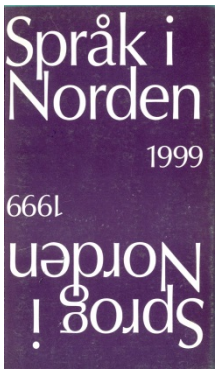
Sprog i Norden

Titel: Datamaskinell skrivestøtte

Forfatter: Koenraad de Smedt & Victoria Rosén

Kilde: Sprog i Norden, 1999, s. 20-32

URL: http://ojs.statsbiblioteket.dk/index.php/sin/issue/archive



Søgbarhed

Artiklerne i de ældre numre af Sprog i Norden (1970-2004) er skannet og OCR-behandlet. OCR står for ’optical character recognition’ og kan ved tegngenkendelse konvertere et billede til tekst. Dermed kan man søge i teksten. Imidlertid kan der opstå fejl i tegngenkendelsen, og når man søger på fx navne, skal man være forberedt på at søgningen ikke er 100 % pålidelig.

Datamaskinell skrivestøtte

Koenraad de Smedt & Victoria Rosén

Innledning

I dag inneholder nesten alle ordbehandlingssystemer standardverktøy for språklig skrivestøtte. For engelsk finnes det et utvalg av verktøy som tar seg av bl.a. automatisk stavekontroll, grammatikkontroll, orddeling, søking etter synonymer og oppsummering. For mindre språk, deriblant de skandinaviske språkene, er et mindre utvalg av verktøy tilgjengelig per idag. Noen særegne aspekter av de skandinaviske språkene, bl.a. produktivitet gjennom sammensetning og språklig variasjon gjennom subnormer, støttes enten ufullstendig eller ikke i det hele tatt av systemer som opprinnelig ble utviklet for andre språk.

Språklig støtte på datamaskin er et resultat av betydelig datalingvistisk forskning på åttitallet. Det ble bl.a. utviklet ulike algoritmer for gjenkjenning av feilstavede ord og "robuste" systemer for analyse av setninger med morfosyntaktiske feil. Mange prototyper av fullstendige systemer ble testet på brukere. Resultatene var blandet og flere prosjekter (f.eks. CRITIQUE) og produkter (f.eks. GRAMMATIK) ble stoppet. Noen resultater har likevel funnet veien inn i nittitallets ordbehandlingssystemer. Det pågår fremdeles mye aktivitet på dette feltet der nye aktører har kommet inn. Ofte er det små, spesialiserte bedrifter som tar opp utvikling av produkter for mindre språk, f.eks. WORDFINDER, MORPHOLOGIC, LINGSOFT og NYNODATA. I tillegg til programvarebedrifter jobber også forlags- og avisbransjen med egen utvikling av språkstøttesystemer, hovedsakelig beregnet til eget bruk.

At utsiktene for kommersielle anvendelser av korrekturlesningssystemer igjen er lysere, lar seg delvis forklare av at pc-er endelig har fått tilstrekkelig kapasitet til å lagre ordlister som omfatter et helt språk. I tillegg har nittitallets webmani ført til at elektronisk informasjonsformidling, som i all hovedsak inneholder språklig informasjon, er blitt enda viktigere. Datastøttede

prosesser for skriving og formidling er blitt vanlige, og dermed kan også språkverktøy etter hvert bli en del av normal tekstbehandling. Tilsynelatende setter lesere pris på språklig kvalitet, og forfattere ønsker seg verktøy som støtter språklig kvalitetssikring. Selv om få språkbrukere er fullt fornøyd med dagens systemer, er det liten tvil om at disse systemene er i stadig utvikling og kommer til å bli bedre og kraftigere, iallfall for store språk med en betydelig markedsandel.

Avis- og forlagsbransjen har et klart behov for ulike former for automatisert skrivestøtte. Korrekturlesning er dyrt; mange bedrifter har derfor gått over til automatiske systemer for rettelse av skrivefeil, selv om de ikke er perfekte. Orddeling er en ganskje omfattende jobb i aviser med smale spalter, og overlates derfor ofte til automatiske systemer. Ved hjelp av riktige verktøy kan orddeling med fordel automatiseres, slik at vi ikke lenger får feilaktige orddelinger som f.eks. *has-tverk* eller *jaz-zutdannelse*. Avis- og forlagsbransjen har dessuten noen spesielle profesjonelle behov. Programmet IJLEX, utviklet ved Institutt for Journalistikk i Gamle Fredrikstad, advarer profesjonelle språkbrukere mot bruk av farlige ord, vanskelige ord eller moteord. Programmet er basert på en liste med ord og begreper med tilknytning til krenkelser fra de siste 25 års rettspraksis i Norge.

Når systemer for automatisk korrekturlesning for skandinaviske språk i fremtiden etter forventning oppnår en bedre kvalitet, vil folk flest sikkert komme til å bruke slike systemer privat såvel som i det meste innen kommersiell og offentlig tekstproduksjon. For bruk på skolen kan man dessuten tenke seg at skrivestøtte i overskuelig fremtid integreres med systemer for språklæring, slik at elevene får språklig veiledning og opplæring tilpasset personlig språklig ytelse og behov. Erfaringer som IBM har gjort, har forresten vist at det er store forskjeller mellom ulike brukergrupper med hensyn til i hvilken grad de setter pris på skrivestøtte. Undersøkelser med IBMs tidlige engelske system Critique viste bl.a. at førsteårsstudenter var mest tilfreds og fant gjennomsnittlig 72 % av systemets kommentarer

korrekte. Minst tilfreds var profesjonelle forfattere, som fant bare 39 % av kommentarene korrekte. Skribenter med engelsk som andrespråk fikk 54 % korrekte kommentarer men fant likevel at hele 87 % av kommentarene var informative og brukbare (Richardson & Braden-Harder, 1988).

Korrekturlesning er alltid normerende, enten denne prosessen utføres manuelt eller ved hjelp av datamaskiner. De fleste ønsker nok å holde seg stort sett til den offisielle rettskrivningen, men det vil ikke si at vedtakene i Norsk språkråd uten videre kan brukes som grunnlag for korrekturlesning. På den ene siden vil skribenter gjerne kunne bruke nyord, lånord, navn og nye sammensetninger. På den andre siden inneholder begge offisielle skriftspråk store variasjonsmuligheter som ingen ønsker å blande vilkårlig. Den norske skriftspråksituasjonen er såpass komplisert at det er grunn til å tro at det ikke finnes språkteknologiske systemer i utlandet som tar vare på disse behovene. Ut fra et ønske om å utvikle bedre tilpassede verktøy ble et omfattende forskningsprosjekt startet i 1996 med EU-støtte. Partnere i dette prosjektet, som går under navnet SCARRIE, er Uppsala Universitet, Center for Sprogteknologi i København, Universitetet i Bergen, WordFinder og Svenska Dagbladet. Erfaringer fra dette prosjektet, som tar sikte på å utvikle korrekturlesningsverktøy for skandinaviske språk (norsk, dansk og svensk), vil fungere som eksempler i resten av denne artikkelen.

Denne artikkelen fokuserer på korrekturlesning. Selvsagt bør skrivestøtte ikke begrenses til korrekturfunksjoner. Søking etter synonymer er en annen slags skrivestøttefunksjon mange setter pris på og flere tesaurusbaserte systemer er allerede i bruk. Systemer for å lage automatiske sammendrag har et stort potensiale men er i dag bare brukbare for å få et grovt inntrykk av tekstens innhold, ikke for å forfatte et publiserbart sammendrag. For å illustrere dette, har vi brukt Microsofts program for engelsk på en tekst om det norske språket skrevet av Lars Vikør. Denne teksten var på nesten 1 600 ord, dvs. ca. 4 normale

sider, og ble redusert til ca. 150 ord eller 10 prosent. Sammendraget inneholdt bl.a. følgende avsnitt:

> *Christian III's Danish Bible was introduced in Norway, too; no Norwegian translation was made until modern times. This became the basis of the variety called Nynorsk.*

Selv om systemet er i stand til å plukke ut setninger med en viss grad av relevans, er det tydelig at strategien for å sette sammen en ny tekst ikke er særlig vellykket.

Korrekturlesning: problemer og strategier

Automatisk korrekturlesning er i dag den viktigste funksjonen i skrivestøttesystemer. Det er store forskjeller mellom mennesker og maskiner når det gjelder hvilke feiltyper de er i stand til å håndtere. Repetisjon av et ord er en hyppig feil som merkelig nok ofte er vanskelig å oppdage for en menneskelig korrekturleser. En maskin kan derimot oppdage gjentatte ord lynraskt. Derimot ville det være mye vanskeligere for en maskin å oppdage at det mangler et ord i *skipfartsnæring som lever å frakte andres handel*, mens en menneskelig leser stusser over det og skjønner hvilket ord som mangler. Dagens systemer kan raskt søke etter ord men har vanskelig for å avgjøre i hvilken kontekst et ord skal eller ikke skal brukes.

Skrivefeil kan deles opp i ulike typer avhengig av ulike kriterier, bl.a. eventuell opprinnelse og mulige korreksjonsstrategier. Typografiske feil er feil som oppstår ved bruk av tastaturet. Slike feil kan for eksempel bestå av forveksling av to bokstaver (f.eks. *øybelikk* istedenfor *øyeblikk*). Ortografiske feil, derimot, oppstår ved manglende kunnskap om ordets stavemåte (f.eks. *prinsippiell* istedenfor *prinsipiell*). Selv om denne forskjellen ikke alltid er klar – den siste feilen kunne også ha vært en typografisk feil – er det viktig å skille mellom disse feiltypene. Typografiske feil er nemlig lettere for forfatteren å oppdage enn ortografiske feil. I tillegg gir ortografiske feil ofte et dårligere inntrykk på leseren enn typografiske. Derfor er det generelt sett

viktigere å kunne oppdage ortografiske feil automatisk enn typografiske, og det er viktigere å kunne rette dem siden forfatteren vangligvis ikke vet hva den riktige stavemåten er.

Ortografiske feil har nesten alltid en uttale som ligner det påtenkte ordets uttale. Slike feil er ikke bare frekvent i barns skriftspråk (f.eks. *sentimeter* istedenfor *centimeter*, *sjøre* istedenfor *kjøre*, fra en undersøkelse på NTNU) men også i mediene (f.eks. *papparazzi, pappariziene*). Strategier for å rette særlig ortografiske feil kan med fordel bruke søkemekanismer basert på fonologisk likhet, mens typografiske feil oppdages best med søkemekanismer basert på grafemisk likhet (se oversikt i De Smedt & Van Berkel, 1998). Mekanismer som kombinerer fonembaserte og grafembaserte søkemekanismer har vist gode resultater (Vosse, 1994). Alle strategier har en større ytelse jo lengre ordet blir. Feil i veldig korte ord (f.eks. *emd* istedenfor *med*) er vanskelige å rette siden en liten feil kan gi et forholdsvis stort avvik, både fonemisk og grafemisk, fra det påtenkte ordet.

For å kunne oppdage en skrivefeil og foreslå en passende korreksjon er det avgjørende om et feilstavet ord består av et ukjent ord eller et eksisterende ord. Er det siste tilfelle (f.eks. omen time istedenfor om en time), er feilen umulig å oppdage, med mindre man foretar en analyse av konteksten, gjerne en syntaktisk analyse eller til og med en semantisk og pragmatisk analyse. Selvsagt gjelder dette også vanlige feil i bruk av kjente ord, f.eks. brant istedenfor brent eller lagt istedenfor ligget, grammatisk inkonsistens, avbrutte setninger og andre feil som går utover ordnivået.

Selv om en dekkende behandling av slike feil er utenfor rekkevidde i dag, finnes det typer av grammatiske feil som lar seg til en viss grad oppdage og rette ved hjelp av teknikker basert på kontekst. Noen systemer kan korrigere en predefinert rekke ord til en annen rekke ord. På denne måten kan f.eks. den hyppige feilen *vel og merke* rettes til *vel å merke*. På samme måte kan f.eks. stavemåten *accent* godkjennes i uttrykk som *accent aigu, accent grave*, osv., men ellers rettes til *aksent*. Slike mekanismer

med begrenset kontekst er likevel ikke tilstrekkelig for å oppdage og rette de fleste grammatiske feilene.

Kongruensfeil er en type grammatiske feil som er vanligere i norske tekster enn mange tror. Selv om nordmenn nesten aldri ville si *den IT-teknologisk forskning*, oppstår slike editeringsfeil gjerne når man endrer på en del av en setning uten å kontrollere om hele setningen fremdeles stemmer. For å kunne rette kongruensfeil brukes det "robuste" programmer som kan takle setninger med morfosyntaktiske feil (Vosse, 1992). Arbeid med Scarrie har ikke bare vist at slike systemer er brukbare for å oppdage mange typer av kongruensfeil, men også at det er ulike preferanser for hvordan programmet skal foreslå å rette dem (*den IT-teknologiske forskningen* eller *IT-teknologisk forskning*).

En kompliserende faktor for behandling av grammatiske feil er at det finnes flere systemer for genus med hensyn til bøyning og kongruens mellom artikkel, adjektiv og substantiv. Det er tillatt å skrive både *en liten bok* og *ei lita bok*. Selv om det ikke finnes offisielle regler om konsistens, ville nok de fleste som skriver *ei lita bok* i en tekst foretrekke å skrive *boka mi* i den samme teksten og ikke *boken min*. For de som ønsker støtte for slik konsistens er det mulig å definere ulike sett med kongruensregler som er gjensidig utelukkende. Kongruensregler er avhengig av en ordliste der alle ordformer står oppført med utførlig informasjon om ulike subnormer og kongruensmuligheter.

Mange typer av kunnskap må legges inn og kombineres for å klare ulike oppgaver innenfor korrekturlesning. Grammatikkregler og en ordliste er allerede nevnt. Det viktigste er antagelig at alle ordformene i språket blir gjenkjent og at systemet har nøyaktig informasjon om hver ordform. Ethvert tilfelle der datamaskinen klager på et ord som faktisk er riktig, er forvirrende eller irriterende for brukeren. Selvsagt skal datamaskinen ikke bare gjenkjenne oppslagsord som står oppført i en vanlig ordbok, men også alle bøyningsformer. Det finnes ulike løsninger for dette, som enten består av at systemet analyserer ordformene når teksten blir korrekturlest, eller at man legger alle

bøyningsformer inn i systemets ordliste på forhånd. I de fleste systemer blir den siste løsningen valgt på grunn av at tidsbesparelse er viktigere enn plassbesparelse.

I begge tilfeller er man i stor grad avhengig av omfattende ordlister som inneholder nøyaktig og tilstrekkelig informasjon. Scarrie bygger på ordlister utviklet av NorKompLeks-prosjektet ved NTNU. NorKompLeks var det første prosjektet som forsøkte å lage en liste med alle bøyningsformer på grunnlag av materiale fra eksisterende ordbøker. Naturligvis vil det alltid forekomme feil i store ordbøker. Men den samme feilen kan være mye alvorligere i en elektronisk ordliste som brukes av et dataprogramm enn i en trykket bok som leses av et menneske. F.eks. er verbet *stette* kodet i NorKompLeks som i Bokmålsordboka med bøyningskoden v2 istedenfor den riktige v1. Menneskelige brukere vil sannsynligvis aldri oppdage denne feilen og vil ikke komme til å bøye ordet feil på grunn av den. Når slike feil uten videre overføres til en ordliste med alle bøyningsformer vil flere feilaktige former, som f.eks. *stettt* eller *stettte*, lett kunne dukke opp som rettelsesforslag.

Produktive sammensetninger er en spesiell utfordring på skandinaviske språk som på mange andre språk, mens engelsk slipper unna dette problemet. Man kan ikke regne med at mange sammensetninger, som f.eks. *elgprøve, datagruvedrift* og *eurotegn*, finnes som ferdige oppslag, verken i en vanlig ordbok eller i en elektronisk ordliste. På den ene siden må nye sammensetninger tillates og gjenkjennes. På den andre siden må man kunne oppdage typografiske feil der to ord tilfeldigvis er skrevet i ett ord (f.eks. *eromtrent* istedenfor *er omtrent*). Det er derfor nødvendig med detaljerte regler som beskriver såvel mulighetene som begrensningene for å lage nye sammensetninger. Slike regler blir for tiden utviklet ved Tekstlaboratoriet på UiO. Skrives likevel en sammensetning i to ord istedenfor ett (f.eks. *klone sauen*) så er denne feilen vanskelig å oppdage hvis delene også finnes som selvstendige ord. I så fall krever korrektur generelt sett en syntaktisk analyse. Nesten umulig å oppdage er sammensetninger som etter reglene er mulige, men ikke

tilsiktet (f.eks. *skalddyrssalat* istedenfor *skalldyrsalat* eller *skaldyrsalat*).

Korrekturlesning og skriftspråknormer

Alle de ovennevnte mekanismene, dvs. ordsøking, kongruensrettelse og gjenkjenning av sammensetninger, må ta hensyn til forfatterens "stil" eller bruk av en bestemt subnorm. Ta f.eks. denne setningen som inneholder en typografisk feil:

En hohtlønnet lærer får sjelden vondt i maven.

Både *høgtlønnet* og *høytlønnet* er mulige korreksjoner i bokmål (selv om ingen av dem ble foreslått i den aktuelle versjonen av Microsoft Word). Likevel ville det ikke være passende å foreslå *høgtlønnet* som korreksjon, siden denne "radikale" formen passer dårlig i denne setningen som ellers er skrevet på en "konservativ" subnorm. På norsk finnes det ikke noen lettvint løsning på dette problemet. På engelsk går de fleste tekstbehandlingssystemer ut fra et klart skille mellom en britisk og en amerikansk ordliste, men denne løsningen er tilsynelatende ikke tilstrekkelig for norsk. Ved hjelp av et antall urelaterte ordlister er det mulig å oppdage en ordvariant som ikke passer i den ønskede subnormen, men det er ikke alltid mulig å foreslå en passende korreksjon siden variantene kan være svært forskjellige. Variantene av et ord må være eksplisitt relatert til hverandre i ordlisten for at systemet skal kunne foreslå en pålitelig korreksjon i alle tilfeller.

Forøvrig går variasjonen utover ordnivået. På bokmål finnes det tre hovedsystemer for genus med hensyn til bøyningsendelser og kongruens i nominalfraser. I tillegg til et komplett tregenussystem finnes det et ofte brukt to-genussystem som verken har feminine bøyninger eller determinativer. Dessuten finnes det et system med feminine bøyninger og etterstilte determinativer, dog ikke feminin ubestemt artikkel, altså et system som befinner seg mellom to og tre genus. Dette oppsummeres i tabellen nedenfor.

3	2,5	2
*en (liten) bok	en (liten) bok	en (liten) bok
ei (lita) bok	*ei (lita) bok	*ei (lita) bok
boka (mi)	boka (mi)	*boka (mi)
*boken (min)	boken (min)	boken (min)

Legg merke til at det ikke er mulig for et korrekturlesningssystem å oppnå konsistens i en tekst ved å bare erstatte en form med en annen form. Særlig i 2,5-systemet finnes det en større frihet siden både feminine og maskuline bøyninger av feminine substantiver er akseptable. Det er dog visse prefereranser. F.eks. er bruk av formen *boka* relativt mer frekvent enn *mora*; derfor impliserer bruk av *mora* vanligvis at også *boka* brukes, men ikke nødvendigvis omvendt.

Innenfor Scarrie-prosjektet er det blitt foretatt utførlig arbeid med hensyn til leksikalsk og annen variasjon relatert til subnormer, siden de fleste brukere av automatisk korrekturlesning, særlig profesjonelle brukere i avis- og forlagsbransjen, setter pris på slikt. Ofte ønsker man å holde seg til en stil som er betraktelig trangere enn alt som er tillatt i bokmål. Omvendt må praktiske korrekturlesningssystemer også ta høyde for utstrakt bruk av noen ordformer som ikke er tillatt i den offisielle rettskrivningen (f.eks. *hverken*).

I dag finnes det ikke ordbøker som inneholder utførlig informasjon om subnormer, selv om språkbrukere klart har en tendens til å holde seg til slike subnormer. Den leksikalske variasjonen innenfor bokmål er enorm. Hvis en stamme med to mulige skrivemåter kombineres med to mulige endelser er det allerede fire ordformer; f.eks. er *melken, melka, mjølken* og *mjølka* alle tillatt på bokmål. Hver av disse kan godtas, men de har forskjellig stilverdi. Noen blir betraktet som mer "konservative" eller nærmere riksmål mens andre blir oppfattet som mer "radikale" eller nærmere nynorsk. I noen tilfeller blir det verre. Her er et eksempel på et verb som har tre mulige stammer; hver av dem kan kombineres med 11 endelser:

slokk	slukk	sløkk
slokka	slukka	sløkka
slokke	slukke	sløkke
slokkede	slukkede	sløkkede
slokkende	slukkende	sløkkende
slokker	slukker	sløkker
slokkes	slukkes	sløkkes
slokket	slukket	sløkket
slokkete	slukkete	sløkkete
slokt	slukt	sløkt
slokte	slukte	sløkte

Selv om alle disse formene er tillatt i bokmål, virker noen av dem kunstige og er sannsynligvis ikke i bruk hos noen forfattere. Formen *slukka* er for eksempel en kombinasjon av en konservativ stamme og en radikal bøyningsendelse. Omvendt er *sløkkede* en kombinasjon av en radikal stamme og en konservativ bøyningsendelse. De kombinatoriske konsekvensene av en bred normering kan være ganske dramatiske for muligheten til å bygge effektive språkteknologiske systemer. Poenget er at mye språkteknologisk arbeid er nødvendig for å dekke alle godkjente variasjonsmulighetene på tross av at ingen vil bruke disse mulighetene fullstendig. I diskusjoner om variasjon i offisiell rettskrivning er såvidt vi vet dette punktet ikke blitt tatt i betraktning.

Subnormering i forhold til korrekturlesning kan takles på ulike måter. Én måte er å identifisere alle ulike faktorer som viser variasjon, f.eks. stammer, substantivendelser, verbendelser, genussystemer, osv. Så lar man brukeren innstille egne preferanser for enhver språklig opsjon og fritt kombinere disse innstillingene. En annen måte er å definere brede, konsistente subnormer. I Scarrie-prosjektet har vi valgt å definere slike brede normer. Normene er basert på at ord har en stilverdi på en skala mellom radikal og konservativ, og at enhver ordform enten er godkjent som hovedform, likestilt form eller sideform, eller er en ikke-tillatt form i den offisielle rettskrivningen.

1. Nøytralt bokmål: bruk hovedformer, sideformer og likestilte former, unntatt alle former som er utpreget radikale eller utpreget konservative. Bare et fåtall veletablerte ikke-tillatte former godtas. Det brukes et 2,5-genussystem.
2. Konservativt bokmål: bruk hovedformer, sideformer og likestilte former som har en konservativ eller nøytral verdi, men ikke former som blir oppfattet som radikale. Ikke-tillatte konservative former (riksmål) godtas. Det brukes et 2-genussystem.
3. Radikalt bokmål: bruk hovedformer, sideformer og likestilte former som har en radikal eller nøytral verdi, men ikke former som blir oppfattet som konservative. Ikke-tillatte radikale former godtas. Det brukes et 3-genussystem.
4. Læreboknormalen: tillatt alle hovedformer og likestilte former, men verken sideformer eller ikke-tillatte former, uansett om ordene blir oppfattet som radikale, konservative eller nøytrale. Alle genussystemer er akseptable.
5. "Anything goes": tillatt alle former som er akseptable under stilene ovenfor. Dermed blir bare ikke-tillatte former med fremmede stavemåter (f.eks. *grease*) forkastet.

Legg merke til stilene 1 til og med 3 er definert slik at et automatisert system kan støtte en konsistent stil. Læreboknormalen, som er den eneste offisielle subnormen innenfor bokmål, er derimot ikke definert slik at konsistens er mulig. Læreboknormalen inkluderer dermed noen former som ikke er i bruk i noen stiler (f.eks. *fordyping*) og tillater også blandinger av radikale og konservative orddeler, f.eks. *rødgraut*. Basert på de ovennevnte normene har det blitt utviklet en ordliste som inneholder informasjon om hvordan enhver ordform eventuelt kan oversettes til varianter under en viss stil.

Denne ordlisten virker i takt med en grammatikk som tar seg av kongruens avhengig av genussystemet. Et slikt system, som samtidig skal ta vare på ordform, grammatikk og subnorm, blir raskt innviklet. For et feilstavet ord kan det være flere alterna-

tiver ut fra fonologisk og grafemisk likhet. For hver av dem skal det kontrolleres om ordet tilhører den valgte stilen eller skal erstattes med en mer passende variant. Siden korreksjonsforslaget kan være et ord som har andre syntaktiske trekk en det som passer i setningen, må kanskje andre ord i setningen også endres. Mange deler av dette høyst kompliserte systemet måtte spesialutvikles for norsk.

Konklusjon

Systemer for språklig skrivestøtte er nyttige når de kan øke effektivitet og kvalitet i skrivingen. Dagens systemer nærmer seg en grense der ikke bare naive språkbrukere men også profesjonelle skribenter vil ta dem i bruk. Fortsatt er mye forskning nødvendig for å nå dette målet. Noen aspekter av denne forskningen er rettet mot mer teknologisk sofistikerte løsninger. Man bør likevel ikke glemme andre, lingvistiske aspekter. Fremfor alt må arbeidet med pålitelige ordlister fortsette. Videre bør språkbrukernes operative normer studeres gjennom lingvistisk forskning på hvordan folk faktisk skriver og hvordan kompetente språkbrukere evaluerer det (Dyvik, 1998). Dette er nødvendig for å kunne utarbeide språklig intelligente systemer som er i takt med aktuelt språkbruk. Gjennom korpusbaserte studier og analyser av brukerbehov må vi finne ut hvilke ordformer som er i bruk og hvordan de grupperer seg i subnormer som har relevans for utvikling av skrivestøttesystemer. Fremtidig normering kan også dra nytte av slike undersøkelser. Dermed kan lingvistisk forskning bidra til et bedre samsvar mellom språkbruk, språknormer og språkteknologi. Det er verd å understreke at utvikling av språkteknologiske systemer for mindre språk ikke kan bestå av en enkel tilpasning av systemer som ble utviklet for verdens majoritetsspråk hvis alle nyansene av språket skal få aktiv støtte.

Referanser

De Smedt, K. & B. van Berkel, 1988. Triphone analysis: A combined method for the correction of orthographical and typographical errors. *Proceedings of the Second Conference on Applied Natural Language Processing, Austin* (pp. 77–83). Association for Computational Linguistics.

Dyvik, H., 1998. Hva bør en språktknologisk satsing inneholde? *Språknytt 4/98*. Norsk språkråd.

Richardson, S. D. & L. C. Braden-Harder, 1988. The experience of developing a large-scale natural language text processing system: CRITIQUE. *Proceedings of the Second Conference on Applied Natural Language Processing, Austin* (pp. 195–202). Association for Computational Linguistics.

Rosén, V. & K. de Smedt, 1999 (under utgivelse). SCARRIE: Automatisk korrekturlesning for skandinaviske språk. *Norsk språkvitskap: utvalde artiklar frå Mons 7.* Novus forlag, Oslo.

Vosse, T., 1992. Detecting and correcting morpho-syntactic errors in real texts. *Proceedings of the Third Conference on Applied Natural Language Processing, Trento* (pp. 111–118). Association for Computational Linguistics.

Vosse, T., 1994. *The word connection.* Neslia Paniculata, Enschede.