Innledende del av foredraget:
Skriftspråk og formell grammatikk:
et forsømt objekt og et omdiskutert redskap.
(MONS2017)

Helge Dyvik

Emnet for dette foredraget er, som tittelen antyder, skriftspråk som forskningsobjekt og bruken av formell, generativ grammatikk som et redskap for å studere det. Allerede formuleringen av dette emnet reiser noen spørsmål, og la meg si litt om dem først, før vi ser nærmere på det konkrete språket og den konkrete generative grammatikken.

Skriftspråk er selvfølgelig et veletablert studieobjekt innenfor filologi og lingvistikk. Nettopp det er utvilsomt en del av grunnen til den mer stemoderlige behandling skriftspråket får i den retorikken vi gjerne bruker når vi forteller hva moderne språkforskning handler om. For skriftspråket synes der å være noe vi helst vil distansere oss fra. Vi har frigjort oss fra tradisjonens og filologiens begrensning til skriftspråket, sier vi, og etablert talespråket (og tegnspråket) som det egentlige språket, det som virkelig finnes. Skriftspråket kan da bli redusert til ufullkomment transkribert tale. I noens munn blir det bare en refleks av individers internaliserte talespråkregler, supplert med og hemmet av normative regler og tillærte stilistiske forestillinger, og uten noen selvstendig status verdig en lingvists oppmerksomhet – i hvert fall ikke hvis lingvisten har generative og teoretiske tilbøyeligheter.

Spørsmålet er om dette ikke er å gå litt langt. Men det er klart at prioriteringen av talespråk og tegnspråk ikke er grunnløs. Oppsummeringen i Wikipediaartikkelen om Linguistics er jo korrekt:

«Most contemporary linguists work under the assumption that spoken data and signed data is more fundamental than written data. This is because:

- Speech appears to be universal to all human beings capable of producing and perceiving it, while there have been many cultures and speech communities that lack written communication;
- Features appear in speech which aren't always recorded in writing, including phonological rules, sound changes, and speech errors;
- All natural writing systems reflect a spoken language (or potentially a signed one) they are being used to write, with even pictographic scripts like Dongba writing Naxi homophones with the same pictogram, and text in

writing systems used for two languages changing to fit the spoken language being recorded;

- Speech evolved before human beings invented writing;
- People learnt to speak and process spoken language more easily and earlier than they did with writing.»

Men enkelte profesjonelle lingvister går noe lenger, f.eks. på spørsmål-svar-siden Linguistics Stack Exchange:

«Linguistics is the scientific study of language. A language is narrowly defined as the set of rules that "speakers" (speaking or signing) acquire when they are very, very young. Written language has no internalized rules, it reflects only the speakers' internal rules and, possibly, a speaker's regard for the prescriptive rules of style; written language has to be learned explicitly (often laboriously). "Written language" is an artifact of (some) human cultures *who already had spoken languages*. Spoken language is definitional of our *species*, and it is this kind of species-specific capacity that linguists study.» (Alexis Wellwood)

Her blir konklusjonen at skriftspråk ikke er språk – eller i beste fall redningsløst relegert til kategorien E-språk, dette heterogene, eksterne kaoset som oppstilles som motstykke til våre internaliserte I-språk gjennom Chomskys kjente dikotomi. Særlig aktuell er denne dikotomien for lingvister som arbeider med generativ grammatikk, der studieobjektet normalt identifiseres som I-språket. Bruk av generativ grammatikk for å studere skriftspråk synes derfor å kreve noen kommentarer.

Skriftspråk har åpenbart flere egenskaper som adskiller det fra talespråk. Skriftspråk tilegnes senere enn talespråket, og det har en tidsdybde som talespråket savner – en «synkron» tidsdybde, som en følge av at tekster fortsatt kan bli lest lenge etter at opphavspersonene er borte. Noenlunde stabile skriftspråk har slik en mer transtemporal eksistens enn talespråket. Dette medfører også et akkumulert ordforråd som normalt overstiger det aktive ordforådet i en persons talespråk, og et større syntaktisk og stilistisk repertoar enn man normalt finner der. Det enorme sivilisasjonsfremskritt som et utviklet skriftspråk innebærer, ligger nettopp i at det blir et repositorium for den kollektive hukommelse i et samfunn. Dette gjelder både innholdet i tekstene og skriftspråket selv, med sitt forråd av uttrykk.

Skriftspråk er avhengige av kulturinstitusjoner for å bli opprettholdt og videreført – skoler, forlag, presse osv. Dette gjør dem mer sårbare enn

talespråk. Mens talespråk klarer seg selv, kan skriftspråk i verste fall gå i oppløsning. De er avhengige av en form for «omsorg» som talespråk ikke trenger. Skriftspråkets operative normer ligger i universet av normdannende tekster, og ettersom ingen kan lese alt, og mange leser lite, kan dokumenterende ordbøker, grammatikker og lignende ressurser være viktige for å opprettholde det, i tillegg til strukturert opplæring. Jeg bør kanskje også understreke at å opprettholde det ikke innebærer å konservere det. Det er nettopp den jevne tilførselen av nye ord og uttrykksformer som fører til rikdommen i et skriftspråk – selv om en viss gjenkjennelighet i stavemåter og bøyningsformer er en forutsetning for å opprettholde dets identitet over tid.

Det er åpenbart at jeg her beskriver skriftspråket som en felles ressurs, et felles kunnskapsobjekt som den enkelte kan ha mer eller mindre partiell kunnskap om – som noe som befinner seg i et kollektiv av språkbrukere, et språksamfunn – og ikke som et individuelt I-språk. Språksamfunnet rundt et skriftspråk er også tydelig forskjellig fra språksamfunnet rundt et talespråk. Det består av skrivende og lesende mennesker spredt over et stort område både geografisk, sosialt og i tid, og det er et språksamfunn der ikke alle medlemmer er like sentrale som normbærere, eller som informanter. De som skriver eller utgir mye og blir lest av mange, har større innflytelse på den persiperte norm enn de som skriver og leses lite. Og de som leser mye, er mer innforlivet med denne normen enn de som leser lite, og derfor mer interessante som informanter. Dette hierarkiet er et uomgjengelig, om enn udemokratisk, faktum.

Likevel er det tale om en faktisk, operativ norm som styrer lesende menneskers korrekthetsoppfatninger og syn på hva som tilhører skriftspråket, og hva som ikke gjør det. Lesende mennesker har intuisjoner om hva som er mulig i skriftspråket, intuisjoner som ikke synes å være av en vesentlig annen art enn de intuisjonene de har om eget talespråk. Det er, kort sagt, ikke på noen måte selvsagt at ikke også skriftspråkskunnskap skal kunne anskues som I-språk på samme måte som talespråkskunnskap, hvis man skulle være interessert i det. For vår skriftspråkkunnskap har ikke preg av et møysommelig memorert heterogent kaos av tilfeldige normative påfunn kombinert med vag forankring i egen tale. Vi kan vite ganske presist at noe er en velformet skriftlig setning, uten å ha lest akkurat den setningen før, og selv om vi aldri ville ha brukt den muntlig. Dette er kunnskap vi er innforlivet med – internalisert kunnskap. Skriftspråket, slik vi kan det, er rikt strukturert både syntaktisk, morfologisk og leksikalsk, og minst like tilgjengelig for presis grammatisk beskrivelse som talespråket.

Det kunne innvendes at siden skriftspråket tilegnes senere, og gjennom strukturert opplæring, vil ikke den enkeltes skriftspråkkompetanse reflektere vår medfødte språktilegnelsesevne på samme måte som vårt talte førstespråk vil. Men at det skulle være en vesensforskjell her, er en deduktiv konklusjon på grunnlag av en hypotese om hvordan førstespråk, men ikke senere språk, tilegnes; det er ikke et empirisk resultat.

Uansett er det den eksterne forståelsen av skriftspråket som et felles kunnskapsobjekt jeg vil legge til grunn i det følgende – og egentlig tror jeg ikke at jeg da i praksis arbeider svært forskjellig fra mainstreams generative grammatikere, til tross for deres programmatiske henvisninger til I-språk.

Med et eksternt skriftspråkbegrep finner vi språket manifestert i tekster. Hva den korrekte grammatikken for et slikt språk er, vil da avhenge av hvordan vi har avgrenset det – altså inndelt universet av tekster i ulike språk. Som kjent kommer ikke den språklige virkelighet allerede vakkert partisjonert i diskrete språk; den jobben må vi gjøre selv. Avgrensningen blir da i stor grad avhengig av formål: hva man vil si noe om. Generaliseringer om såkalt «engelsk» kan først skje etter at vi har bestemt oss for om både britisk, amerikansk, australsk og andre slags engelsk skal være med, eller om vi begrenser oss til et subsett av disse. Generaliseringer om norsk talespråk kan på lignende måte skje først etter at vi har klargjort hvor grensene for dette går – norsk talespråk kan f.eks. defineres som klassen av nordgermanske dialekter som snakkes som nedarvet morsmål innenfor grensene til kongeriket Norge; det er jo slik adjektivet ‘norsk’ ofte brukes om talespråk. Med denne rimelige avgrensningen er det f.eks. ikke selvsagt at vi finner noe fellestrekk ved norsk talespråk som ikke også deles av svensk eller dansk. Den avgrensede klassen av varianter som utgjør ‘norsk’, er da ikke basert på indre språklige kriterier. Derfor kan man være pessimistisk overfor muligheten for å skrive én generativ grammatikk for norsk talespråk forstått slik; her er E-språk-motforestillingene berettigede.

På området norsk skriftspråk hevdes det noen ganger at bokmål og nynorsk ikke er å anse som to skriftspråk, men som to måter å skriftfeste det samme språket ‘norsk’ på. Den påstanden baserer seg da heller ikke på språklige kriterier, men på en utvendig avgrensning av ‘norsk’, først og fremst på grunnlag av statsgrenser, sammenlignbar med den nevnte avgrensningen av ‘norsk talespråk’. Språklig sett er det ikke klart mer urimelig å betrakte bokmål og nynorsk som to skriftspråk enn det er å betrakte norsk, svensk og dansk som tre.

Til tross for variasjonen innenfor hver av bokmål og nynorsk lar tekstuniverset av utgitte tekster seg rimelig entydig partisjonere i bokmålstekster og nynorsktekster, med få eller ingen grensetilfeller. Å si at de to da er samme språk, er mer et kulturpolitisk enn et lingvistisk utsagn. Ikke desto mindre kan vi i tilfellet nynorsk og bokmål langt på vei utarbeide en felles generativ grammatikk for de to på området syntaks – mens leksikon og morfologi selvsagt er noe annet.

Uansett hvordan vi avgrenser et språk, er det neppe mulig å unngå å anta intern variasjon, fonologisk, leksikalsk, morfologisk og syntaktisk. Det komplett homogene språksamfunn er som kjent en idealisering, og allerede det gjør en konsekvent I-språk-synsvinkel vanskelig. Men det behøver ikke å forhindre en innsiktsgivende generativ beskrivelse, der variasjon kan takles blant annet gjennom underspesifisering. Dette kunne oppstilles som ett av kriteriene for individuering av skriftspråk for et gitt formål: Et skriftspråk er språket i et univers av tekster der variasjonen ikke er større enn at en tilstrekkelig underspesifisert grammatisk og leksikalsk beskrivelse fortsatt er innsiktsgivende og nyttig – for det gitte formål.

En grunn til at vi ikke ønsker å definere oss bort fra variasjon, er at også variasjonen kan ha en struktur som vi ønsker å innfange i beskrivelsen. Variasjon er ikke nødvendigvis kaos; den kan ha en lingvistisk interessant struktur – og ikke bare sosiolingvistisk. Dette er også en del av vår språklige kompetanse – kunnskap om strukturer i variasjonen, og hva strukturene uttrykker – og den kompetansen kunne ikke bli innfanget ved hjelp av et monolittisk språkbegrep som idealiserer seg bort fra variasjon.

En generativ grammatikk over et skriftspråk gir da en formell analyse av de grammatiske normene som konstituerer det. Det de syntaktiske og morfologiske reglene beskriver, er setningenes *analyserbarhet* – altså i prinsippet noe potensielt. De innebærer ingen påstand om at den enkelte språkbruker analyserer i samme grad av detalj når han eller hun skriver og leser. Grammatikken er ikke en psykolingvistisk teori eller en teori om I-språk, hvis I-språk skal forstås som en komponent som direkte inngår i prosessene bak språkproduksjon og språkforståelse. Men antagelsen vil være at en språkbruker kan *bringes til* å se den mer detaljerte analysen ved behov, f.eks. ved misforståelser som skyldes flertydighet. Det er mulig å erkjenne flere egenskaper enn man hadde erkjent tidligere, ved noe man allerede kan. I praksis kan den enkelte språkbruker huske og gjenbruke mer eller mindre komplekse fraser, på linje med ord, uten å analysere dem til bunns hver gang –

eller kanskje i det hele tatt; i hvilken grad dette skjer, er psykolingvistikkens domene. Grammatikken beskriver den prinsipielle analyserbarheten i vårt felles språk.

Det kan også finnes tilfeller der analysen av en konkret setning i en viss tekst blir forskjellig avhengig av hvordan vi har avgrenset det språket som vi antar at teksten er formulert i – f.eks. genuskategorisering av substantiv i bokmål. Formen *gaten* i en tekst vil være hankjønn hvis språket omfatter alt skriftlig bokmål, der vi finner tre kjønn. Men samme form i samme tekst vil være felleskjønn hvis språket er avgrenset til tekster der hunkjønnskongruerende former som *ei*, *lita* og *mi* ikke benyttes, slik at språket bare har to kjønn. Disse alternativene er da naturligvis uavhengige av I-språket hos skribenten.

[Utelatt hoveddel av foredraget om den komputasjonelle grammatikken NorGram og trebanken NorGramBank, med eksempler på søk og analyser av data.]

Generativ grammatikk brukt på skriftspråk er faktisk under press fra to sider: både fra lingvistikken og fra feltet NLP – Natural Language Processing. Motforestillingene fra deler av lingvistikken har jeg nevnt: Skriftspråk er ikke skikkelig språk, og en generativ grammatikk over skriftspråk forteller oss i hvert fall ikke noe om strukturen i kompetansen bak språkproduksjon og språkforståelse hos mennesket. Motforestillinegn fra NLP er at regelbaserte grammatikker er overflødige og hemmende i systemer for automatisk språkforståelse, oversettelse osv. – maskinene analyserer langt raskere og bedre med andre metoder: rent statistiske, eller i det siste, dype nevrale nettverk, som faktisk gir imponerede resultater. Så tiden er kanskje kommet til å pakke sammen?

Jeg tror ikke det.

Svaret til lingvistene er, som tidligere nevnt: Nei, ganske riktig, de grammatikkene vi skriver på denne måten, er ikke teorier om strukturen i den kompetansen som inngår i tale, skriving og avkoding. Grammatikken forteller om strukturen og analyserbarheten i *produktet* av denne kompetansen, et felles kunnskapsobjekt, og ikke direkte om hvordan kunnskapen om objektet er representert hos den enkelte. Dette kan f.eks. være nyttig i oppbyggingen av språkressurser, som vi har sett.

Svaret til forskerne innenfor NLP er analogt med dette: Regelbaserte grammatikker er ikke ideelle for sanntidsprosessering av språk i maskiner heller, men det er ikke først og fremst til slik bruk de skrives. Grammatikkene beskriver strukturen i produktet, tekstenes språk, og den informasjonen kan være nyttig også i utviklingen av systemer for sanntids språkprosessering, for eksempel i produksjon av analyserte, taggete, trenings- og testdata for maskinlæringssystemene.