Lydskrivning over landegrænser

En studie i interskandinavisk genbrug af sprogteknologi

Af Peter Juel Henrichsen

Danish, Norwegian, and Swedish phonetics are in fact so similar that trilateral transfer of lexical and phonetic resources seems a linguistically well-motivated enterprise. Not only is the overlap of the three phone inventories substantial, the toneme systems of Swedish and Norwegian correspond closely to the Danish *stød* (an instance of creaky voice), and the principles of stress assignment are almost identical. These conditions suggest that a single Scandinavian phonetics accompanied by a small number of language-specific transformations could be effective for several purposes, linguistic as well as technological. In this paper we present a pilot experiment in transferability. We produced a phonetic transcription of the 900k word Norwegian speech corpus NoTa using an extensive phonetic database for Danish as a catalyst. The core components of the transfer system were the two algorithms NO2DO (Norwegian Orthography to Danish Orthography) and DP2NP (Danish Phonetics to Norwegian Phonetics) developed with standard machine learning techniques.

Indledning

Rent politisk bliver de skandinaviske hovedsprog opfattet som helt uafhængige af hinanden. I EU skal der oversættes ind og ud af dansk med samme intensitet som for svensk – og nordmændene vil næppe nøjes med mindre, hvis de følger trop en dag. Enhver tale om indskrænkning til ét skandinavisk fællessprog ville sætte den offentlige opinion i alarmberedskab; men det har formentlig mere at gøre med psykologisk integritet end med lingvistisk rationale. Materialiter er dansk og norsk hovedstadssprog næppe mere forskellige end to nationale dialekter, svensk kun en smule fjernere, og derfor kunne det fra et økonomisk og teknologisk synspunkt være en alvorlig overvejelse værd at udvikle en række grundlæggende lingvistiske ressourcer én gang i det skandinaviske sprogområde og så lade den sprogspecifikke forankring tage form af kompenserende småregler.

I det følgende præsenterer vi en konkret case, et pilotforsøg med anvendelse af dansk fonetik i et norsk lydskrivningsprojekt. Som vi vil demonstrere, ligner de to sprog hinanden så meget at man kan opnå en anvendelig fonetisk transskription af et norsk talesprogskorpus ved at genbruge eksisterende danske lydskrivningsalgoritmer. Bortset fra selve lydskriften har projektet genereret metoder og erfaringer som kan finde videre anvendelse i oversættelsessystemer og i interskandinavisk taleteknologi.

Artiklen begynder med en refleksion over formålet med at lydskrive. Derefter følger afsnit med introduktion af det norske talesprogskorpus samt de danske fonetiske ressourcer som vi vil bringe i spil. Inden for den germanske sprogfamilie består lydskrivning særligt i to aktiviteter: fonvalg og tryksætning. Disse to bliver derfor forberedt og kommenteret i en vis detalje. Så følger en præsentation af de faktisk anvendte algoritmer, samt en evaluering af det færdige resultat. Der er tale om et pilotforsøg, og vores konklusioner er måske nok foreløbige – men absolut optimistiske.

Hvorfor lydskrive?

En ortografisk transskription af det spontane talesprog er som et fotografi med grove raster. Billedet fungerer i en vis observationshøjde, men mange detaljer er forsvundet. Ortografien kan gøre rede for de leksikalske identiteter og deres rækkefølge, men kun i mindre målestok for det syntaktiske forhold mellem dem og slet ikke for betoningen, prosodien og udtalevariationen. Derfor er det en forbedring at have en lydbaseret transskription ved siden at den ortografiske. En fonetisk eller prosodisk repræsentation af talen har en anden og finere registrering, den er tættere på lyden og på den enkelte talers udtale, tættere på *parole* og fjernere fra *langue*. Kombinationen af ortografi og lydskrift giver derfor en mere dækkende repræsentation af talen end nogen af delene kan gøre enkeltvis.

Ideelt set skal en fonetisk repræsentation afledes ved aflytning og transskription, men desværre er deskriptiv lydskrivning overordentlig kostbar. Det kan let tage en time at lydskrive et minuts spontantale når man medregner alle kontrollytninger og drøftelser. Arbejdet stiller langt større krav til transskriptørens tålmodighed og ekspertise end den enklere ortografiske transskription som kun kræver identifikation af leksemerne. Derfor er kun de færreste store talesprogskorpora – slet ingen i Norden – forsynet med ægte aflyttet lydskrift.

Et ofte anvendt, langt billigere alternativ er automatisk genereret lydskrift. En genereret lydskrift bygger på forventninger og normalantagelser i stedet for aflytning, og den kan derfor ikke tage hensyn til variationen fra den ene taler til den anden. Dette udelukker naturligvis en del anvendelser, men alligevel kan den være et udmærket supplement til en ortografisk transskription. Ved at søge i en genereret lydskrift kan man for eksempel let samle eksempler på realiseringen af et givet fonetisk mønster. Desuden kan den genererede lydskrift tjene til forberedelse af et senere lydskriv-

ningsprojekt baseret på aflytning. Det er ofte hurtigere for en transskriptør at korrigere en eksisterende lydskrivning end at nyskrive hvert ord. For informanter uden særlige fonetiske særpræg kan transskriptøren endda nøjes med at godkende den lydskrivning der allerede findes, i alt fald for en god procentdel. I Danmark findes en hel del eksempler på talesprogskorpora som er forsynet med genereret lydskrift og på denne måde er blevet søgbare efter lydlige kriterier, sådan som fx det store korpus LANCHART (Gregersen 2007).

Det norsk-danske lydskrivningseksperiment

Det store norske spontantalekorpus NoTa (Norsk Talespråkskorpus, Bondi et al. 2008) blev præsenteret for det nordiske forskersamfund ved en konference i Oslo kort før nytår 2007. NoTa var transskriberet ortografisk og forsynet med PoS-tagging,[1] men det stod klart at man ikke ville få råd til en deskriptiv lydskrift skønt interessen ikke manglede. Derfor besluttede forfatteren at prøve en genvej, nemlig at lydskrive NoTa ved hjælp af eksisterende danske metoder i en forsøgsopstilling baseret på norsk-dansk og dansk-norsk transfer. På den ene side var udgangspunktet gunstigt, ethvert resultat ville jo være bedre end ingenting. På den anden side var projektet skræmmende, at lydskrive 900.000 ord – i størrelsesordenen 100 timers tale – på et yderst begrænset budget.

På Center for Computational Modelling of Language (CMOL, Copenhagen Business School) findes betydelige materialer for dansk af samme art som skulle bruges for norsk, nemlig store mængder af transskriberet talesprog, omfattende databaser med danske udtaler, applikationer til lemmatisering, morfologisk analyse og skrift-til-lyd-afbildning, samt ekspertise inden for taleteknologi. Projektets mulighed var altså at basere den norske lydskrivning på import fra Danmark. Som før nævnt er Oslo-norsk og Københavner-dansk tæt beslægtede sprog. Men tæt nok? Dette var det vigtigste spørgsmål i NoTaFon-projektets første fase.

Overvejelser og principper

Udrustet med de danske materialer – og med en norsk ordliste til verifikation – gik vi i gang med at udvikle en lydskrivningsalgoritme. Udviklingsarbejdet blev baseret på automatiske træningsprocedurer. Da det jo er sammenhængende tale og ikke kun isolerede ord der skal lydskrives, må

lydskriften også rumme oplysning om talens trykforhold. På norsk som på dansk spiller placeringen af hovedtrykkene en væsentlig rolle for talens prosodi. Hvis den fonetiske korpusversion virkelig skal give nye søgemuligheder, må det være muligt at søge efter prosodiske konturer og trykmønstre, ikke kun fonetiske symboler.

Vi delte derfor projektet op i to spor: tryksætning og fonsætning (vi taler om *foner* frem for *fonemer* da projektets formål er fonetisk, altså lydnær, transskription).

Som basis for tryksætningen valgte vi det lydskrevne danske PAROLE-korpus. Generelt er det naturligvis ikke særlig heldigt at anvende oplæsning (*scripted speech*) som udgangspunkt for annotation af spontantale; vi mente dog at det er forsvarligt til netop tryksætning, da fænomener som tryktab i komposita, morfo-syntaktisk betingede tryktab og enhedstryk er fælles for alle talesprogsgenrer. PAROLE omfatter såvel aflyttet lydskrift med trykmarkering som PoS-annotation i manuelt-verificeret kvalitet. Da ordklasseinformation er sprogneutral i højere grad end ortografisk information, var det naturligt at udtrykke og eksportere de danske trykregler på PoS-niveauet. Af den grund blev det nødvendigt at undersøge to forhold, for det første om den danske PoS-taksonomi og NoTa's PoS-taksonomi er forligelige, og for det andet om dansk og norsk tryksætning følger de samme principper. Hvis disse to forhold kunne bekræftes, var det tredje trin at etablere en række afbildningsregler fra trykmønstre til PoS-mønstre for dansk – og derefter anvende dem i modsat retning for norsk.

Det sværere problem med grafem-til-fon-afbildning måtte igen løses med træningsalgoritmer og dansk-norsk overføring. Her var hypoteserne at norsk ortografi kan afbildes på dansk ortografi – og dansk lydskrivning på norsk lydskrivning – begge dele med så stor sikkerhed at en væsentlig del af ordene i NoTa-transskriptionen kunne opspores i den danske fonetiske database DanPO, omsættes til dansk lydskrivning, og afbildes tilbage til norsk. Naturligvis kunne dette import-eksport-scenarie ikke forventes at fungere for *alle* ord, kun for de mest indlysende paralleller. Derfor måtte strategien under alle omstændigheder suppleres, dels med en håndskreven ordliste med de allerhyppigste og fonetisk uregelmæssige ord (bl.a. pronominer og konjunktioner), dels med en traditionel norsk-norsk skrift-til-lyd-algoritme. Sidstnævnte kunne vi udvikle i forlængelse af tilsvarende projekter for dansk (bl.a. i forbindelse med Gregersen 2007). For at gøre overgangen fra dansk til norsk lydskrift så glat som muligt valgte vi at bruge IPA-baseret transskription (International Phonetic Alphabet) i begge sprog, nærmere bestemt »grov IPA« (Grønnum 1998) i den – for

computerarbejde – meget praktiske notation der kaldes SAMPA (Speech Assessment Methods Phonetic Alphabet). SAMPA-notationen er udviklet netop med henblik på computerbaseret fonetisk repræsentation, analyse og afbildning mellem sprogene. Se definitionerne for den danske og (øst-) norske SAMPA i http://www.phon.ucl.ac.uk/home/sampa/.

Herunder opsummerer vi projektets arbejdshypoteser, ressourcer og praktiske begrænsninger. I de næste afsnit bliver de fem arbejdshypoteser undersøgt nærmere.

Arbejdshypoteser

1. Norsk og dansk kan beskrives med samme PoS-taksonomi
2. Principperne for norsk og dansk tryksætning er ækvivalente
3. Tryksætning er, i god tilnærmelse, en mange-til-en afbildning af PoS
4. Norsk ortografi kan, i god tilnærmelse, afbildes på dansk ortografi
5. Dansk fonetik kan, i god tilnærmelse, afbildes på norsk fonetik

Begrænsninger

1. Projektets mål er ikke-deskriptiv lydskrivning
2. Ingen mulighed for træning af en automatisk norsk lydskriver og tryksætter pga. manglende fonetisk træningsmateriale (»gold corpus«)
3. Begrænset mulighed for manuel oversættelse og lydskrivning

Ressourcer

1. NoTa (www.tekstlab.uio.no/nota/)
   - ortografisk transskription
   - PoS (automatisk annoteret[2])
2. NorKompLex – norsk lydskreven database (Nordgård 2000)
3. DanPO (Skadhauge et al 2005), dansk lydskrivningsapplikation med
   - kompositumanalyse
   - lemmatisering
   - overfladeregler til skrift-til-lyd-afbildning som fallback-strategi
4. Det danske PAROLE-korpus (Henrichsen 2007a) med bl.a.:
   - ortografiske tekstord og interpunktion
   - PoS (manuelt annoteret)
   - tryksætning (aflyttet)
   - indlæsning (1 mandlig speaker, 100k tokens)
   - lydskrivning (aflyttet)
5. Korpus OSLO, blandede tekstarter (www.tekstlab.uio.no/norsk/bokmaal/).

PoS for dansk og norsk

Den projekterede tryksætter byggede på en antagelse om at de danske og norske morfologiske taksonomier er nær-identiske. Skønt den antagelse ikke er kontroversiel, havde vi brug for et konstruktivt bevis, nemlig en effektiv *afbildning* mellem danske og norske PoS (the proof of the pudding is in the eating, som bekendt). Som empirisk materale kunne vi trække på dels det danske PAROLE-korpus, dels frekvensdata for det store Oslo-korpus af blandede tekstarter på bokmål.

Det danske PAROLE-korpus er manuelt opmærket med det pan-europæiske PAROLE-tagset (Keson 1999). PAROLE-tags er hierarkisk opbygget; hver tag består af en række bytes ordnet efter stigende specificitet sådan at den første er den mest generelle (hovedordklassen). For substantiver (hovedgruppe `N`) er ordningen for eksempel *kategori-underkategori-køn-tal-kasus-bestemthed*. Et ord som »partiernes« er annoteret med `NCNPG==D`, svarende til værdierne Noun–CommonNoun–Neuter–Plural–Genitive–*void*–*void*–Definite. De ubenyttede værdier (position 6 og 7) er det pan-europæiske systems fingeraftryk: De svarer til træk som dansk morfologi ikke koder for, mens fx finsk gør. Det gennemtænkte kodningssystem gør at man let kan undersøge morfologiske fænomener på tværs af de europæiske sprog. PAROLE-tags er kompakte, gode at arbejde med for datalingvisten, men – det skal indrømmes – temmelig anstrengende at læse.

Oslo-tags følger et andet notationsprincip med løsere ordning af de morfologiske træk og friere mulighed for underspecifikation. I Oslo-korpusset forekommer fx følgende PoS-annotation:

partienes »parti« subst appell nøyt be fl gen

der er så letlæst at den ikke behøver forklaring. Ordet i citationstegn er den leksikalske indgangsform.

Trods de to forskellige notationskulturer er det muligt at afbilde mellem DPoS (dansk PAROLE) og NPoS (Oslo-tags) uden større besvær eller informationstab i nogen retning. Både i teori og praksis er dansk og norsk morfologisk tagging altså stort set i en-til-en-forhold. De få undtagelser omtaler vi kort i det følgende. Vi vælger at se bort fra de små forskelle der er mellem NoTa-korpussets og Oslo-korpussets anvendelse af NPoS-taksonomien.

For enkelte kategorier er NPoS mere finkornet end DPoS, fx for proprier, der i begge systemer er markeret for kasus (genitiv/umarkeret), men i NPoS også for køn: Værdierne mask og fem bruges til personnavne, værdien nøyt til appellativer med konventionel propriumbrug, såsom »Stortinget«, »Senterpartiet« og »Middelhavet«. For proprier er NPoS-til-DPoS altså en mange-til-en-afbildning (dvs. lider et informationstab); se Tabel 1.

NPoS	DPoS	Kategori	Eksempel
subst prop	NP.-U==-	*Umarkeret*	»Norge«
subst prop gen	NP.-G==-	*Genitiv*	»Guds«
subst prop mask	NP.-U==-	*Maskulinum*	»Erik«
subst prop mask gen	NP.-G==-	*Maskulinum + genitiv*	»Benjamins«
subst prop fem	NP.-U==-	*Femininum*	»Anne«
subst prop fem gen	NP.-G==-	*Femininum + genitiv*	»Solveigs«
subst prop nøyt	NP.-U==-	*Neutrum*	»Middelhavet«
subst prop nøyt gen	NP.-G==-	*Neutrum + genitiv*	»Nordens«

Tabel 1 Afbildning NPoS-til-DPoS af proprier

På enkelte andre områder er DPoS rigere end NPoS. Eksempelvis markerer kun DPoS ordinaltal som sådan (type »syttende«) mens de i NPoS er analyseret som almene adjektiver.

De lukkede ordklasser er i en del tilfælde kodet forskelligt i DPoS og NPoS (fx konjunktioner, grammatiske partikler, adverbier), men uden at dette fører til større konflikter i afbildningen.

Kun i ganske få tilfælde er der tale om sprogrelaterede forskelle. At genus-kodningen for appellativer har tre værdier i NPoS og kun to værdier i DPoS, skyldes at norsk skelner mellem maskulinum og femininum mens denne forskel i dansk er kollapset til kategorien utrum (fælleskøn). »Bil« er således hankøn på norsk, fælleskøn på dansk.[3]

Hvad angår lydskrivningsprojektet, er det største problem ved de to systemer imidlertid en mangel som de begge har til fælles: Ingen af dem markerer *hjælpeverber* som sådan. Mere om dette i næste afsnit. Se også Nivre et al. (2007), som diskuterer anvendelsen af skriftsprogets PoS-taksonomi på spontantale.

Tryksætning

Tryksætning er et notorisk vanskeligt område af talesprogslingvistikken, vanskeligt ikke mindst fordi lingvister ofte tror det er let. Hvor ligger trykkene i sætningen »det ved jeg ikke«? Gennem årene har jeg spurgt mindst en snes danske lingvister og filologer. De fleste er ikke i tvivl, men deres sikkerhed afspejler ikke en faktisk enighed. Enigheden rækker stort set kun til ordet *jeg*, som næsten alle mener, er trykløst i en neutral oplæsning. Derudover har jeg høstet alle varianter af et, to og tre hovedtryk, placeret på *det*, *ved* og *ikke* i næsten enhver kombination.

Der kan naturligvis opstilles regler for tryksætning – fx disse tre skoleeksempler:

1) I præpositionsforbindelsen har kun ét led hovedtryk. Hvis styrelsen er et pronomen, bevarer præpositionen typisk sit leksikalske tryk, i andre tilfælde tabes trykket til styrelsen.

1a. jeg arbejder ^for $_0$ham *(^ er tryk, $_0$ er tryktab)*
1b. jeg arbejder $_0$for ^Bo

2) Hovedverbet i det transitive verballed taber sit tryk til det direkte objekt hvis objektet er ubestemt og samtidig mangler et udtrykt determinativ.

2a. hun $_0$køber ^øl
2b. hun har $_0$købt ^øl
2c. hun ^køber en ^øl
2d. hun ^køber ^øllen

3) Sammensatte proprier har kun ét hovedtryk idet kun det sidste led bevarer sit leksikalske tryk.

3a. ^Birte
3b. ^Weiss
3c. $_0$Birte ^Weiss

Selv om trykreglerne 1-3 (her let forenklede, se også Henrichsen 2004) er blandt de mest veletablerede både i den teoretiske lingvistik og i den almene sprogfornemmelse, kan de alle tre – og enhver anden leksikalsk-

grammatisk baseret trykregel for dansk – altid overtrumfes af den langt stærkere regel for *emfatisk* tryk.

4) Ethvert ord kan tage emfatisk tryk.

I en passende kontekst er det helt naturligt at flytte trykket fra »for« til »ham« i 1a herover, eller fra »Bo« til »for« i 1b, fx som svar på spørgsmålet »arbejder du sammen med Bo?«, og tilsvarende for eksemplerne i 2 og 3. Endda er billedet endnu mere diffust, idet emfase blot er en ekstrem grad af *prominens*, som på sin side forekommer i alle grader fra svag til stærk. Da prominens altså ikke er en binær egenskab, er det ikke så overraskende at aflyttere og afskrivere i mange konkrete tilfælde er uenige om et ord har tryk eller ej.

Kan danske trykregler genbruges for norsk?

Hvordan kan man, givet alle disse usikkerheder omkring dansk tryksætning, etablere sammenhængene mellem dansk og norsk? Til afklaring gennemførte vi et eksperiment med fem erfarne norske transskriptører som fik til opgave at vurdere tryksætningen i et lille testkorpus, Parole65, bestående af de første 65 tekstafsnit i PAROLE.[4] Det første tekstafsnit i Parole65:

> <To <kendte <russiske his<torikere An<dronik Mirgan<jan og <Igor <Klamkin <tror <ikke , at <Rusland kan <udvikles <uden en »<jernnæve«.

De fem deltagere blev bedt om at indsætte, fjerne og flytte tryktegn i teksten for at bringe den i overensstemmelse med norsk tryksætning. De blev opfordret til at være konservative, altså kun ændre uacceptable tryksætninger – dette dels for at begrænse opgavens frihedsgrader (som nævnt er tryksætning i sig selv ikke let) og dels for at undersøge den specifikke hypotese at danske principper for tryksætning kan benyttes uændrede for norsk.

Af grunde vi allerede har nævnt, havde vi to forventninger til svarene: Stor forligelighed mellem de to sprogs tryksætning og samtidig betydelig spredning i de norske vurderinger.

Vores forventninger blev ikke skuffet. Det viste sig at 849 ud af ialt 910 tokens (fraregnet interpunktionstegn) – svarende til 93,3% eller 14 ud af 15 – er uændrede hos alle fem deltagere. Reelt er overensstemmelsen endda

højere idet en del ændringer er rent leksikalsk motiveret. Hvis vi fraregner de trykstærke tokens hvor trykket blot er flyttet (originalens »<repræsentativ« er fx ændret til »repræ<sentativ« hos tre deltagere) eller antallet af tryk er justeret (»<der<efter« er ændret til »<derefter« hos to deltagere), når accept-raten op på 95,1% (865 tokens). I over halvdelen af tekstafsnittene (36 ud af 65) er tryksætningen uændret hos alle fem deltagere. Når man dertil lægger at ikke et eneste af de 910 tokens er ændret af alle fem deltagere, er der ingen umiddelbare tegn til større forskelle mellem dansk og norsk tryksætning. En nærmere dataanalyse bestyrker dette indtryk.

Uenighed	Antal tokens	Uenighedsprofil	Antal tokens
0	865	`.....`	865
1	30	**`....x`**	1
		`..x..`	7
		`.x...`	15
		`x....`	7
2	10	**`.x..x`**	3
		`.xx..`	1
		`xx...`	6
3	4	**`xx..x`**	3
		`xxx..`	1
4	1	**`xxx.x`**	1
5	0	-	-

Tabel 2 'Uenighed' = antal deltagere som har skiftet status for et token fra trykstærkt til tryksvagt eller vice versa. 'Uenighedsprofil' viser deltagerne 1-5 i rækkefølge, sådan at . betyder enig (dvs. bevaret trykstatus) og x betyder uenig (ændret trykstatus). Profiler der udelukkende repræsenterer ændringer fra tryksvag til trykstærk, er vist i fed font.

Bemærk i Tabel 2s tredje kolonne at deltager 4 er noteret for *nul* tokens ændrede fra trykstærk til tryksvag eller omvendt. Dette bør ikke tages som tegn på sløseri eller manglende skønsomhed, for samme deltager har i syv tilfælde ændret et ords tryk*placering* uden at forandre dets status som trykstærkt, heraf to som ingen andre deltagere har ændret (»<allerede« → »aller<ede« og »Mediterra<neo« → »Mediter<raneo«). Der er snarere tale om en mere konservativ holdning i den betydning vi indførte herover, end

hos fx deltager 1 (24 ændringer) og deltager 2 (39 ændringer). Deltagerne 3 og 5 ligger i midten af feltet med hhv. 14 og 8 ændringer, og dermed kan man ikke udpege én deltager som atypisk. Dette bekræftes også af at kun fem af Parole65's tokens er ændret af tre eller flere deltagere.[5]

Hvilke ændringer er mest signifikante, trykstærke ord ændret til tryksvage eller omvendt? Som diskuteret er den trykstærke variant næsten altid et gyldigt alternativ til den tryksvage i dansk (i en passende kontekst), mens det omvendte ikke gælder. Derfor synes danske stærktryk erstattet med norske svagtryk at være de mest informative til vores formål, fordi de belyser sprogforskellen klarest. Et konkret eksempel: I Parole65 forekommer navnet Birte Weiss i to tekstafsnit, begge med tryksætningen »Birte <Weiss«, altså kun tryk på efternavnet. Ingen af deltagerne har fjernet et tryk, mens fire af deltagerne har tilføjet et eller to:

Deltager 1	»<Birte <Weiss«	(..)	»<Birte <Weiss«
Deltager 2	»<Birte <Weiss«	(..)	»<Birte <Weiss«
Deltager 3	»Birte <Weiss«	(..)	»<Birte <Weiss«
Deltager 5	»<Birte <Weiss«	(..)	»Birte <Weiss«

Disse ændringer må være motiveret i semantisk-kontekstuelle snarere end i sprogtypiske forskelle (andre af Parole65's sammesatte proprier er uændrede af alle) og kaster altså ikke nyt lys over hypotesen. Noget tilsvarende gælder for hovedparten af ændringerne svag→stærk hos de fem deltagere. Som det ses i Tabel 2, er de mere signifikante ændringer stærk→svag imidlertid kun repræsenteret ved to deltagere, nemlig 1 og 2. De er altså ikke generelt udbredt, hvorfor vi heller ikke fra denne vinkel ser modargumenter mod det dansk-norske trykfællesskab.

Til sidst nogle sammenfattende bemærkninger. Netop fordi vi har begrundet tillid til deltagernes dømmekraft, er det interessant at en ekstremt stor del af testkorpus – 99,5% – er dækket af de *små* uenighedsprofiler hvor to eller færre deltagere har ændringer. Læg dertil at størsteparten af de norske ændringer er fuldgyldige danske alternativer, og konklusionen er klar: For det første må vi aldrig forvente enighed af vores transskriptører i tryksætningen af danske og norske tekster; for det andet skulle der være grønt lys for tryksætning af NoTa efter danske spilleregler.

Implementering af en dansk tryksætningsautomat

Der er ingen tvivl om at PoS-distribution og trykdistribution er korreleret.

For nogle grammatiske kategorier kan man opstille regler der gælder næsten undtagelsesfrit (så længe der ikke er emfase med i spillet): Substantiver er trykstærke, adjektiver er trykstærke, præpositioner med ikke-pronominel styrelse er tryksvage, partikler med særlige grammatiske funktioner er tryksvage (»at«, »om«, »som«, »der«), letledsadverbialer er tryksvage (»jo«, »da«, »vel«, »nok« osv.), numeralier er trykstærke, konjunktioner er tryksvage et cetera.

Den første version af den automatiske tryksætter blev derfor lavet ved, for hver symbol *P* i PAROLEs PoS-inventar, at optælle fordelingen af trykstærke og tryksvage *P*-ord i korpus. Hvis samtlige *P*-ord var tryksvage, indførtes reglen: »*P* koder for tryktab«. Hvis andelen af tryksvage lå mellem 50% og 100%, indførtes den samme regel, men nu gjaldt den naturligvis kun for en approksimation. I de resterende tilfælde bevaredes den leksikalske trykfordeling uændret, dvs. her valgtes reglen »*P* koder ikke for tryktab«.

Procedure for opstilling af trykregler

For hver grammatiske kategori *P* i PAROLEs PoS-inventar:

$|\text{tryksvage } P\text{-ord}| > |\text{trykstærke } P\text{-ord}| \rightarrow$ »*P* koder for tryktab«

$|\text{tryksvage } P\text{-ord}| \leq |\text{trykstærke } P\text{-ord}| \rightarrow$ »*P* koder ikke for tryktab«,

hvor $|X|$ står for antallet af *X* i korpus PAROLE.

Er tryksætning en funktion af PoS?

Den næste hypotese til evaluering var at (dansk) tryksætning kan beskrives som en mange-til-én-afbildning af PoS på trykgraderne trykstærk/tryksvag. Hvis afbildningen gjaldt uden undtagelser, ville den automatiske tryksætter fra sidste afsnit fungere perfekt.

Vi har dog allerede diskuteret to modeksempler, nemlig enhedstrykket (som i »$_0$køber øl« versus »^køber øllen«) og tryksænkningen i sammensatte navne (»$_0$Birte ^Weiss« versus »^Birte«) der begge afhænger af syntaktiske forhold, så reelt samler interessen sig om disse tre spørgsmål:

I. I hvor høj grad kan tryksætning implementeres som en funktion af PoS?
II. For hvilke PoS fungerer beskrivelsen dårligst?
III. Hvor findes de billigste forbedringer?

Evalueringen foregår ved at sammenligne PAROLE-korpusset i to udgaver, dels med de manuelt annoterede tryk (den såkaldte *gold standard*), dels trykopmærket med de simple PoS-regler fra sidste afsnit. Vi måler virkningsgraden for PoS-til-tryk-afbildningen over to parametre kaldet Konsekvens (K) og Forventede Fejl (FF).

$$K_P = \frac{|2t - a|}{a} \qquad FF_P = \frac{a - |2t - a|}{2}$$

I begge formler er a antal forekomster af tokens med PoS-tag P, hvoraf t er trykstærke i det manuelt annoterede PAROLE-korpus.

K-værdier ligger mellem 0 (minimal konsekvens) og 1 (maksimal konsekvens) for alle værdier af P. Hvis fx 100 tokens har tag *P'* (a=100) og samtlige forekomster er trykstærke (t=100), er $K_{P'} = 1$ som udtryk for at afbildningen *P'→trykstærk* er 100% præcis. Hvis samtlige forekomster er tryksvage (t=0), er igen $K_{P'} = 1$ idet afbildningen *P'→tryksvag* er præcis. Den minimale K-værdi indtræffer for t=50, fordi en PoS-til-tryk-afbilder i dette tilfælde ikke kan forvente at score bedre end den tilfældige tryktilskrivning. K-værdier kan altså tages som lingvistiske pejlemærker: For hvilke leksemer og grammatiske kategorier er tryksætning sværest at forudsige, dvs. mest informationsrig – og K-værdierne tilsvarende lavest?

FF-værdier måler det forventede antal fejl i tryksætningen af tokens med tag P for den optimale PoS-til-tryk-regel. FF-værdier ligger mellem 0 og t/2. Antag at 100 tokens har tag *P''*, heraf de 49 trykstærke. Den optimale trykregel er altså *P''→tryksvag*; men for 100 ukendte forekomster af *P''* kan alligvel $FF_{P''} = 49$ forekomster forventes at være tryksat forkert. Også for t=51 er den forventede fejl $FF_{P''} = 49$, men nu målt ud fra trykreglen *P''→trykstærk*. FF-værdier giver således information om hvor stort fejlvolumen hver PoS-regel repræsenterer og, følgelig, hvor revision af PoS-regler og lydskrift vil have mest synlig virkning.

Bedømt på K-værdien er tryksætning vanskeligst at forudsige for PAROLE-taggen `PP2C.N-NP` med 63 forekomster i PAROLE hvoraf 30 trykstærke, svarende til K=0,048; denne tag rummer kun én ordform, nemlig »De« – den høflige variant af pronominet »du«. Næstvanskeligst er `PP3NSU-NU` (a=2.838, t=1326, K=0,066), også med kun én ordform, nemlig »det«. Herefter følger `PP3.PN-NU` (a=849, t=393, K=0,074), igen med én ordform: »de«. Blandt de i alt 18 kategorier med K<0,5 er de ni pronominalformer. Vi finder derudover især verbalformer: imperativ (K=0,119), infinitiv (K=0,162), præteritum-aktiv (K=0,169), præsens-aktiv (K=0,335), perf. participium (K=0,362), og desuden proprier (a=8.359,

t=6164, K=0,475). Resten af kategorierne er atypiske (»udenlandske ord«) eller ubetydelige (*a*<10).

Vi konstaterer altså at pronominer, verber i alle former, samt proprier er vanskeligst at tryksætte mekanisk, hvad der igen tyder på at tryksætningen for disse kategorier bærer en betydelig semantisk information. Det bekræfter iagttagelser vi allerede havde gjort i afsnittet Tryksætning herover. Samtidig bemærker vi at de nævnte kategorier varierer betydeligt i størrelse, fx de to førstnævnte med *a*=63 (»De«) og *a*=2.838 (»det«).

Til sammenligning viser Tabel 3 de 10 kategorier med det største fejl-*volumen*.

PoS	Kategori	*a*	*t*	Regel	FF_{PoS}	K_{PoS}
VADR=----A-	verber, præsens	14.205	4.722	*tryksvag*	4.722	0,335
VADA=----A-	verber, præteritum	6.665	2.770	*tryksvag*	2.770	0,169
VAF-=----A-	verber, infinitiv	5.953	3.458	*trykstærk*	2.495	0,162
RGU	adverbier	13.391	10.945	*trykstærk*	2.446	0,635
NP—U=--	proprier	8.359	6.164	*trykstærk*	2.195	0,475
SP	præ-positioner	21.301	1.889	*tryksvag*	1.889	0,823
PP3NSU-NU	»det«	2.838	1.326	*tryksvag*	1.326	0,070
VAPA=S.I.-U	verber, perf.part.	3.878	2.641	*trykstærk*	1.237	0,362
PI-CSU--U	ubestemte pronominer, fælleskøn {»en«, »anden«, »nogen«, ...}	3.782	764	*tryksvag*	764	0,596
PI-NSU--U	ubestemte pronominer, intetkøn {»et«, »andet«, »noget«, ...}	1.932	508	*tryksvag*	508	0,474
Alle	*Summering over alle kategorier*	177.761	97.430	-	25.007	-

Tabel 3 Forventede fejl ved simpel PoS-til-tryk-afbildning, sorteret efter volumen (FF-værdi).

En forbedret tryksætningsautomat

Som det ses, repræsenterer *verber* og *proprier* en særlig stor del af fejlvoluminet; præsensformer alene står for næsten en femtedel af alle fejl. Her bør en redningsaktion sætte ind. *Adverbier*, RGU, er en slet defineret kategori i PAROLE-systemet og vanskelig at operationalisere, mens *præpositioner*, SP, faktisk er ret præcist tryksat (K=0,823) og kun bidrager mærkbart til fejlvoluminet på grund af det høje forekomsttal (*a*=21.301).

Smertensbarnet er – som altid – pronominet »det« der i sine mange grammatiske funktioner forekommer med alle grader af prosodisk prominens, såvel i oplæst tale som i fri diskurs.

Vi ønsker altså at indføre nogle supplerende tryksætningsregler for verber og proprier. Da det især er hjælpe- og modalverberne som giver de dårlige FF-værdier ved at gå imod den generelle tendens for verber til at være tryk*stærke*, indfører vi en særregel $VERB_{AUX} \rightarrow tryksvag$ for $VERB_{AUX} \in$ {VÆRE, BLIVE, HAVE, KUNNE, SKULLE, VILLE, MÅTTE, TURDE, BURDE} i alle bøjninger, mens øvrige verber følger reglen *VERB→trykstærk*. Selvom denne definition af »hjælpeverbum« er ret grov og ikke tager højde for fx den possessive brug af HAVE, er forbedringen markant (se Tabel 4). Vi forventer iøvrigt at kunne anvende særreglen for norsk også: Af Parole65's 52 $VERB_{AUX}$-forekomster er trykket kun ændret for to tokens, i hvert tilfælde af bare én deltager.

For NP-kategorien indfører vi en kontekstfølsom regel der kun appliceres for proprier efterfulgt af et proprium: $PROP_{PRÆ} \rightarrow tryksvag$. Denne særregel tager højde for tryktab i flerleddede personnavne af typen »$_0$Birte ^Weiss« og »$_0$Danny $_0$Lund ^Jensen«.

Med disse to justeringer falder det globale fejlvolumen fra 25.007 til 18.731, svarende til en forbedring fra 14,1% til 10,5% fejl. Bemærk i Tabel 4 de fine K-værdier for især hjælpeverber og for proprier der ikke efterfølges af proprier. Af pladsgrunde dækker tabellen kun tre PoS, men gevinsten er tilsvarende for de øvrige verbalformer.

PoS	Kategori	*a*	*t*	Regel	FF_{PoS}	K_{PoS}
VADR=----A-	hjælpeverber, præsens	8.338	623	*Tryksvag*	623	0,851
	øvrige verber, præsens	5.867	4.099	*Trykstærk*	1.768	0,397
VADA=----A-	hjælpeverber, præteritum	2.984	233	*Tryksvag*	233	0,844
	øvrige verber, præteritum	3.681	2.537	*Trykstærk*	1.144	0,378
NP--U==-	proprium efterfulgt af proprium	2.738	663	*Tryksvag*	663	0,516
	øvrige proprier	5.638	5.515	*Trykstærk*	123	0,956

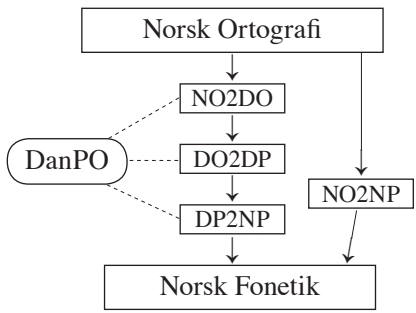
Tabel 4 Forbedrede FF- og K-værdier for uddifferentierede verbal- og propriumformer.

Naturligvis kunne man nå en mere præcis tryktilskrivning ved at supplere de PoS-specifikke regler med leksemspecifikke regler og erstatte de simple token-til-tryk-afbildninger med *n*-gram-baserede afbildninger eller andre kontekstsensitive regler (Henrichsen 2001). Leksemspecifikke regler ville imidlertid kræve en dansk ord-til-ord-oversættelse af NoTa, mens kontekstbaserede regler kræver en forudgående analyse af dansk-norsk syntaktisk overensstemmelse. Begge dele var uden for praktisk rækkevidde i det nuværende projekt. Derfor stiller vi os foreløbig tilfreds med tryk-reglerne herover. De giver et målbart kvalitetsniveau på ca. 90% korrekte hovedtryk hvad der virker fornuftigt i lyset af de uafklarede teoretiske præmisser for tryksætning. En fordel ved disse simple regler er at de kan forventes at være konstante hen over varierende sprogarter, hvad der er nødvendigt i betragtning af stilforskellen mellem PAROLE og NoTa.

Norske skrift-til-lyd-regler

I dette afsnit giver vi en kortfattet præsentation af den transfer-algoritme som vi har udviklet til lydskrivning af de ortografiske former i NoTa. Algoritmen består af adskillige delstrategier hvoraf den mindst konventionelle bruger danske leksikalske ressourcer til at generere lydskriften for de norske ord der har paralleller i dansk.

Som vist i figur 1 består den norsk-dansk-norske strategi af en følge af afbildninger. Hvert modul der indgår, har et teknisk navn sammensat af symbolerne *N* for 'norsk' (*Norwegian*), *D* for 'dansk' (*Danish*), *O* for 'ortografisk form' (*Orthographic form*), *P* for 'fonetisk form' (*Phonetic form*), *2* for 'oversættes til' (*to*), samt *COMP* og *RECOMP* for hhv. kompositum-analyse og -rekonstruktion (*Compound analysis*, *Re-composition*).



Figur 1 Principskitse af de vigtigste strategier i den norske skrift-til-lyd-afbildning.

Først afbildes en norsk ortografisk form på en dansk ditto (NO2DO). Hvis formen derefter kan genkendes i det danske leksikon eller lydskrives pålideligt (DO2DP), bliver den ekspederet videre til afbildning fra dansk til norsk fonetisk form (DP2NP).

Strategien giver gode resultater – de første er fremlagt i Henrichsen (2007b) – men kan naturligvis kun anvendes for norske former der ligger tæt på danske former. Derfor må den suppleres med andre strategier hvis algoritmen som helhed skal være komplet. Ellers vil *NoTaFon* blive fuld af huller. Af supplerende strategier er de væsentligste: NO2NP (norsk ortografi til norsk fonetik) og manuel lydskrivning. Desuden har vi implementeret en norsk kompositumanalyse kaldet NO-COMP, der inkluderer de hyppigst forekommende forstavelser og endelser, samt regler for konkatenering af stammer med nul-, s- og e-fuge.

Forholdet mellem dansk og norsk ortografi

Som dansker kan man opleve at læse sig et pænt stykke ind i en norsk tekst (på bokmål) før man opdager at den ikke er dansk. I de fleste tekster er langt de fleste norske ord identiske med deres danske oversættelser. Nogle få eksempler er »notat«, »notere«, »ignorerer« og »ignorant«. Andre ord afviger kun overfladisk fra de danske, såsom »infisere« (dansk *inficere*), »notasjon« (*notation*), »ignorerte« (*ignorerede*) og »ignorantene« (*ignoranterne*). Kun en meget lille restgruppe af norske stammer har slet intet dansk modstykke, hvad man kan forvisse sig om i enhver almindelig frekvensordliste. Disse observationer har vi taget som udgangspunkt for NO-2DO, hvoraf et lille udsnit er vist i Figur 2.

NO2DO – afbildning af norsk ortografi på dansk ortografi

Alfabetiske varianter

kj→k	kt→gt	øy→øj	au[dt]→ød

Morfologiske varianter

Substantiver:	skap→skab	sjon→tion	sak→sag
Adjektiver:	aktig→agtig	ert$→eret	a$→et

Figur 2 Afbildninger fra norsk ortografi til dansk ortografi ("$" markerer slutning af ordet).

I de tilfælde hvor et norsk ord via NO2DO kan afbildes på et genkendt dansk ord – evt. efter kompositumanalyse (NO-COMP) og gensammensætning (NO-RECOMP) – er den danske fonetiske form tilgængelig via CMOL's lydskrivningsapplikation. Derefter mangler kun en afbildning tilbage fra dansk fonetik til norsk fonetik. Også her er der betydelige ligheder mellem sprogene, men dertil nogle karakteristiske forskelle som DP-2NP må tage højde for. I norsk udtales klusilerne /p/, /t/, /k/ fx altid som [p], [t], [k], mens de på dansk ofte svækkes til [b], [d], [g], afhængigt af deres placering i stavelsen (se fx Grønnum 1998 og referencer dér).

	Norsk	*Dansk*
»titte«	[t"it0]	[t`id0]
»statsmakter«	[st'A:tsmAkt0r]	-
»statsmagter«	-	[sd`{:?dsmAgdC]
»straffbart«	[str"AfbA:rt]	-
»strafbart«	-	[sdr`AfbA:?d]

Figur 3 Norsk-danske lydskrifter. [ˈ] = norsk tonelag 1, [ˈˈ] = norsk tonelag 2, [`] = dansk hovedtryk, [?] = dansk stød, [:] = vokalforlængelse.

Det danske stød modsvarer i det store og hele det norske tonelag 1, dog med to vigtige forskelle:

- Adskillelsen af norsk tonelag 1 og 2 giver kun mening i ord med to eller flere stavelser. Der er ingen tilsvarende begrænsning for det danske stød.
- Dansk stød forekommer kun i to fonetiske kontekster: enten ved en lang vokal, eller ved en kort vokal efterfulgt af en sonorant konsonant; norsk har ingen tilsvarende begrænsning for tonelag 1 og 2.

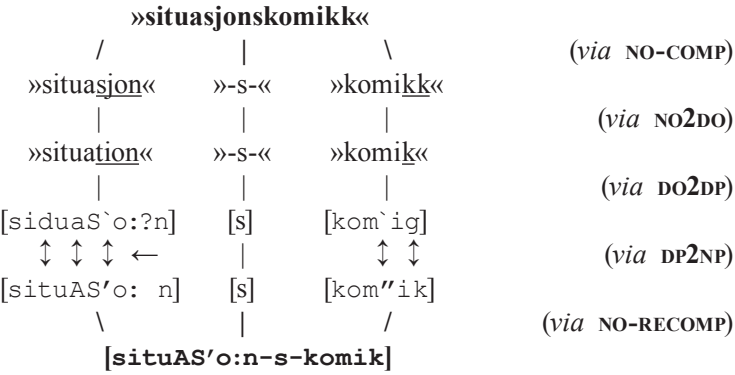
Disse og andre systematiske forskelle er der taget højde for i DP2NP-reglerne, hvoraf nogle af de mest produktive er vist i Figur 4.

DP2NP – afbildning af dansk fonetik på norsk fonetik

[{|a|A] → [A]	Danske »a«-foner afbildes alle på norsk [A]
[`*V*:?] → ['*V*:]	Dansk stød v/lang vokal → norsk tonelag 1
[`*VC*?] → ['*VC*]	Dansk stød v/kort vokal → norsk tonelag 1
[d] → [t] *iff* »t« i ortografi	Dansk »t« realiseret som [d] → norsk [t]
[D] → [d] *iff* »d« i ortografi	Dansk »d« realiseret som [D] → [d]
[g] → [k] *iff* »k« i ortograf	Dansk »k« realiseret som [g] → [k]
[w?] → [N] *iff* »gn« i ortograf	... som i »vogn«, »sagn«

Figur 4 Afbildninger fra dansk fonetik til norsk fonetik.

I Figur 5 herunder følger vi ordet »situasjonskomikk« gennem alle transformationer.



Figur 5 Eksempel på derivationshistorie: lydskrivning af et kompositum.

Læg mærke til at det danske [d] rekonstrueres som norsk [t] som signaleret i ortografien. Bemærk også at det danske stød [ʔ] korrekt selekterer norsk tonelag 1 [ˈ]. Det forkerte tonelag 2 [ˈˈ] signaleret af det manglende danske tryk i »komik« er heldigvis usynligt i resultatet, hvor kun første orddel bevarer sit leksikalske tryk. SAMPA-kyndige vil bemærke at vores valg af fonetiske symboler ligger tættere på det danske alfabet end det norske, naturligvis uden at dette har betydning for lydskrivningens kvalitet.

Nu er ingen kæde stærkere end sit svageste led, og hvert led i strategien NO2DO-DO2DP-DP2NP kan indføre en lydskrivningsfejl. Nogle fejl er usynlige i resultatet (som i eksemplet herover), mens andre viser sig som uregelmæssigheder i lydskriften. Det er naturligvis vigtigt at få bestemt både mængden og arten af lydskrivningsfejl; dette vender vi tilbage til herunder.

Traditionelle skrift-til-lyd-regler til supplement

Selv om langt de fleste norske ord i almindelig løbende tekst har danske ækvivalenter, er det ikke altid hensigtsmæssigt at oversætte dem ved hjælp af regler. En regel der beskriver sammenhængen mellem norsk »tvil« og dansk »tvivl« ville fx let kunne postulere en oversættelse af norsk »sal« til

dansk »savl«. Fremmedord er ofte bevaret med udenlandsk stavning i dansk (»bassin«, »orange«) mens norsk ortografi generelt er mere lydret (»basseng«, »oransje«). Og naturligvis findes der norske stammer som i dansk helt mangler eller er gået af brug (»kanskje«, »slik«).

I alle disse tilfælde giver translitteration mellem dansk og norsk ikke megen mening, og derfor omfatter oversættelsesalgoritmen også et traditionelt norsk skrifttegn-til-lyd-modul (NO2NP) udviklet med maskinlæringsteknikker, med Nordgård (2000) som reference.[6]

Desuden har vi indført en ordliste med manuelle lydskrivninger af de hyppigste ord i NoTa's frekvensliste som den øvrige algoritme ikke behandler tilfredsstillende. Det drejer sig især om de lukkede ordklasser: pronominer (fx »jeg« [j'Ei]), præpositioner (»med« [m'e:]), konjunktioner (»og« ['O:]) , numeralier (»tolv« [t'Ol]) og så videre.

En ytring fra det lydskrevne NoTa-korpus

Her præsenterer vi et citat fra NoTaFon. Hver linje repræsenterer et ord i ytringen »og det er ikke så vanlig å se sandblåste tømmerbygninger altså«. Den genererede lydskrivning er vist i kantparenteser [...] mens oplysningen i krølleparenteser {...} er derivationens historie. Derefter følger den ortografiske transskription, og resten af linjen udgør PoS-taggen.

[O:]	{M}	og	konj
[de:]	{M}	det	pron nøyt ent pers 3.
[Er]	{M}	er	verb pres
["ik0]	{D}	ikke	adv
[s'O:]	{D}	så	adv
[v"A:nli]	{D}	vanlig	adj nøyt ent ub pos
[O:]	{M}	å	inf-merke
[s'e:]	{D}	se	verb inf
[s"AnblO:st0]	{D+O}	sandblåste	adj be sup *(sic)*
[t'øm0rbygniN0r]	{D+D}	tømmerbygninger	subst mask fl appell ub
['AlsO]	{D}	altså	adv

Figur 6 Eksempel fra NoTaFon.

{M} betyder manuel lydskrivning – disse ord er aflæst direkte i den supplerende ordliste omtalt ovenfor. {D} er lydskrivning baseret på en dansk

ækvivalent ordform mens {O} er afbildning fra norsk ortografi direkte på norsk lydskrivning.

Som eksemplet viser, er NoTaFon hæderlig, men ikke perfekt. Der er fejl både i lydskrivning og PoS-tagging (jf. *sic* i Figur 6). Med de nuværende simple regler er tryksætningsalgoritmen blind for fænomenet enhedstryk, og derfor er »se« blevet lydskrevet med tryk, [s'e:]. Dette er vel ikke en umulig realisering, men giver en fornemmelse af kontrasttryk eller anden emfase som ikke har noget belæg i ytringen selv. En bedre tryktilskrivning ville undlade hovedtrykket på »se« med henvisning til at det grammatiske objekt, »sandblåste tømmerbygninger«, er nøgent – altså ubestemt og uden determinativ (fx Mørch 2006 og referencer der). Dette ræsonnement er imidlertid uden for rækkevidde af de nuværende tryksætningsregler, som ikke omfatter syntaktisk analyse.

Til gengæld er »er« (formentlig) korrekt tryksænket takket være særreglen for hjælpeverber (se Tabel 4). »Så« og »altså« lyder, med tryk, mildt emfatiske, uden at de dog helt kan afvises som mulighed; her må aflytning til. »Sandblåste« er korrekt analyseret af NO-COMP som sand+blåst/e, og de to dele er hhv. behandlet af NO2DO-DO2DP-DP2NP (»sand« kan genkendes som et dansk ord) og NO2NP (»blåst« bliver ikke genkendt som det tilsvarende danske »blæst«). NO-RECOMP vælger korrekt tonelag 2 i analogi med det danske stødtab for »sand-«.

»Tømmer« og »bygning/er« indgår et lykkeligt ægteskab efter visit hos de danske slægtninge.

Evaluering

Det var med stor spænding at vi lod en norsk fonetiker læse et tilfældigt udvalg af lydskrevne NoTa-ytringer (100 ialt), i første omgang for at få en umiddelbar vurdering af lydskrivningskvaliteten. Svaret var at lydskriftens niveau svarede til det man kan forvente af en bedre fonetikstuderende på afsluttende BA-niveau. De hyppigst forekommende fejlmønstre skønnedes at være

- forveksling af [e] og [E] (som i dansk »hende«—»henne«)
- lang vokal noteret som kort og vice versa (som i dansk »mene«—»minde«)
- tonelag 1 noteret som tonelag 2 og vice versa (som i norsk »bønder« versus »bønner«).

Disse forvekslinger er angiveligt også blandt de hyppigste fejl i norske fonetikstuderendes transskriptioner. Fra et perceptionssynspunkt er fejltyperne relativt harmløse, idet de kun vedrører lokal-fonematiske forhold; de kan betegnes som finjusteringer af den enkelte vokalkvalitet og indebærer fx ikke forkert placering af hovedtryk eller udeladelse af foner. Resultatet syntes altså godt nok til at lydskrivningen var en systematisk evaluering værd.

Vi bad derefter tre erfarne norske lydskrivere, alle ansat ved Tekstlab på Oslo Universitet, om at analysere et materiale af 65 tilfældigt udvalgte NoTa-ytringer (782 tokens). Lydskriften blev præsenteret i to versioner,

I. med hvert ord markeret for ±tryk og uden anden fonetisk information
II. de fuldstændige lydskrivninger.

Deltagerne blev bedt om at revidere lydskriften i en konservativ ånd, idet instruktionen var formuleret sådan: »indfør kun ændringer hvor du mener udtalen er usandsynlig eller umulig«. De fik besked på at færdiggøre serie I før de begyndte på serie II. Revisionen blev gennemført uden aflytning, der jo ikke giver mening som evalueringsgrundlag i forhold til en automatisk genereret lydskrift.

Deltagernes vurderinger er resumeret i Tabel 5 herunder. Ord som ingen af de tre har ændret, benævnes »OK«, mens ord som højst én deltager har ændret, benævnes »acceptable«.

Filter	Lydskrevne ord godkendt af							
	3 deltagere *(OK)*		2 deltagere		1 deltager		Ingen	
alle foner	152	23,0%	217	32,8%	285	43,1%	8	1,2%
ikke {ˈˈ}	287	43,4%	139	21,0%	231	34,9%	5	0,8%
ikke {:}	308	46,5%	221	33,3%	127	19,2%	6	0,9%
ikke {ˈˈ:}	536	81,0%	64	9,7%	59	8,9%	3	0,5%
ikke {eE}	190	28,7%	223	33,7%	241	36,4%	8	1,2%
ikke {ˈˈ:eE}	599	90,5%	46	6,9%	14	2,1%	3	0,5%

Tabel 5 Tre evalueringer af NoTaFons lydskrivning. Kolonnen Filter viser hvilke foner som er taget i betragtning; i rækken »ikke {ˈˈ}« er altså opregnet antallet af lydskrevne ord der anses for korrekte når tonelaget ignoreres. Tilsvarende i de øvrige rækker, med udeladelse af hhv. vokalforlængelse, vokalerne {eE}, samt kombinationer heraf.

I serie I blev 78,0% af ordene vurderet som OK, 91,9% som acceptable. Kun 16 ord blev ændret af alle tre deltagere, hvad der igen viser at tryksætning vitterligt er vanskelig, for norsk som for dansk. I serie II blev kun 662 ud af de 782 tokens faktisk evalueret, da resten er leksikalsk og fonetisk atypiske (ufuldstændige ord, suk, tøven etc). Ud af disse blev kun 23,0% bedømt som OK og 55,8% som acceptable. Langt den største del af ændringerne var imidlertid af de førnævnte typer: tonelag (men ikke *trykplacering*), vokallængde, samt [e]/[E]-forveksling (sjældnere [0]/[e]). Hvis man ser bort fra disse som nævnt mindre alvorlige fejltyper, stiger succesprocenten markant til 90,5% ord bedømt som OK og ikke mindre end 97,4% som acceptable.

Konklusion

Det er naturligvis risikabelt at vurdere NoTaFon-lydskriften – og dermed hele det dansk-norske lydskrivningsscenarie – alene baseret på de her fremlagte tal. Som nævnt blev evalueringen udført i en konservativ ånd med færrest mulige ændringer (inden for rimelighedens grænser), og det kan have forskønnet resultatet. På den anden side mener vi at visse optimistiske konklusioner er sikre at drage. Selv om det konkrete antal af lydskrivningsfejl kan være vurderet for lavt, burde mængden og arten af fejl*typer* være troværdige. Det er derfor interessant at de fundne fejltyper er få, lokale og fra et perceptionssynspunkt ret harmløse. Med andre ord, givet at vi kan finde effektive kompenserende løsninger til et lille antal specifikke residualproblemer, har vi opnået en lydskrivning af tilstrækkelig kvalitet til at opfylde projektets praktiske hovedformål: fonetisk søgbarhed.

Vi har i løbet af 2009 gennemført en række pilotforsøg som sandsynliggør at op imod to tredjedele af de resterende lydskrivningsfejl kan rettes med enkle midler. Nærmere bestemt skønner vi at 55-65% af de regulære lydskrivningsfejl i den nuværende NoTa-lydskrift kunne korrigeres med en indsats på et par mandmåneder.

Der er ingen apriorisk grund til at tro at sproggrænsen mellem dansk og svensk skulle volde større principielle vanskeligheder end den dansk-norske. Svensk tryktilskrivning, tonemsystem, fonvalg, PoS-kategorier og syntaks lægger sig tæt op ad de dansk-norske. En mere markant forskel finder man i den fonetisk-prosodiske *realisering*, men dette vil næppe være et problem i et lydskrivningsprojekt som jo vedrører fonvalg og ikke

konkret udtale. De største udfordringer kan nok ventes i den svenske morfologi (med fx fuldvokaler, hvor norsk og dansk har schwa og [ɔR]), samt i den større leksikalske afstand; men der findes allerede en mængde praktisk orienteret litteratur om leksikalsk-morfologiske ligheder og forskelle som et svensk projekt kan tage udgangspunkt i (fx Fjeldstad et al 1989, Ambrosius 2000, Henrichsen et al 2005). Vores håb og anbefaling er derfor at det nuværende pilotforsøg bliver opgraderet til et egentlig produktionssystem til lydskrivning af de skandinaviske hovedsprog og vigtigste dialekter. Pålidelige transfer-funktioner mellem de skandinaviske sprog er attraktive både som instrumenter for grundforskningen og som vitale komponenter i oversættelsessystemer og taleteknologi.

Tak til Janne Bondi Johannessen, Kristin Hagen, Torbjørn Nordgård og Frans Gregersen for udlån af nødvendige materialer – og især for deres interesse for lydskrivningsprojektet både før, under og efter. Tak også til den anonyme reviewer af denne artikels første version for de præcise og kvalificerede kommentarer. Til slut en venlig tanke til de studerende ved et par semestres kurser i natursprogsbehandling og taleteknologi på CBS for deres engagerede deltagelse i analyserne af de dansk-norske trykforhold. Adskillige studerende opnåede, til egen forundring, en erkendelse af at selv så nørdet en aktivitet som transfer af fonetisk transskription mellem dansk og norsk kan føre til dybe indsigter i sproghistorie, lingvistisk perception, oversættelsesprocesser og interkulturel kommunikation.

Den lydskrevne version af NoTa udlånes af forfatteren ved personlig henvendelse, forudsat en skriftlig, personlig godkendelse fra Oslo Tekstlab. Kontaktinformation findes i http://www.tekstlab.uio.no/nota

Litteratur

Ambrosius, J. (2000) *Dansk-Svensk-Dansk Ordlista*; Malmö: Corona Förlag

Bondi Johannessen, J.; K. Hagen (eds) (2008) *Språk i Oslo. Ny Forskning omkring Talespråk;* Oslo: Novus forlag. Se også http://www.tekstlab.uio.no/nota/oslo/

Fjeldstad, A.; K. Hervold (1989) *Norsk for Svensker*; Lund: Studentlitteratur

Gregersen, F. (2007) *The LANCHART Corpus of spoken Danish, a corpus in progress*; in Toivanen et al (eds.) (2007)

Grønnum, N. (1998) *Fonetik og Fonologi – Almen og Dansk*; København: Akademisk Forlag

Henrichsen, P. J. (2001) *Transformation Based Learning of Danish Stress Assignment*; Eurospeech-2001

Henrichsen, P.J. (2004) *The Twisted Tongue, Tools for Teaching Danish Pronunciation Using a Synthetic Voice*; in Henrichsen (ed) (2004)

Henrichsen, P.J. (ed.) (2004) *CALL for the Nordic Languages – tools and methods for Computer Assisted Language Learning*; Copenhagen Studies in Language 30/04

Henrichsen, P.J. (2007a) *The Danish PAROLE corpus – a merge of speech and writing*; in Toivanen, J. et al (eds.) 2007

Henrichsen, P.J. (2007b) *A Norwegian letter-to-sound engine with Danish as a catalyst;* NODALIDA-2007, Tartuu

Henrichsen, P.J.; J. Allwood (2005) *Swedish and Danish, Spoken and Written Language – a statistical comparison*; International Journal of Corpus Linguistics. 10:3/2005, 367-399

Jensen, J.N.; O. Ravnholt; J. Schack (2006) *Ordet Fanger – Festskrift til Pia Jarvad i anledning af 60-års-dagen*; Dansk Sprognævns Skrifter 37

Keson, B. (1999) *Vejledning til det danske morfosyntaktisk taggede PAROLE-korpus*; Det Danske Sprog- og Litteraturselskab (se http://korpus.dsl.dk/e-resurser/parole-korpus.php)

Mørch, I.E. (2006) *Bestemhed og Nøgen Form*; in Jensen et al (eds.) (2006)

Nivre, J.; L. Grønqvist (2007) *Tagging a Corpus of Spoken Swedish*; in Toivanen et al (eds) (2007) (reprint from International Journal of Corpus Linguistics)

Nordgård, T. (2000) *NorKompLeks. A Norwegian Computational Lexicon*; COMLEX-2000, Patras

Skadhauge, P.R.; P.J. Henrichsen (2005) *DanPO – a transcription-based dictionary for Danish speech technology*; NODALIDA-2005, Joensuu

Toivanen, J.; P.J.Henrichsen (eds.) (2007) *Current Trends in Research on Spoken Language in the Nordic Countries*, vol 2., Oulu Univ. Press

Noter

1 Part-of-speech, svarende til ordklasse og bøjning samt i nogle tilfælde morfo-syntaktisk funktion (fx for participier), stil (fx *formel*), brug (fx *høflig*) og aktualitet (fx *arkaisk*).

2 NoTa's PoS-annotation er genereret automatisk med en PoS-tagger trænet på Oslo Tekstlab's store tekstkorpus (Kristin Hagen, personlig kommunikation). PoS-kvaliteten er moderat (ved stikprøvekontrol), markant lavere end for Oslo-tekstkorpusset; dette er ikke overraskende givet den betydelige stilforskel mellem de to korpora. Nivre et al. (2007) har en diskussion om de teoretiske og praktiske problemer ved at anvende skriftsprogstagging på talesprog.

3 Kønsmarkeringen har næsten tabt sin indflydelse på det daglige talesprog i Oslo-området. I vores sammenhæng spiller genustrækket ikke nogen rolle, da det ingen systematisk indflydelse har på ordenes udtale.

4 Et tekstafsnit i PAROLE svarer oftest til »en helsætning afgrænset af punktum«, men kan også være kortere fragmenter (fx »Afsløret tilfældigt«), i enkelte tilfælde 1 token. PAROLE-med-tryk, hvoraf Parole65 er en lille del, er tryksat manuelt i forbindelse med projekt Dansk Syntetisk Tale af Nina Grønnum og to forskningsassistenter. Trykstærke stavelser blev præfigeret med '<'.

5 To af deltagerne har udtrykt tvivl om visse tryksætninger eller angivet flere alternativer – dog ikke for de samme tokens.

6 Torbjørn Nordgård stillede venligt sin fonetiske database NorKompLeks til rådighed for lydskrivningsprojektet, på den betingelse at der ikke citeredes fra databasen, men at denne udelukkende anvendtes til reference- og kontrolformål i forbindelse med træning af de automatiske lydskrivningskomponenter i projektets første fase. Derefter blev projektgruppens arbejdskopi slettet. NorKompLeks spillede derfor ingen rolle i den senere del af projektperioden hvor selve lydskrivningen blev produceret.