Å STUDERE GAMLE SPRÅK MED DATAMASKINER

foredrag på møte
7. april

av professor Dag Trygve Truslew Haug, Institutt for filosofi, idé- og kunsthistorie og klassiske språk, Universitetet i Oslo.

Foredraget mitt handler om å studere gamle språk med datamaskiner, men jeg skal begynne med å si litt om hvorfor vi studerer gamle språk i det hele tatt. Det er mange grunner til det. Først og fremst er jo språk som gresk, latin, kirkeslavisk, klassisk armensk, sanskrit, norrønt og så videre kulturbærere – tekster skrevet på disse språkene er vår viktigste kilde til kunnskap om gresk og romersk kultur og så videre. Og det er ikke så enkelt at man bare kan lære seg f.eks. gresk slik som man lærer seg et moderne språk og så sette seg ned og lese og forstå greske tekster. Når man leser en engelsk tekst, kan man i prinsippet alltid spørre en med engelsk morsmål hva noe betyr. Men vi kan ikke spørre en romer eller en gammelgreker hva en setning betyr. Og derfor er det slik at mye av den litterære, kulturelle og historiske fortolkningen av gamle tekster dreier seg om det rent språklige: hva betyr kildeteksten faktisk? Dette leder oss til vitenskapelig utforskning av disse språkene.

Hva vi vet og kan vite om gamle språk, avhenger naturligvis av hvor mye data vi har. Thesaurus Linguae Latinae er et samarbeid mellom forskjellige vitenskapsakademier – inkludert dette – som tar sikte på å dokumentere all latin fram til år 600 etter Kristus. Prosjektet starta lenge før datamaskiner var påtenkt og har samla materialet på små papirsedler, en for hvert ord, ca. ti millioner ord i alt. Thesaurus Linguae Graecae, som dekker gresk fra Homer til Byzants fall i 1453 har per i dag 103 675 506 ord. Slik sett er gresk og latin heldig stilt i forhold til en del andre gamle språk, hvor det ofte finnes mange færre tekster, i en del tilfeller bare én.

Men i sammenlikning med moderne språk er ikke 10 eller 100 millioner ord spesielt mye. Det norske Big Brother-korpuset som inneholder alt som ble sagt under den første sesongen av Big Brother på norsk TV våren 2001, inneholder litt over en halv million ord. Så den totale mengden av ord i

Thesaurus Linguae Latinae utgjør ikke mer enn 20 sesonger med Big Brother. Enda viktigere er det selvfølgelig at kildematerialet til norsk språk i prinsippet er ubegrenset og ikke bare består av faktiske ytringer, men av språkbrukernes intuisjoner om hva som er korrekt norsk. Så det er ikke til å komme fra at studiet av gamle språk, selv godt belagte språk som gresk og latin, er vanskelig, ikke minst når vi tar i betraktning at kildene våre kommer fra et stort tidsspenn, og at språket hele tida var i endring. Det begrenser hvor godt vi kan forstå språkene, og dermed hvor godt vi kan forstå tekstene som er skrevet på disse språkene.

Dette problemet blir om mulig enda større når vi studerer gamle språk av en annen grunn, nemlig fordi de er interessante kilder til lingvistisk innsikt. Språkvitenskapen som vitenskap forsøker bl.a. å karakterisere hva som er mulige og umulige språk, og også hva som er mulige og umulige språkendringer. Da er selvfølgelig utdødde språk i prinsippet likeverdige kilder som moderne språk, i den grad vi kan få sikker viten om dem. Og det er her datamaskinene kommer inn og kan hjelpe oss, som vi skal se.

Grovt sett kan vi dele språkvitenskapen inn i fire områder: fonologi (lydlære), morfologi (formlære), syntaks (setningslære) og semantikk (betydningslære). Den tradisjonelle tilnærmingen til språkvitenskapen la stor vekt på lydlære og formlære. Dette er ting som er relativt enkelt å observere i gamle språk. Nå kan det virke rart å si at det er lett å observere lydlæren til et utdødd språk, for selvsagt kan vi aldri vite helt nøyaktig hvordan et slikt språk ble uttalt. Men vi har likevel en lang rekke kilder: i mange tilfeller finnes det gode beskrivelser av hvordan lyder ble uttalt i den antikke litteraturen; i andre tilfeller finnes det lån av ord mellom forskjellige språk som vi kan bruke til å si noe om hvordan de to språkenes lydsystemer var forskjellige; ikke minst er selvfølgelig skriftsystemet en viktig kilde til lydsystemet. Dette gjelder i særlig grad språk som er skrevet i et system som fremdeles er i bruk, som f.eks. de greske og latinske alfabetene. Hvis også språket har vært i kontinuerlig bruk fram til i dag, så kan vi også benytte oss av det faktum at lydendringer – i alle fall når de sees i makroperspektiv – er systematiske og unntaksløse. Ved å bruke alle disse beviskildene kan vi med stor grad av sikkerhet rekonstruere lyd- og bøyningssystemet i mange gamle språk og til og med bakover til språk vi ikke har, som det indoeuropeiske grunnspråket.

I lang tid var rekonstruksjonen av lydsystemer og av de systematiske endringene mellom systemene språkvitenskapens hovedfokus. I tillegg var studiet av bøyningsformer, morfologi, viktig. Dette ga språk som gresk og latin med deres rike morfologi en selvsagt plass i lingvistikken, særlig fordi fokuset så ofte var på historisk utvikling.

Men fonologi og morfologi er ikke alt. Det har det selvfølgelig heller aldri vært. Det er en lang tradisjon for å studere syntaksen i gresk og latin. ”Sýntaxis” betyr ”sammensetning” på gresk, og det handler om hvordan vi kan sette ord sammen til setninger. Ordets lingvistiske betydning har vært i bruk i alle fall siden Apollonius Dyscolus, som arbeidet ved biblioteket i Alexandria i det andre århundre etter Kristus. Og klassiske filologer har alltid vært opptatt av syntaks fordi det er helt essensielt for å forstå gresk og latin: vi kan ikke forstå en setning uten å vite hvordan den er bygd opp. Så på den filologiske sida har det alltid vært gjort grundig arbeid med syntaksen, men den lingvistiske interessen for syntaks var ikke så stor før 1957.

1957 var altså året da Noam Chomsky ga ut sin første bok, *Syntactic Structures*. Jeg skal ikke komme inn på detaljer i generativ grammatikk her, men Chomskys nye innsikt var å bruke ideer fra formell språkteori som går tilbake til den norske matematikeren Axel Thue, på naturlige språk. Poenget er da at man ser på språket som en mengde (i matematisk forstand) setninger over et gitt vokabular. Syntaksen er da en formelt spesifisert mekanisme som genererer de grammatiske setningene gitt vokabularet. Her ser vi allerede mange koplinger til informatikken og teorien for formelle språk, som vi skal komme tilbake til.

Chomskys bruk av formell språkteori på naturlige språk førte til en eksplosjon i interessen for syntaks på sekstitallet og utover. Men for døde språk er det en utfordring å se på syntaksen på denne måten. Den mengden av greske og latinske setninger som vi vet er grammatisk korrekte, er endelig, selv om den er ganske stor. Så vi kunne lage en grammatikk simpelthen ved å ta denne mengden av setninger til å definere språket. Men hele forskningsprogrammet er selvfølgelig uinteressant dersom mengden av mulige setninger er endelig. Vi trenger en mekanisme som kan generere en uendelig mengde setninger. Men det innebærer at vi må ha hypoteser om at en gitt setning på latin er grammatisk selv om den ikke er belagt noe sted. Og det er enda verre, for det er den minste sak i verden å lage en formell grammatikk som f.eks. aksepterer alle kombinasjoner av ord som en grammatisk setning. Vanskeligheten ligger typisk i å sette opp grammatikken slik at den utelukker ugrammatiske setninger. Men de 100 millionene greske ord vi har, inneholder selvfølgelig ikke en liste over ugrammatiske setninger. Korpuset kan strengt tatt aldri gi oss data om hva som er umulig gresk.

Nå må jeg selvfølgelig skynde meg å si at klassiske filologer alltid har hatt sine oppfatninger om hva som er umulig gresk eller latin. En klassisk øvelse innen gresk- og latinstudiet er stilskriving, hvor elever og studenter

oversetter tekster til god latinsk eller gresk stil, mens læreren svinger pennen over resultatet. Læreren bygger korrekturen på egne intuisjoner om hva som er god latin, og hele poenget med øvelsen er å overbringe disse intuisjonene til studenten. Men selv om dette er en vanlig og høyst forsvarlig pedagogisk praksis, kan man ikke bruke slike intuisjoner som vitenskapelige data fordi de ikke er etterprøvbare. I alle fall ikke om man ikke kan være sikker på at hele forskersamfunnet deler intuisjonen, og det vil stort sett bare skje i de uinteressante tilfellene. For eksempel kan vi være enige om at et gitt verb tar et objekt i dativ, og at det er ugrammatisk å konstruere dette verbet med en akkusativ. Men moderne syntaktiske forskningsspørsmål er ofte mye mer finkornede og krever hypoteser som delte forskerintuisjoner ikke kan svare på, simpelthen fordi dataene er vanskelige eller umulige å overskue selv for dem som er svært godt inne i den latinske litteraturen. Det er her data-maskinene kommer inn. Sagt på en annen måte: når alt vi har av data, er ti eller hundre millioner ord, så trenger vi å strukturere disse dataene så godt som overhodet mulig for å få mest mulig ut av dem.

Dette var en innsikt som klassiske filologer tidlig hadde, og mange pionerer på feltet datamaskinell språkbehandling har vært klassikere. I 1946 fikk jesuitten Roberto Busa ideen om å digitalisere alle tekstene til Thomas Aquinas, samt kommentarene til Thomas og noen andre tekster. Til sammen utgjør dette materialet 10 600 085 ord, altså like mye som den klassiske latinen i Thesaurus Linguae Latinae. Datamaskinen var knapt oppfunnet, men i 1949 møtte Busa Thomas Watson, grunnleggeren av IBM, og overtalte ham til å støtte prosjektet. I 1967 var hele teksten over på hullkort, som veide flere tonn. Busas mål var å lage en konkordans, dvs. en liste over alle ord-forekomster i materialet, med litt kontekst rundt. I løpet av syttitallet ble denne utgitt, som bok, men den er nå tilgjengelig på nett. Andre tidlige pionerprosjekter var Thesaurus Linguae Graecae, som begynte å digitalisere greske tekster i 1972. Også i Norge har vi hatt pionerer: Knut Kleve fikk lagd en konkordans over verkene til den greske filosofen Filodem, som ble brukt til å begrunne hypoteser om hva som mangler i de delvis ødelagte papyrustekstene vi har fra denne filosofen.

Disse første digitaliseringsprosjektene var rene tekstbaser som kunne brukes til å generere konkordanser, gjøre fulltekstsøk osv. Dette er svært nyttig, men den rike morfologien i gresk og latin skaper også mange problemer. Hvis vi f.eks. er interessert i hvordan et bestemt ord brukes i tekstene, kan vi ikke uten videre finne dette gjennom et fulltekstsøk uten å søke på en lang rekke former. Svaret på dette er lemmatisering, altså å knytte hvert enkelt ord opp til et lemma, dvs. en oppslagsform, gjerne slik at vi får

en morfologisk analyse også. F.eks. ønsker vi å vite at *fuit* er tredje person entall perfektum indikativ av lemmaet *sum*, som betyr "å være". Det er ikke helt trivielt å få en datamaskin til å gjøre dette i et språk med rik morfologi, men siden morfologien stort sett er regelstyrt, og unntakene ikke er uendelige selv om de er mange, så lar det seg gjøre. Problemet oppstår først og fremst når man har å gjøre med former som er tvetydige. Ved hjelp av statistikk kan vi trene datamaskinen opp til å gjette hvilken analyse som sannsynligvis er den riktige i en bestemt kontekst, og vi kan få den til å gjette riktig i omlag nitti prosent av tilfellene. Det er ikke helt galt, gitt at det er 805 forskjellige analyser å velge mellom i gammelgresk, og 653 i latin. Men 90 % riktig er naturligvis ikke et godt nok utgangspunkt for lingvistisk utforskning av tekstene, så det må også manuelt arbeid til.

Et korpus som har morfologisk analyse og lemmatisering, gir oss veldig mye mer informasjon enn en rein tekstbank. F.eks. vil vi kunne søke opp alle forekomster av et bestemt begrep, uavhengig av hvilken bøyningsform det har. Eller vi kan søke opp alle forekomster av en bestemt bøyning, uavhengig av hvilket ord de forekommer i. Med litt kreative søk kan man på denne måten finne data til stor hjelp for språkforskningen. Men fremdeles er det en stor del av den lingvistiske strukturen vi utelater, nemlig alt som har å gjøre med syntaks. For å kunne søke også i denne trenger vi mer sofistikerte digitale korpus, såkalte trebanker. I en trebank forsøker man også å kode setningers syntaktiske form. Dette krever mer kompliserte datastrukturer som tradisjonelt har hatt form som trær, derav navnet trebank. Man begynte å lage slike trebanker for moderne språk på begynnelsen av nittitallet. De siste fem til ti årene har det vokst fram slike trebanker også for eldre stadier av de største europeiske språkene, som engelsk, tysk og fransk, samt for en del andre språk som islandsk, irsk og arabisk. Ved Universitetet i Oslo har vi lagd en trebank over eldre indoeuropeiske bibeloversettelser, som inkluderer tekster på gammelgresk, latin, gotisk, klassisk armensk og kirkeslavisk. Vi har også arbeidet med ikke-bibelske tekster på gresk og latin, eldre stadier av germanske og romanske språk, samt med norrøne tekster i samarbeid med Universitetet i Bergen. Resultatet er en trebank som så langt dekker 1,3 millioner ord fordelt på elleve språk, hvorav gresk, latin og norrønt utgjør mesteparten av materialet. Vi er langt fra å ha en representativ dekning av noen av språkene, men dataene åpner allerede helt nye muligheter i forskningen. Noe av det mest spennende med det opprinnelige materialet vårt er at det består av samme tekst – Det nye testamentet – på fem forskjellige språk. Det åpner for kontrastive og sammenliknende studier av disse språkene på helt nye måter som vi skal se.

Vi kan forstå en setnings syntaktiske struktur som to relasjoner mellom ordene i setningen. Den ene relasjonen er presedens: altså hvilket ord følger etter hvilket. Dette er en enkel, lineær ordning. Den andre relasjonen er dominans eller styring, altså at f.eks. et verb krever et subjekt og et objekt. Dette er en partiell ordning som i de fleste tilfeller danner en trestruktur over ordene i setningen, som vist i figur 1. I enkle tilfeller vil det være en sammenheng mellom disse to relasjonene, slik at vi kan tegne det hierarkiske treet over den lineære strukturen uten at det oppstår kryssende greiner, som vist i figur 2. Dette svarer til at de ordene som hører sammen også står sammen.

Fig. 1.

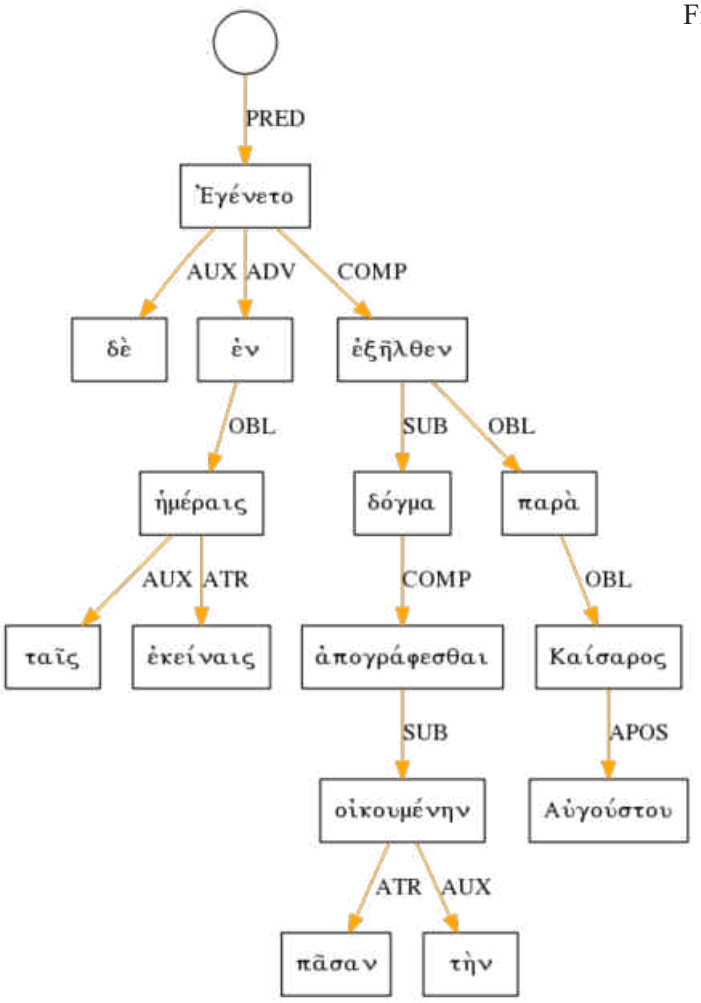
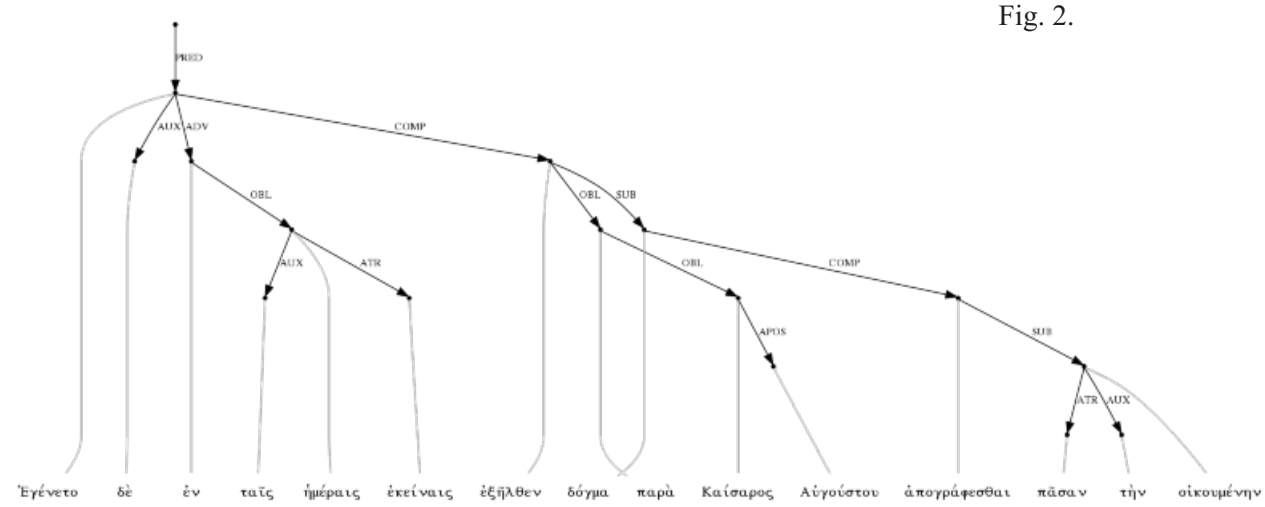


Fig. 2.



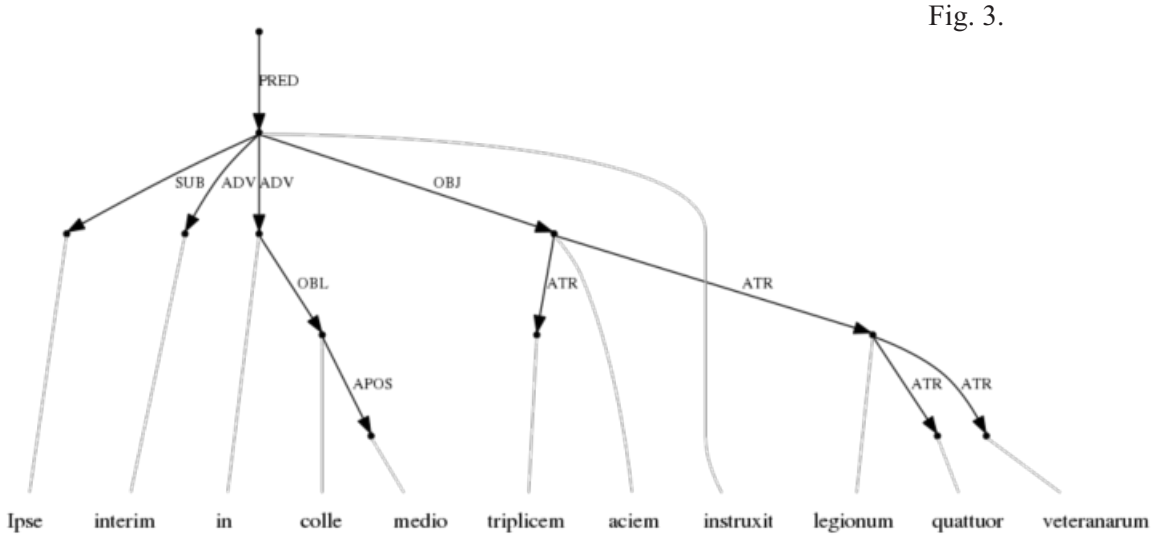
Men et av de mest interessante trekkene ved gresk og latin er at det slettes ikke alltid er slik. Forholdet mellom de to relasjonene kan være komplisert, og typisk mye mer komplisert enn det vi finner i norsk, engelsk eller noen av de vanlige europeiske språkene, som har mye strengere regler for ordstilling. Den frie ordstillingen er noe av det som gjør gresk og latin interessant for språkvitenskapen, men det er også noe av det som tradisjonell filologisk analyse har hatt vanskeligst for å håndtere. På norsk kommer vanligvis subjektet foran verbet, som kommer foran andre argumenter, slik at vi sier: ”Peter så ulven.” På andre språk vil verbet komme aller først eller aller sist i setningen. Forskjellige språk velger altså forskjellige muligheter her. Men på gresk og latin er ”alle” rekkefølgene mulig. Og faktum er at det ikke en gang er enighet om hvilken ordstilling som er den vanligste.

Nå er det ikke sikkert at det er spesielt interessant å vite hvilken ordstilling som er den vanligste i en bestemt tekst. Men likevel er det grunn til å tro at dersom vi ikke kan bli enige om slike helt grunnleggende fakta, så klarer vi ikke å komme særlig langt i forskningen. Det er derfor interessant å vite noe om hvorfor selv et såpass enkelt eksperiment med såpass små datamengder går galt. Det er selvfølgelig mange kilder til forskjeller her: én er skrive- eller regnefeil, en annen er at det kan være forskjellige utgaver som ligger til grunn for de forskjellige tabellene. (Jeg kan skyte inn at greske og latinske tekster ikke er gitt for alltid, de kommer i stadig nye utgaver som tar hensyn til framskritt i tekstforskningen.) Men sånne forskjeller vil alltid

være der og er ikke særlig interessante. Mer interessant er det at selv et såpass enkelt spørsmål kan forstås på så mange forskjellige måter. Nøyaktig hva regner vi som et objekt? Regner vi med alle slags setninger? Hvis ikke, hvilke utelukker vi? Hva gjør vi med de setningene som f.eks. har et diskontinuerlig objekt med flere underledd som befinner seg på begge sider av verbet, altså der hvor det er kryssende greiner i treet, som i figur 3? Bare de færreste studiene svarer uttømmende på slike spørsmål. Det er mulig å gjøre det – men selv da oppnår vi ikke helt det vi ønsker, for det vi faktisk vil ha, er avhengighetene mellom de forskjellige svarene og resultatene vi får for hvert enkelt svar. Altså, hva er konsekvensene dersom vi teller diskontinuerlige subjekter på den ene eller andre måten, osv. Og slike avhengigheter lar seg rett og slett ikke enkelt håndtere uten datamaskiner. Med datamaskiner og trebanker blir derimot saken mye enklere.

Dette var bakgrunnen for at vi i 2008 begynte å bygge et parallelt korpus over de tidlige indoeuropeiske bibeloversettelsene. Dette var et prosjekt – – – – – kalt PROIEL – med finansiering fra NFR, og i løpet av de fire årene prosjektet varte, var femten studentassistenter med på å bygge trebanken ved å legge inn all den informasjonen som ikke kunne ekstraheres automatisk. – – – I tillegg arbeidet en programmerer, to postdoker og to stipendiater i prosjektet. Etter prosjektperioden har vi fortsatt arbeidet i mindre skala, nå med lokal finansiering fra IFIKK ved HF, UiO. Men en

Fig. 3.

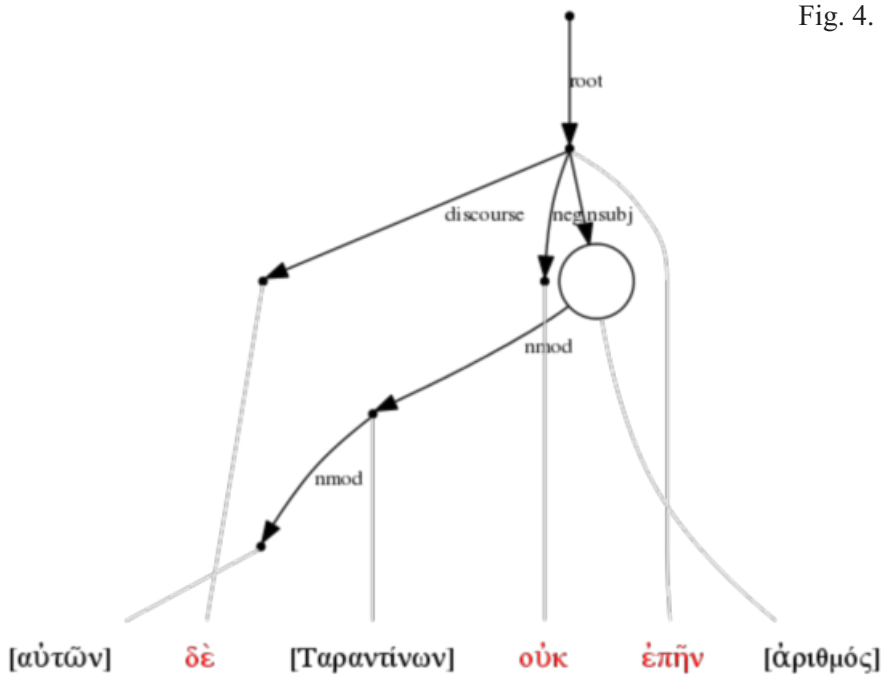


ting var å få dataene på plass, mer interessant er jo den forskningen vi har kunnet gjøre med basis i dataene. Jeg skal nå gi to eksempler på slik forskning.

Som vi nettopp har sett, så er syntaktiske diskontinuiteter en utfordring for formelle grammatikker. Årsaken er at vi må ha en mekanisme som kan dele informasjon mellom noder i det syntaktiske treet. Dette betyr at en algoritme som skal sjekke om en gitt streng aksepteres av grammatikken vår, blir mer komplisert og potensielt trenger lengre tid. I en typisk diskontinuerlig frase som i figur 3, trenger vi f.eks. å sende informasjon "over verbet" *instruxit*, siden objektet står på "begge sider". I andre tilfeller trenger vi å sende informasjon over to "hull" i strukturen, se figur 4. Subjektet i denne setningen, markert med klamme-parenteser, befinner seg på tre forskjellige steder i strukturen. Dette er komplisert å håndtere, men så lenge grammatikken vår gir oss en begrensning på hvor mange syntaktiske diskontinuiteter som i prinsippet kan være åpne til enhver tid, så er det mulig å sjekke om en streng er velformet uten at tidsbruken øker eksponensielt.

Nå vet vi at gresk og latin inneholder mange syntaktiske diskontinuiteter. Vi kan faktisk bruke trebanker til å sjekke det. *Universal dependencies* er et

Fig. 4.



internasjonalt prosjekt som tar sikte på å skape trebanker som er sammenliknbare på tvers av språk, siden de er annotert på samme måten. Hittil finnes trebanker på mer enn tredve språk, og det er mulig å sjekke hvor vanlig syntaktiske diskontinuiteter er. Vi finner at ikke-lokale avhengigheter er vanligere i gresk enn noe annet språk i universal dependencies-trebankene. Men et interessant poeng er hvorvidt det finnes en prinsipiell grense for hvor mange diskontinuerligheter en setning kan inneholde. For å se litt på det spørsmålet må vi si noe om såkalt rekursive kategorier.

Som nevnt er det et viktig poeng at grammatikken ikke legger noen begrensninger på hvor lange setninger kan være. Måten vi oppnår det på i et formelt system, er rekursjon. Et helt sentralt poeng da, er at vi kan putte én setning inn i en annen setning. F.eks. vet vi at "Per så Kari" er en grammatisk setning på norsk. Da kan vi lage en ny setning "Jon sa at Per så Kari" og enda en ny setning "Mari trodde at Jon sa at Per så Kari" osv. Vi sier at setningskategorien er rekursiv, siden vi trenger en regel som lar en setning inneholde en annen setning, og denne regelen kan da anvendes igjen i den underordnede setningen, slik at vi får en potensielt uendelig rekke setninger under hverandre. Det som da blir et interessant spørsmål, er hvorvidt rekursive kategorier – som setninger – kan være diskontinuerlige. Hvis de kan det, så vil rekursjonsmekanismen kunne sørge for at antallet diskontinuerligheter er potensielt uendelig, og det kan vi ikke håndtere uten at tidsbruken øker eksponensielt. Latin og særlig gresk er nettopp slike språk hvor man kunne tenke seg at dette skjer, siden det er såpass mange diskontinuerligheter. Og med en trebank så kan vi nå sjekke dette.

Haken er selvfølgelig at det vi ønsker å studere, er hvorvidt noe er umulig på gresk. Det kan vi strengt tatt aldri gjøre, siden det alltid kan være en tilfeldighet at noe ikke finnes i de 100 millionene ord vi har – eller enda mer sannsynlig, i de par hundre tusen ordene vi har i en trebank. Men likevel: hvis fenomenet vi ser på er vanlig nok, så kan vi ved hjelp av statistisk analyse i alle fall antyde at noe er umulig. Og i dette tilfellet ser vi på en svært vanlig kategori, nemlig setninger, så vi kan være desto sikrere på konklusjonen.

Det er noen andre problemer også. Det er f.eks. ikke helt opplagt hva som er en setning, og det er heller ikke alltid helt opplagt hvor et element hører hjemme. En setning består av et subjekt og et predikat. Men i mange tilfeller deler flere predikater ett subjekt, og da kan man lure på hvor subjektet egentlig hører hjemme. Det er f.eks. tilfelle med infinitivsetninger som *Per vil lære latin*.

Her er *Per* subjekt både for *vil* og for *lære latin*, men det er rimelig klart

at det befinner seg strukturelt i hovedsetningen, altså at det hører sammen med *vil*. Og dette har man antatt er et språklig universale, at et subjekt som er delt mellom et finitt og et infinitt verb alltid må realiseres i den finitte setningen.

Det som nå viser seg når vi ser på dataene fra trebanker, er at finitte setninger – den viktigste typen rekursive kategorier – ikke kan være diskontinuerlige. Dette er et negativt resultat, men siden vi har tusenvis av observasjoner, kan vi med statistiske tester sannsynliggjøre at fraværet av diskontinuerlige setninger ikke er tilfeldig. Dette er et interessant resultat fordi det antyder at selv for språk som gresk og latin, som er svært glade i diskontinuerlige strukturer, så finnes det begrensninger. Det har igjen implikasjoner for hvor kraftig grammatikkmekanismen må være, og gjør det mer sannsynlig at vi kan klare oss med en grammatikk som faktisk setter en øvre grense for hvor mange diskontinuerligheter som kan være åpne på en gang: når vi krysser en setningsgrense, må alle diskontinuiteter være lukket. I den grad de mekanismene vi bruker i grammatikkformalismene våre, reflekterer egenskaper ved den menneskelige språkforståelsen, sier det også noe om hvor komplekse strukturer vi kan håndtere.

Et annet interessant resultat vi får med oss på veien, er at dette kontinuerlighetsresultatet kun gjelder for en viss type infinitte setninger dersom vi antar at disse kan ha subjekter – i strid med hva man tidligere har antatt er mulig. Begge disse resultatene er altså eksempler på at utforskning av gamle språk ved hjelp av datamaskiner kan gi resultater som både er interessante for vår forståelse av de aktuelle språkene, samtidig som de har bredere implikasjoner for allmenn språktypologi og mer formelle og matematiske aspekter ved språkmodellene. Denne forskningen ville ikke vært mulig uten bruk av datamaskiner. Datamaterialet må systematiseres på en måte som rett og slett ikke lar seg gjøre med penn og papir, f.eks. fordi man må holde rede på avhengigheter mellom hvor subjektet antas å høre hjemme, og hvorvidt de resulterende strukturene er kontinuerlige.

Trebanken vår inneholder det nye testamentet på fem forskjellige språk, og jeg skal gi et eksempel på hvordan vi har kunnet utnytte dette i forskningen vår. Det første vi da må gjøre er å parallellstille tekstene våre. Her er vi i utgangspunktet heldig stilt fordi alle bibeltekster bruker det samme systemet for inndeling av tekst, basert på bøker, kapitler og vers. Vi vet da f.eks. at innholdet i Markus 6.38 er mer eller mindre det samme i alle oversettelsene. Men det er ikke nok: vi ønsker å vite ikke bare hvilke setninger som svarer til hverandre, men hvilke ord som svarer til hverandre.

Heldigvis kan datamaskinen finne ut av dette nesten på egen hånd. Den

grunnleggende tanken er at ord som er oversettelsesekvivalenter, har større sjanse til å forekomme i samme bibelvers enn ord som ikke er det. Vi kan derfor telle opp slike forekomster og gjøre statistiske tester for å rangere ord etter hvor sannsynlige oversettelsesekvivalenter de er. På denne måten kan vi lage en sannsynlighetsvekta ordbok. Når vi så kombinerer denne ordboka med annen informasjon, som f.eks. posisjon i setningen, grammatisk analyse osv., kan vi med høy grad av sikkerhet fastslå hvilke ord som svarer til hverandre. Rundt 98% av disse automatiske parallellstillingene viser seg å være korrekte når vi etterpå går gjennom dem manuelt.

Med denne parallellstillingen i hånd kan vi så kontrastere språkene i korpuset vår. En problemstilling som ofte er aktuell, er at vi forstår den greske teksten bedre enn en bestemt annen versjon av teksten – tross alt har vi hundre millioner ord på gresk og mye mindre av de andre språkene. Det vi da kan gjøre, er å se etter hvilke mønstre som svarer til et bestemt fenomen i gresken.

F.eks. er slaviske språk kjent for å ha en kategori vi kaller aspekt, som forenkla sagt uttrykker om en handling ble fullendt eller ikke, jf. forskjellen i engelsk på *He read the book* og *He was reading the book.* I moderne slaviske språk uttrykkes denne distinksjonen ofte ved at man føyer til en preposisjon foran verbet. Dette systemet finner vi også i den kirkeslaviske teksten, men samtidig finner vi et system med bøyningsformer på slutten av verbet. Spørsmålet som da oppstår, er hva som egentlig uttrykker aspekt i kirkeslavisk. Siden det meste av forskningen som er gjort på dette, har blitt gjort av morsmålsbrukere av moderne slaviske språk, er det kanskje ikke en overraskelse at de har konkludert med at det gammelslaviske systemet likner helt på det moderne systemet.

Vi har utfordret denne konklusjonen. Uten å gå i detaljer kan systematiseringen av data i trebanken vår la gresk grammatikk – som er forholdsvis godt forstått – kaste lys over kirkeslavisk grammatikk, hvor det er mye vi vet, og resultatet er nye svar på gamle spørsmål. Igjen er statistisk analyse sentralt: i dette tilfellet gjelder det umuligheten av å oversette en spesiell gresk form på en bestemt måte. Resultatet er interessant ikke bare fordi det åpner for en bedre forståelse av de slaviske dataene, men også fordi det kaster lys over hvordan kategorien aspekt har utviklet seg. Særlig interessant er det at vi kan belegge et mellomstadium hvor det er to morfologiske eksponenter som begge uttrykker aspekt.

Jeg håper disse to eksemplene viser hvordan bruk av datamaskiner åpner nye muligheter i utforskningen av døde språk, fordi vi kan systematisere og analysere dataene på helt nye måter. Det er mye snakk om digital humaniora

for tida, men jeg vil gjerne betone at det ikke dreier seg om en ny disiplin, men om nye metoder. Forskningsspørsmålene vi svarer på her, er tradisjonelle humanistiske spørsmål – i dette tilfellet fra historisk lingvistikk – men metodene er nye. Det er en utfordring at de nåværende ressursene dekker såpass liten del av de dataene vi har. Men her har vi i alle fall én fordel av at det ikke finnes mer enn ti millioner ord klassisk latin: det er realistisk at vi i løpet av denne generasjonen vil få trebanker som dekker alle tekstene.