" adv
"<skrive>"
    "skrive" verb inf tr1 i1 tr11 pa1 d1 pa5 pa3
"<feil>"
    "feil" adj pos m/f ub ent
    "feil" adj pos be ent
    "feil" adj pos fl
    "feil" subst mask appell ub ent
    "feil" subst mask appell ub fl
    "feil" adj pos noeyt ub ent
    "feile" verb imp i1 tr6
```

De taggene som ikke er korrekte i konteksten, elimineres deretter av en disambiguerende tagger. Taggeren som brukes, er en omarbeidet versjon av den taggeren som tidlige

Figur 4: Disambiguert setning: *Det er fort gjort og skrive feil*

```
"<Det>"
    "det" pron pers 3 noeyt ent
"<er>"
    "være" verb pres a5 pr1 pr2 <aux1/perf_part>
"<fort>"
    "fort" adj pos <adv>
"<gjort>"
    "gjøre" verb perf-part tr1 rl9 pr3
"<og>"
    "og" konj
"<skrive>"
    "skrive" verb inf tr1 i1 tr11 pa1 d1 pa5 pa3
"<feil>"
    "feil" adj pos noeyt ub ent
    "feil" adj pos fl
    "feil" adj pos m/f ub ent
    "feil" subst mask appell ub fl
    "feil" subst mask appell ub ent
```

Til slutt blir selve grammatikkontrollprogrammet kjørt, og eventuelle feil funnet. Grammatikkontrollen er laget etter en metode som kalles *Constraint Grammar* eller *føringsbasert grammatikk* (Karlsson et al 1995), og her vil vi vise noen av Constraint Grammar-regelene som er laget for å få grammatikkontrollen til å oppdage feil. Totalt inneholder grammatikkontrollen ca. 700 regler. Nedenfor er en forenklet versjon av regelen som leter etter feilen i eksempelsetningen. Til høyre for regelen står en forklaring på det regelen gjør.

Figur 5: Forenklet versjon av regelen som leter etter feil bruk av "og"

```
(@w =! (@ERR)                    feilmarkere alle forekomster av
(0 "OG"-KONJ)                    ordformen og hvis og er en konjunksjon,
(NOT -1 INF)                     hvis det første ordet til venstre ikke er en infinitiv og
(1C INF))                        hvis ordet umiddelbart til høyre kun har infinitivslesning
;TITLE «og» i stedet for «å»
```

Denne regelen ber programmet om å velge merkelappen (@ERR), dvs. feilmarkere alle forekomster av ordformen *og* som tilsvarer følgende kriterier:

- Hvis *og* er en konjunksjon og det første ordet til venstre ikke er en infinitiv (*Han vil spise og drikke)*
- Hvis ordet umiddelbart til høyre kun har infinitivslesning

Så gir grammatikkontrollen brukeren en melding om hvilken type feil som er funnet, ber brukeren om å kontrollere ordet (eventuelt ordene) der feilen er lokalisert, og kommer dersom det er mulig, med et rettingsforslag.

I eksempelsetningen *Det er fort gjort og skrive feil* får brukeren beskjed om at det skal være *å* i stedet for *og* foran *skrive*. I tillegg til den korte forklaringen kan brukeren få opp en lengre generell forklaring som i dette tilfellet redegjør for bruken av *og* og *å* foran infinitiv. Slike forklarende hjelpetekster er knyttet til hver feiltype:

Figur 6: Forklarende hjelpetekst for bruk av *og* i stedet for *å*

«og» i stedet for «å»

«Å» kalles «infinitivsmerke». Det innebærer at «å» skal etterfølges av et verb i infinitiv. «Infinitiv» er den formen av verbet som gjerne brukes når en skal «nevne» hvilket verb man snakker om. Derfor kalles infinitiven også for «oppslagsformen» eller «ordboksformen» av verbet. Eksempler på infinitivsmerke og verb i infinitiv er: «å være», «å skrive», «å synge» og «å hoppe».

Konjunksjonen «og» binder sammen to like ledd. Den skal dermed brukes mellom to infinitiver, slik som her: «Han liker å spille og synge». Ellers er det infinitivsmerket «å «som skal brukes foran infinitiv og ikke konjunksjonen «og»:

Feil: «Det er sunt og drikke vin.»
Riktig: «Det er sunt å drikke vin.»

Feil: «Jeg liker og spise is.»
«Riktig: «Jeg liker å spise is.»

Feil: «B-gjengen prøvde og sprenge pengebingen i lufta.»
Riktig: «B-gjengen prøvde å sprenge pengebingen i lufta.»

Infinitivsmerket «å «kan bare brukes foran verb i infinitiv. I de følgende eksemplene er derfor «å «feil brukt og konjunksjonen «og» må brukes i stedet:

Feil: «Jeg traff Per å Kari.»
Riktig: «Jeg traff Per og Kari.»

Feil: «Jeg sitter å leser avisa.»
Riktig: «Jeg sitter og leser avisa.»

Feil: «Jeg har sittet å lest avisa.»
Riktig: «Jeg har sittet og lest avisa.»

3. Svakheter ved grammatikkontrollen

Den største svakheten ved grammatikkontrollen er kanskje at den ikke finner alle grammatiske feil i en tekst, men bare leter etter feiltypene som ble beskrevet i figur 1. Dette betyr at selv om mange feil kan være luket ut etter at grammatikkontrollen er kjørt, kan teksten fremdeles inneholde feil, for eksempel kommafeil. Til tross for at feiltypene listes opp under *Stavekontroll og grammatikk* i *verktøy*-menyen, vil noen brukere sikkert forvente at en grammatikkontroll skal finne *alle* grammatiske feil i en tekst, slik som stavekontrollen finner nesten alle stavefeil. Dessverre er det per i dag umulig å lage en grammatikkontroll som finner alle mulige feil, noe brukerne nødvendigvis ikke er klar over.

3.1 Manglende feilmelding på grunn av flertydighet

Noen ganger overser grammatikkontrollen også feil den egentlig er laget for å finne. Årsaken kan være flertydighet. Som vi nettopp så, fant grammatikkontrollen feilen i setningen *Det er fort gjort og skrive feil*. Feilen i denne setningen vil derimot ikke bli påpekt: *Det er fort gjort og rette feil* fordi *rette* fremdeles er flertydig etter at teksten er disambiguert, det vil si at her står *rette* igjen med adjektiv- og substantivlesninger i tillegg til verblesningen:

Figur 7: Flertydighet etter disambiguering: *Det er fort gjort og rette feil*

```
"<Det>"
      "det" pron pers 3 noeyt ent
"<er>"
      "være" verb pres a5 pr1 pr2 <aux1/perf_part>
"<fort>"
      "fort" adj pos <adv>
"<gjort>"
      "gjøre" verb perf-part tr1 rl9 pr3
"<og>"
      "og" konj
"<rette>"
      "rett" adj pos fl
      "rette" verb inf tr1 tr11 pa1 d5 d5/til a7 rl9
      "rette" subst mask appell ub ent
      "rette" subst fem appell ub ent
"<feil>"
      "feil" adj pos noeyt ub ent
      "feil" adj pos fl
      "feil" adj pos m/f ub ent
      "feil" subst mask appell ub fl
```

Uten noen klar infinitivsindikator som modalverb eller infinitivsmerke i perioden, har ikke den disambiguerende taggeren klart å entydiggjøre *rette* som infinitiv fordi taggeren ikke "forstår" teksten den skal disambiguere. I tilsvarende grammatisk kontekst kunne *rette* også ha vært adjektiv slik som her: *Læreren sa at det er pent tegnet og rette linjer overalt*.

Som vi så tidligere, krever grammatikkontrollregelen for denne feiltypen (se figur 5) at infinitiven skal være entydig etter *og*, og dermed får ikke setningen *Det er fort gjort og rette feil* noen feilmelding. Dette er et valg vi har tatt fordi vi ikke ønsket at brukeren skulle få falske feilmeldinger, for eksempel i setningen med *rette linjer*. Men ulempen er altså at noen opplagte feil ikke får noen feilmelding.

Flertydighet fører også til at grammatikkontrollen ikke oppdager at det er et finitt verb etter hjelpeverbet i dette eksempelet: *Han kan biler*. I eksempelet *Hun vil spaserer* avsløres imidlertid feilen med regelen i figur 8:

Figur 8: Forenklet regel som ser etter finitt verb etter hjelpeverb

```
(@w =! (@ERR)                    feilmarkere alle forekomster av ordet
(0 PRES/PRET)                    hvis det er et verb i presens eller preteritum
(NOT 0 IKKE-VERB)                hvis ordet ikke har annet enn verblesninger
```

```
(NOT 0 INF)                          hvis ordet ikke er infinitiv
 (NOT 0 PERF-PART)                   hvis ordet ikke er perfektum partisipp
(-1C M-HJ-VERB))                     hvis det står et modalverb rett til venstre for ordet
;TITLE Finitt verb etter hjelpeverb
```

Denne regelen sier altså at ordet skal feilmarkeres dersom:

- ordet er et verb i presens eller preteritum
- ordet ikke har noe annet enn verblesninger
- ordet verken er infinitiv eller perfektum partisipp
- det står et modalverb rett til venstre

Til slutt gir grammatikkontrollen en feilmelding som sier at dersom et verb styres av et modalverb, bør det stå i infinitiv. I setningen *Han kan biler* finner ikke grammatikkontrollen denne feilen fordi *biler* fremdeles er flertydig etter disambigueringen, og ved nærmere ettertanke er det vel også en flertydig setning?

En god grammatikkontroll bør selvsagt ha god *recall*, det vil si at den finner så mange feil som mulig i teksten. Samtidig bør *presisjonen* være høy, med andre ord bør så mange av feilene som mulig være reelle feil og ikke falske feil slik at brukeren blir alarmert i utide. Vi har testet over 4 millioner ord for å forsøke å finne den rette balansen. Presisjonen til grammatikksjekkeren er per i dag 75 prosent regnet ut fra et korpus på 890 000 ord fra avisene *Nordlys* og *Sarpsborg Blad*.

Det er likevel viktig å være klar over at presisjonstallet kan variere alt etter hvilke type tekst en tester og hvilke regler som er med i testen. 75 prosents presisjon er oppnådd i en test der det kun er reglene fra grammatikkontrollens *standardkontroll* som er testet. Velger en derimot *utvidet grammatikkontroll* fra *verktøy*-menyen, virker alle reglene som er laget, blant annet en regel som gir brukeren melding når det ikke er noe finitt verb i setningen. Når regelen for *ingen verb i setningen* er med i testen, og grammatikkontrollen kjøres på en tekst med mange verbløse overskrifter, øker presisjonen til 91%. Merk at disse tallene ikke tar hensyn til om feilmeldingen som gis til brukeren, er korrekt.

3.2 Feil med gal feilmelding

Ofte finner grammatikkontrollen en feil, men klarer ikke å gi riktig beskjed om hva som er galt. I frasen *de store kaffe koppene* burde diagnosen selvfølgelig vært at brukeren hadde skrevet *kaffe koppene* som to ord i stedet for ett. I stedet får brukeren denne feilmeldingen i figur 9:

Figur 9: Feil med gal feilmelding: *De drakk av de store kaffe koppene*

Siden grammatikkontrollen ikke forstår betydningen av hvert enkelt ord eller har evnen til å gjette hva skribenten har ment, må diagnosen bli at det ikke kan forekomme et substantiv i entall etter flertallsartikkelen *de*. Slike feilmeldinger kan sikkert virke frustrerende på mange brukere, men selv om grammatikkontrollen kan gi gal diagnose og feilmelding, finner den i alle fall feilen. Her er et annet eksempel: *Han kam i dag tidlig*. Her vil brukeren få beskjed om at setningen mangler verb selv om det egentlig er ordet *kom* som er feilskrevet som *kam*.

Figur 10: Feil med gal feilmelding: *Han kam i dag tidlig*

4. Konklusjon

En grammatikkontroll bør ikke gi brukeren for mange falske feilmeldinger, men samtidig bør den finne så mange feil som mulig. Dette er en balansegang. Under arbeidet med grammatikkontrollen har vi stadig stått overfor dette valget: Skal grammatikkontrollen gi flest mulig feilmeldinger og risikere at mange av feilmeldingene er falske, eller skal den bare gi sikre feilmeldinger og risikere at mange feil blir oversett? Mange falske feilmeldinger er forstyrrende og fører gjerne til at brukeren slår av hele kontrollen. Derfor har vi forsøkt å få antallet falske feilmeldinger ned til et minimum selv om dette i noen grad har gått ut over grammatikkontrollens evne til å finne reelle feil.

Vi har altså valgt å lage en grammatikkontroll som gir brukeren så få falske feilmeldinger som mulig fordi mange falske feilmeldinger både er irriterende og gir brukeren inntrykk av at grammatikkontrollen ikke er til å stole på.

Litteratur

Arppe, A. 2000: Developing a grammar checker for Swedish. I Nordgård, T. (red.)
Nodalida'99 Proceedings from the 12th "Nordiske datalingvistikkdager". Department of Linguistics, NTNU, Trondheim, 13-27.

Birn, J. 1999: Detecting grammar errors with Lingsoft's Swedish grammar checker.
I Nordgård, T. (red.) *Nodalida'99 Proceedings from the 12th "Nordiske datalingvistikkdager"*. Department of Linguistics, NTNU, Trondheim, 28-40.

Hagen, K., Johannessen, J. B. & Nøklestad, A. 2000: A Constraint-Based
Tagger for Norwegian. I Lindberg, C.-E. og S. Nordahl Lund (red.): *17th Scandinavian Conference of Linguistics, vol. I.* Odense: Odense Working Papers in Language and Communication, No. 19, vol I.

Hagen, K. & Johannessen, J. B. 1998: Disambiguering uten syntaks. *MONS 7.*
Utvalde artiklar frå det 7. Møtet om Norsk Språk i Trondheim 1997.

Karlsson, F., Vuotilainen, A., Heikkilä, Juha, og Anttila, A. (red.). 1995: *Constraint*
Grammar: A Language-Independent System for Parsing Unrestricted Text. Berlin: Mouton de Gruyter.

Kristin Hagen
Tekstlaboratoriet
Institutt for lingvistiske fag
Postboks 1102
N-0314 Oslo
kristin.hagen@ilf.uio.no

Pia Lane
Tekstlaboratoriet
Institutt for lingvistiske fag
Postboks 1102
N-0314 Oslo
p.m.j.lane@ilf.uio.no

---

[1] Hos Lingsoft var det Trond Trosterud som arbeidet med den norske grammatikkontrollen. Han hadde ansvaret for bl.a. testing, oppdatering av leksikon og for utviklingen av grammatikkontrollmodulen som finner feil i tegnbruk og talluttrykk. På Tekstlaboratoriet skrev Pål Kristian Eriksen utkastet til de forklarende hjelpetekstene.