Novus forlag · eISBN 978-82-8390-082-8

Forskjeller mellom talemål og skriftspråk

Hva kan trebanker fortelle oss?

Signe Laake og Lilja Øvrelid

Denne artikkelen presenterer en sammenlignende studie av norsk talemål og skriftspråk som tar for seg en rekke syntaktiske forskjeller mellom skrift og tale som har blitt foreslått i tidligere studier. Den nylig utviklede LIA-trebanken inneholder transkribert talespråk og er annotert for syntaks i overensstemmelse med den skriftspråklige Norsk Dependenstrebank, med visse utvidelser for å dekke talespråksspesifikke fenomener. Den syntaktiske analysen i disse trebankene tillater presise syntaktiske søk som gjør at vi kan undersøke en rekke syntaktiske fenomener, som sideordning vs. underordning, utelating av syntaktiske ledd og sammenligning av leddstilling, og sammenligne disse fenomenene i skrift og tale i norsk.

Stikkord: LIA-trebanken, Norsk Dependenstrebank (NDT), syntaktisk annotasjon, dependensgrammatikk, skriftspråk, talespråk

1 Introduksjon

Tilgjengeligheten av trebanker muliggjør lingvistiske studier ved søk i manuelt annotert syntaktisk struktur. Den nylig utviklede LIA-trebanken inneholder transkribert talespråk og er annotert i overensstemmelse med den skriftspråklige Norsk Dependenstrebank, med visse utvidelser for å dekke talespråksspesifikke fenomener. Dette muliggjør kvantitativ sammenligning basert på dependensrepresentasjoner.

I denne artikkelen presenterer vi en sammenlignende studie av norsk talemål og skriftspråk som tar for seg en rekke syntaktiske forskjeller mellom skrift og tale som har blitt diskutert i tidligere studier. Vi viser her at den syntaktiske analysen i disse trebankene tillater presise syntaktiske søk som gjør at vi kan fange opp en rekke syntaktiske fenomener, som sideordning vs. underordning, utelating av syntaktiske ledd og sammenligning av leddstilling.

Artikkelen er strukturert som følger: I seksjon 2 introduserer vi den syntaktiske formalismen som danner grunnlag for trebankene, nemlig dependensgrammatikk, og i seksjon 3 presenterer vi de to trebankene (NDT og LIA) som danner datagrunnlaget for denne studien. Seksjon 4 beskriver den empiriske studien, der vi begynner med en generell sammenligning av de to trebankene, før vi ser nærmere på spesifikke syntaktiske mønstre i disse. Avslutningsvis konkluderer vi artikkelen i seksjon 5.

2 Dependensgrammatikk og dependenstrebanker

Syntaktiske representasjoner i form av såkalte dependensstrukturer, uttrykt ved bileksikale relasjoner mellom ord, har mottatt økt oppmerksomhet innen språkteknologisk forskning i løpet av de siste ti til femten årene. Utvikling av algoritmer som effektivt kan behandle disse strukturene, utgjør en viktig grunn til at dependensgrammatikk nå er blitt den dominerende syntaktiske formalismen innen språkteknologi.

Dependensgrammatikk har en lang historie innen teoretisk lingvistikk (Tesnière 1959; Mel'čuk 1988), men kan på ingen måte hev-

des å utgjøre en enhetlig syntaktisk formalisme. Et sentralt kjennetegn ved ulike tilnærminger til dependensgrammatikk er allikevel at syntaktisk struktur ikke beskrives ved konstituentstruktur, men snarere gjennom binære relasjoner mellom ord, såkalte bileksikale relasjoner. Formelt sett er disse representasjonene *trær,* det vil si en graf der alle noder kan nås fra en *rot.* Roten er toppnoden i grafen og den eneste noden som ikke har noen innkommende kanter. Kantene i dependensgrafen er merket med *dependensrelasjoner*, der relasjonene som regel uttrykker syntaktiske funksjoner (subjekt, objekt osv.), men kan også være basert på eksempelvis semantiske roller.

2.1 Trebanker

Trebanker er tekstkorpus som er manuelt annotert med syntaktisk struktur. Slike tekstsamlinger står helt sentralt i dagens språkteknologi siden de muliggjør maskinlæringsbasert utvikling av automatiske verktøy, som ordklassetaggere og syntaktiske parsere. Trebanker kan også være svært nyttige for lingvistiske studier og lingvistisk hypotesetesting fordi de muliggjør søk direkte i manuelt oppmerket syntaktisk struktur.

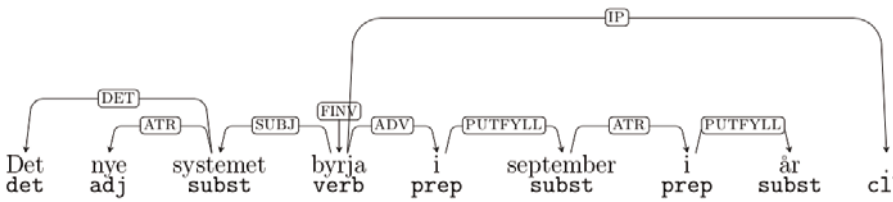
Trebanker har derfor blitt utviklet for en rekke ulike språk, og et stort antall trebanker har i løpet av senere tid blitt annotert med dependensgrammatikk. Der trebanker med konstituentbasert analyse, som den engelske Penn Treebank (Marcus et. al., 1993), lenge var dominerende, har dependenstrebanker i senere år blitt klart i flertall. En utfordring har imidlertid vært at ulike dependensgrammatiske tradisjoner har dannet grunnlag for de ulike trebanksprosjektene, og at analysene derfor har vært langt fra enhetlige. Universal Dependencies-initiativet (Nivre et al. 2016) har vært et dugnadsdrevet initiativ for å samle de ulike trebankene for ulike språk under en felles annotasjonsstandard på tvers av språk og tilgjengeliggjøre disse digitalt.

2.1.1 Norsk dependenstrebank

I 2014 ble Norsk dependenstrebank (NDT; Solberg et al. 2014) gjort tilgjengelig. Den ble utviklet ved Språkbanken (Nasjonalbiblioteket) i samarbeid med Tekstlaboratoriet, UiO og Språkteknologigruppen

ved Institutt for Informatikk, UiO. NDT inneholder totalt 600 000 ord fordelt ca. 50–50 på bokmål og nynorsk. Kildene er i hovedsak nyhetstekst (ca. 85%), men også noe tekst hentet fra offentlige rapporter og bloggdata. Tekstene ble annotert av to lingvister etter pre-prosessering med en automatisk parser (videre beskrevet i Solberg et al, 2014). I forbindelse med trebankarbeidet ble også et sett med retningslinjer utformet (Kinn et al. 2014) som gir detaljert beskrivelse av annotasjonen samt kriterier for syntaktiske valg som ble gjort under annotasjonen. Dependensgrafen i figur 1 viser en analyse fra trebanken. Vi ser at det finitte verbet *byrja* fungerer som rot i setningen og denne har en subjektsnode (SUBJ) *systemet* som dependent og har også en adverbial dependent (ADV) der preposisjonen *i* er hode.

NDT har etter at den ble utgitt, dannet grunnlaget for trening av maskinlæringsbaserte tagger og parsere for norsk (Hohle et al. 2017; Velldal et al. 2017). Den er også blitt konvertert til UD-standarden (Øvrelid og Hohle 2016) og via denne blitt brukt i flerspråklige verktøy som spaCy (https://spacy.io/) og Stanza (https://stanfordnlp.github.io/stanza/).



Figur 1: Dependensgraf fra NDT

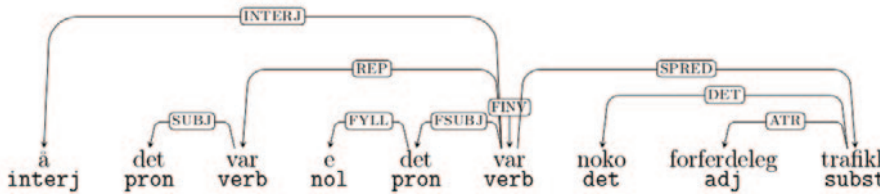
2.1.2 LIA-trebanken

I forbindelse med LIA-prosjektet har en delmengde av LIA-korpuset blitt manuelt annotert med syntaktisk analyse i form av dependensstrukturer. LIA-trebanken (Øvrelid et al. 2018) inneholder 3 651 ytringer og 48 715 løpende ord. Dataene består av transkribert tale (nynorsk), og annotasjonen følger NDTs retningslinjer med noen utvidelser for å gi en syntaktisk analyse av talespråksspesifikke feno-

mener som pauser (som # i (1)), nølelyder (som *e* i (2) og (3)) og reparasjoner (som *det var* i (2)).[1]

(1) ja # og køyrde mjølka ut i byen igjen
(2) å det var e det var noko forferdeleg trafikk
(3) så det var mykje g- e mykje greier

Figur 2 viser dependensgrafen for ytringen i (2). Der ser vi syntaktiske funksjoner som SUBJ (subjekt) og SPRED (subjektspredikat) og også talespråksspesifikke funksjoner som REP (for reparasjoner) og FYLL (for nølelyder).



Figur 2: Dependensgraf fra LIA

3 Vår studie

I studien vår sammenligner to ulike språksystemer: skrift og tale. Det er grunnleggende forskjeller mellom disse to språksystemene som vi viser kommer til syne i de syntaktiske forskjellene vi finner. De to systemene brukes i forskjellige kommunikasjonssituasjoner. Talespråk brukes typisk til kommunikasjon mellom to eller flere som er til stede på samme tid, mens skriftspråk er kommunikasjon mellom sender og mottaker som typisk er adskilt både i tid og rom. Skriftspråk er også preget av monolog i større grad enn talespråk, som gjerne er dialogisk (i hvert fall er materialet i LIA av denne art).

[1] Johannessen og Jørgensen (2006) og Rosén (2008) diskuterer hvor hensiktsmessig det er å analysere slike fenomener på lik linje som syntaktiske forbindelser. I denne artikkelen forholder vi oss likevel til den syntaktiske analysen som foreligger for NDT.

Andre grunnleggende forskjeller er at tale formidles av lydsvingninger, og talestrømmen deles opp av pauser, rytme og tonegang. Ord og lyder glir over i hverandre, og det vil være avbrudd og gjentakelser og korrigeringer. Skriftspråk er grafiske tegn, og vi har tydelige skiller som markere de ulike delene av teksten slik som markering av ord og setninger. I tillegg er det forskjell mellom de to ulike språksystemene når det gjelder planlegging og redigering. Vi har mindre mulighet til å planlegge hva vi sier, og det går heller ikke an å gå tilbake for å forandre. I skrift derimot kan en planlegge i mye større grad og også gå tilbake å redigere og gjøre forandringer underveis.

I vår studie sammenligner vi nynorsk skriftspråk slik det framstår i NDT med det talemålet som finnes i LIA. Det er noen forbehold som må tas, spesielt med tanke på talemålsmaterialet i LIA. I LIA finnes det data fra mange ulike dialekter. Dialektopptakene er også gjort på ulike tidspunkt, så korpuset inneholder derfor dialekter fra ulike tidsperioder. Vi har ikke tatt hensyn til dette i vår studie. Ulike dialekter og tidsperioder har også blitt slått sammen.

I den kvantitative delen av studien har vi gjort flere ulike sammenligninger mellom de to trebankene. Fokuset i studien vår er syntaktiske fenomener, og vi benytter oss derfor i hovedsak av den syntaktiske annotasjonen i korpusene for å utføre sammenligningen. Men siden korpusene også er annotert for ordklasse, benytter vi oss av denne informasjonen siden det ofte er relevant i undersøkelsen av syntaktiske fenomener. Innledningsvis har vi sammenlignet ulike egenskaper ved de annoterte dataene, eksempelvis fordeling av dependensrelasjoner og ordklasser i trebankene, samt setningslengde. Dette diskuteres nærmere i henholdsvis seksjon 3.1 og 3.2. Vi har deretter sett på spesifikke syntaktiske konstruksjoner og gjort søk i trebanken som fanger opp disse. For å søke i trebankene benytter vi oss av skript skrevet i programmeringsspråket Python som leser trebankene inn i en intern datastruktur som gjør det lett å navigere grafene. Merk at de fleste søkene nok også ville kunne gjennomføres i eksisterende søkegrensesnitt som eksempelvis Iness. Når det gjelder de spesifikke syntaktiske fenomenene vi har valgt å fokusere på, har vi i stor grad tatt utgangspunkt i hypoteser og observasjoner presen-

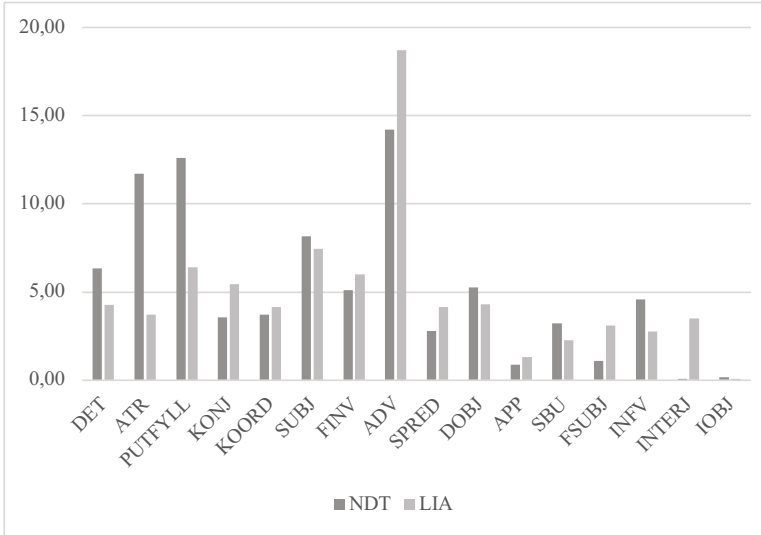
tert i tidligere forskning. Vi begrenser oss videre til nynorsk-delen av NDT fordi LIA-materialet også er transkribert til nynorsk skrift.

3.1 Oversikt: dependensrelasjoner og ordklasser

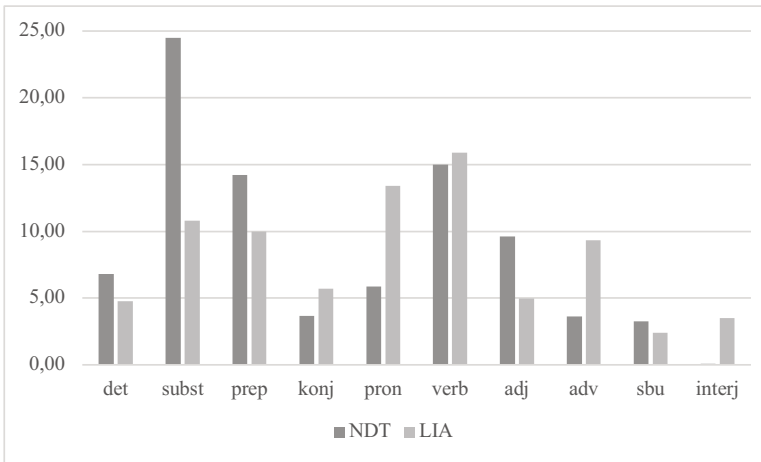
LIA-trebanken inneholder informasjon både om syntaktisk struktur, samt ordklasser, som vist i Figur 1 over. Hvert ord vil derfor være annotert med informasjon om ordklasse og dependensrelasjon, noe som gjør det enkelt å sammenligne distribusjonen av disse i NDT og LIA. Tabell 1 er en oversikt over de ulike dependensrelasjonene i korpusene. Figur 3 og 4 (neste side) viser den prosentvise fordelingen av ulike dependensrelasjoner og ordklasser i de to trebankene.

Dependens-relasjon	*Fullt navn*	*Beskrivelse*
ADV	adverbial	Adverbielle dependenter som modifiserer verb, tidvis også adjektiver, determinativer, adverb osv., både obligatoriske og ikke-obligatoriske adverbialer.
ATR	attributt	Modifiserer substantiver, ofte adjektiver men også preposisjoner.
APP	apposisjon	Appositive dependenter med substantiver og egennavn.
DET	determinativ	Artikler, men også genitiver og kvantitets-substantiver (*en del*).
DOBJ	direkte objekt	Komplementer til verb, både nominale, adjektiviske og verbale ledd (infinitiver, infinitte og finitte leddsetninger).
FINV	finitt verb	Finitte hovedverb, utgjør som regel rot i dependensgrafen.
FSUBJ	formelt subjekt	Ikke-referensielle subjekter i subjektsposisjon i presenteringskonstruksjoner, upersonlig passiv og utbrytinger.
INFV	infinitt verb	Infinitte verb som opptrer med finitt hjelpeverb, infinitiver og partisipper.
INTERJ	interjeksjon	Kan stå som rot uten finitt verb (- *Ja, innimellom*).
IOBJ	indirekte objekt	Brukes kun for nominale elementer i setninger med direkte objekt.
KONJ	konjunksjon	Brukes for konjunksjoner (*og, eller*).
KOORD	koordinasjon	Relasjon for konjunkter utover første konjunkt (som gis det koordinerte utrykkets syntaktiske funksjon).
PUTFYLL	preposisjons-utfylling	Preposisjonsobjekt, nominal dependent på preposisjon.
SBU	subjunksjon	Brukes for subjunksjoner (*at, når*).
SPRED	subjektspredikativ	Predikativelement i kopula-konstruksjoner.
SUBJ	subjekt	Referensielle subjekter i subjektsposisjon.

Tabell 1: Oversikt over dependensrelasjonene i Figur 3 med beskrivelse. For ytterligere detaljer se Kinn et al. (2014).



Figur 3: Dependensrelasjoner. Prosentvis fordeling av dependensrelasjoner over totalt antall ord i trebanken, viser de 16 mest frekvente relasjonene.



Figur 4: Ordklasser. Prosentvis fordeling av ordklasser over totalt antall ord i trebanken, viser de 10 mest frekvente ordklassene

Som det kommer frem av figur 3 og 4, er det noen dependensrelasjoner og ordklasser hvor det er tydelige forskjeller mellom NDT og LIA, mens andre relasjoner og ordklasser har en lik fordeling i NDT og LIA som for eksempel fordelingen av subjekt og finitt verbal. Videre kommer vi til å legge vekt på de dependensrelasjonene og ordklassene der forskjellen i frekvens mellom dem forhåpentligvis kan si oss noe om generelle forskjeller mellom tale og skrift.

Figur 3 viser at den mest frekvente relasjonen i begge korpusene er ADV(erbial). Men det er en tydelig forskjell i frekvens i de to korpusene: den er mye mer frekvent i LIA enn i NDT. En kan se den samme tendensen når vi ser på frekvensen av ordklassen adverb i figur 4. Adverb har en høyere frekvens i LIA enn i NDT.

En annen tydelig forskjell mellom NDT og LIA finner vi i dependensrelasjonene PUTFYLL og ATR. PUTFYLL er komplementet i en preposisjonsfrase, mens ATR brukes på beskrivende dependenter på substantiver, gjerne da adjektiv. Begge disse dependensrelasjonene er mindre frekvente i LIA enn i NDT. Siden både PUTFYLL og ATR er komplementer i fraser, kan dette tyde på at talemålet har færre komplekse fraser enn skriftspråket. Papazian (1984, 142) og Kulbrandstad og Kinn (2016, 50) har foreslått at det er enklere syntaktisk konstruksjoner i talemålet, og våre data støtter dette.

Dependensrelasjonen ATR kan sees i sammenheng med ordklassen adjektiv, hvor vi ser at det er lavere frekvens av adjektiv i LIA enn i NDT. Dette kan tyde på at det er færre utfyllende beskrivelser i talemål. Det kan være en effekt av kommunikasjonssituasjonen der en ikke trenger å beskrive noe grundig hvis en får bekreftelse at samtalepartneren vet hva en snakker om.

I NDT har dependensrelasjonen INTERJ(eksjon) en svært lav frekvens (den relative frekvensen for interjeksjoner er 0,07 %), mens det er en hel del av disse i LIA (den relative frekvensen i LIA er 3,5 %). Dette er heller ikke overraskende siden talemålssituasjonen innbyr til bruk av interjeksjoner i større grad enn i skriftspråksituasjoner. Papazian (1984, 184) påpeker at tale gjerne er mer livlig, impulsiv og emosjonell enn skrift og derfor mer preget av interjeksjoner, slik våre data også tyder på.

Noe overraskende er det kanskje at figur 3 viser at det er flere forekomster av SUBJ(ekter) enn av FINV (finitt verb). En skulle tro at forekomsten av disse to skulle være tilnærmet like siden norsk har et subjektskrav (se seksjon 3.4 for en diskusjon om utelatelse av setningsledd). Men forholdet mellom antall SUBJ og FINV er primært knyttet til de valgene som har blitt tatt i hvordan korpusene blir annotert. Det finitte verbet får relasjonen FINV i helsetninger, mens det finitte verbet i finitte leddsetninger angir relasjonen leddsetningen har til verbet i oversetningen, som for eksempel DOBJ eller ADV. Subjekter derimot blir annoterert både i leddsetninger og helsetninger som SUBJ. Dette fører til at alle subjekter i korpuset er SUBJ, mens ikke alle finitte verb er FINV.

Vi kan også finne forskjeller mellom tale og skrift ved å se på ulike ordklassefrekvenser i NDT og LIA i figur 4. Den største ordklassen, substantiv, viser en klar forskjell i de to korpusene hvor substantiv er mer frekvent i NDT (24 %) enn i LIA (10 %). Denne ordklassen burde sees i sammenheng med en annen ordklasse, nemlig pronomen. Det er mer bruk av pronomen i LIA (13 %) enn i NDT (5 %). Kulbrandstad og Kinn (2016, 50) påpeker at det er mer bruk av deiktiske uttrykk, for eksempel pronomen, i talemål enn i skriftspråk. Dette er ikke overraskende med tanke på at talesituasjonen er preget av en her og nå-situasjon i større grad enn skrivesituasjonen, og dette legger til rette for utstrakt bruk av deiktiske uttrykk i talemålet.

Søkene i korpuset gir oss også mulighet til å kombinere informasjonen vi har om dependensrelasjoner og ordklasser. Dermed kan vi få vite mer om hva slags form setningsleddene har. Vi kan blant annet finne ut hva slags type subjekter vi finner i de to ulike korpusene. Ved å se på typen subjekt kan vi undersøke den tendensen at et subjekt gjerne vil være kjent informasjon. Et typisk subjekt som har kjent informasjon vil ha formen pronomen. Som vi kan se av tabell 2, er det en sterk tendens til at subjektet er pronominalt i LIA der bare ca. 10 % av subjektene har en annen ordklasse enn pronomen. Disse funnene blir støttet av Søfteland (2013, 142), som har undersøkt ulike typer subjekt i spontantale fra fire samtaler i Nordisk dialektkorpus og fant at 83 % subjektene var pronominelle. I NDT derimot finner

vi ikke tilsvarende forhold mellom subjekt og pronomen. I skriftspråkskorpuset er det en overvekt av substantiver som subjekt.

	LIA (N=3622*)*	*NDT (N*=24714*)*
Pronomen	88,9 %	39,7 %
Substantiver	9,7 %	56,5 %
Adjektiver	0,4 %	2,0 %
Determinativer	0,4 %	1,0 %

Tabell 2: Fordelingen av subjekter

3.2 Setningslengde

Både Papazian (1984, 142) og Kulbrandstad og Kinn (2016, 50) hevder at talemålet typisk har kortere ytringer enn skriftspråket. Hverken Papazian (1984) eller Kulbrandstad og Kinn (2016) har gjort kvantitative undersøkelser av setningslengde i skriftspråket sammenlignet med talemålet. Kvantitative undersøkelser ble derimot gjort i Talemålsundersøkelsen i Oslo (TAUS). I TAUS-materialet ble setningslengden i opptak av uformelle intervjuer med folk fra Oslo fra 1971–73, undersøkt. Der fant Hanssen et al. (1978, 138) at den gjennomsnittlige setningslengden var 10 ord.

For å undersøke setningslengden i de to korpusene har vi måttet bruke ulike definisjoner av hva en setning er. I NDT har vi definert en setning som en enhet som blir markert med stor bokstav og punktum selv om dette avviker fra den grammatiske definisjonen av setning i Faarlund et al. (1997, 39) der en setning blir definert som en forbindelse av et nominal som er setningens subjekt, og en verbfrase som må inneholde et finitt verb. En konsekvens av vår definisjon av setning er at ikke alt som er markert med stor bokstav og punktum i skriftlige tekster oppfyller den grammatiske definisjonen av en setning.

I talemål er det notorisk vanskelig å definere enheten "setning". For å komme frem til en definisjon av setning har vi basert oss på transkripsjonen som er gjort av lydfilene. Her er talestrømmen delt

opp i segmenter. I transkripsjonsveiledning til LIA står det at segmentene skal være meningsfulle enheter som skal ligne mest mulig på setninger (Hagen et al. 2018, 14). Derfor har vår definisjon av en setning i LIA vært at et segment er lik en setning. Et segment vil derfor kunne inneholde avbrudd, nølelyder og reparasjoner. Videre i artikkelen bruker vi *setning* om denne enheten i talemål.

Vi finner at den gjennomsnittlige setningslengden i NDT er 17,24 ord, mens den i LIA er 13,34 ord. Med dette kan vi bekrefte antakelsen i Papazian (1984) og Kulbrandstad og Kinn (2016) om at det er kortere setninger i talemålet enn i skriftspråket. Likevel er våre resultater fra NDT og LIA ganske usikre siden vi har sammenlignet to enheter som ikke nødvendigvis er like siden vi bruker forskjellige definisjoner på setning i de to korpusene.

3.3 Sideordning og underordning

Papazian (1984, 142) hevder at det er mer lineær sideordning i talemålet sammenlignet med skriftspråket, og at det er mer underordning i skriftspråket enn i talemålet blant annet ved mer utstrakt bruk av leddsetninger. I følge Papazian fører mer sideordning i talespråket til blant annet en enklere syntaks og at denne enklere syntaksen også henger sammen med de kortere setningene som vi allerede har diskutert i forrige del.

For å undersøke sideordning og underordning på setningsnivå sammenligner vi frekvensen av dependensrelasjonene KONJ(unksjon) og SBU(subjunksjon). Forventningen om mer bruk av leddsetninger i skriftspråket ble bekreftet ved at vi finner en høyere frekvens av relasjonen SBU i NDT enn i LIA. Tilsvarende blir antakelsen om at det er mer lineær sideordning i skrift enn i tale, bekreftet siden vi finner at det er høyere frekvens av relasjonen KONJ i LIA enn i NDT.

3.4 Utelating av syntaktiske ledd

Utelating av syntaktiske ledd som for eksempel subjekt og objekt er typisk for talemål (Papazian 1984, 143; Stjernholm 2008; Nygård 2018). Som eksemplene under viser, kan en i norsk talemål utelate

ulike syntaktiske ledd. I (4) har et referensielt subjekt blitt utelatt, mens i (5) har et formelt subjekt blitt utelatt.

(4) Gjekk ein tur til noko trytetjørner som var litt ifrå for å fiske (LIA)
(5) Var langt t- å gå av garde i kolmørke om morgonen da (LIA)

For å undersøke dette begrenser vi oss til å undersøke setninger der subjektet er utelatt. I norsk er dette sammen med det finitte verbalet det eneste andre obligatoriske setningsleddet (med visse forbehold). Korpuset er ikke tagget for valens eller lignende. Derfor er det ingen mulighet til å for eksempel søke etter setning som skulle ha hatt et objekt, men hvor objektet er utelatt.

Ved å søke etter mønster der roten er et finitt verb uten subjektsdependent og verbet ikke er i imperativ, fant vi at det var forskjell mellom de to korpusene. I NDT var frekvensen av setninger uten et subjekt 1,9 %, mens det i LIA er 6,7 %. Tidligere studier slik som Stjernholm (2008) og Nygård (2018) har ikke undersøkt frekvensen av utelatte subjekt i talemål selv om de begge kommenterer at dette er vanlig. En kunne kanskje ha forventet en enda høyere frekvens av utelatte subjekt i LIA. Uansett støtter funnene våre tidligere forskning som viser at utelatelse av subjekt er mer typisk for talemål enn for skriftspråk.

Til å være et skriftspråkskorpus er det overraskende at NDT har såpass mange setninger uten subjekt, siden en setning i skriftspråket blir definert som en forbindelse av et nominal som er setningens subjekt og en verbfrase som må inneholde et finitt verb (Faarlund et al. 1997, 39). Men her spiller tekstutvalget en rolle. Siden den største delen av tekstene i NDT er avistekster, vil avisoverskrifter stå for en del av setningene uten subjekt. Eksempel (6) viser dette:

(6) Veks forbi Danmark (NDT)

I tillegg inneholder NDT bloggtekster. Disse tekstene er gjerne mer uformelle og talemålspreget, og vi finner setninger uten subjekter her slik som i (7). I tillegg kan vi ikke utelukke setninger av typen som

vises i (8), som kanskje heller viser bruk av stor bokstav og punktum selv om det ikke er en helsetning enn muligheter for utelatelse av subjekt i skriftspråket.

(7) Men ok, skal ikkje gå i detalj på dette her no. (NDT)
(8) Og vart tekne imot så hjarteleg at det sit i kroppen enno. (NDT)

3.5 Subjektdublering

Det er ikke bare mulig i norsk å utelate visse setningsledd, det er også mulig med dublering av setningsledd. Som oftest er det subjektet som er dublert, men det kan også være andre ledd. Det er vanligvis slik at siste leddet (gjerne kalt kopien eller det høyredislokerte leddet) er et pronomen slik som eksempel (9) viser.

(9) **Han** hadde sjølve livet i hendene **han** (LIA)

Borthen (2018, 404) hevder at pronominal høyredislokering (kopien er et pronomen), er et vanlig fenomen i muntlig tale i norsk, men hun har selv bare eksempler fra skriftspråk. Undersøkelsen av LIA og NDT kan bekrefte Borthens antakelse. Ved å søke etter mønster der SUBJ(jektet) har en pronominell dependent med relasjonen APP(osisjon), fant vi at frekvensen av subjektdublering av typen pronominal høyredislokering i LIA i alle setninger er 2,7 % (N=102), mens i NDT er frekvensen 0,1 % (N=35).

(10) Men **eg** var jo glad til **eg** for da fekk eg reise ein tur til Gjøvik (LIA)
(11) **Det** var mykje pengar **det** (LIA)
(12) Men **vi** var nøgde med den førre finansministeren, **vi** (NDT)
(13) For å forklåre motivasjonen for arbeidet, peikar han mellom anna på det at **den gamle typen bibellesarar** er borte; **dei som drøfta tolkingar, usemjer og uklåre tekststader nøye**. (NDT)

Hvis vi sammenligner eksemplene fra LIA og NDT, ser vi at de to eksemplene fra LIA følger den prototypiske subjektdubleringen som Borthen (2018) undersøker. Slike eksempler finnes også i NDT slik som (12) viser, men det er typen som vi ser i (13) som er den vanligste i NDT, altså hvor det høyredislokerte leddet er en spesifisering av subjektet mer enn en dublering eller kopiering. Det er nok en annen type dublering enn den som er frekvent i talemålet.

3.6 Indirekte objekt

I norsk har vi mulighet for å uttrykke de semantiske rollene benefaktiv og mottaker (blant annet) på to ulike måter enten med et nominalt indirekte objekt som i (14) eller som en preposisjonsfrase som i (15):

(14) Hun ga **hesten** en gulrot.
(15) Hun ga en gulrot **til hesten.**

Allerede Western (1921, 140–141) påpekte at det er en forskjell mellom talespråk og skriftspråk i dette henseende: Nominalt indirekte objekt finnes ofte i «det litterære sprog hvor talesproget foretrekker et komplement». Med komplement mener Western her en preposisjonsfrase slik som i (15).

Ved å se på den relative frekvensen for dependensrelasjonen indirekte objekt finner vi at NDT har en frekvens på 1,7 % (*N*=529), mens LIA har en frekvens på 0,6 % (*N*=34). Det er viktig å påpeke at i annotasjonen av korpusene har bare nominale ledd blitt annotert som indirekte objekt. Preposisjonsfraser slik som i (15) får dependensrelasjonen ADV(erbial).

Faarlund et al. (1997, 727) påpeker at valget mellom nominalt indirekte objekt eller preposisjonsfrase er styrt av pragmatiske og referensielle forhold. Et nominalt indirekte objekt vil gjerne ha unik referanse og representere kjent informasjon. Dette kan vi operasjonalisere for å enkelt undersøke om dette stemmer i NDT og LIA. Vi tar utgangspunkt i at pronomen er både kjent informasjon og også har unik referanse, mens substantiv ikke nødvendigvis vil være kjent informasjon i like stor grad. I LIA finner vi at nesten alle nominale

indirekte objekt er pronomen, (94,1 %), mens resten er substantiver. Det må påpekes at det er få nominale indirekte objekt i LIA, bare 35 tilfeller. I NDT derimot har vi langt flere nominale indirekte objekt (*N*=529), og det ser ikke ut som tendensen til at nominale indirekte objekt typisk er pronomen her er like sterk som i talespråkskorpuset. 57,8 % av de nominale indirekte objektene er pronomen, mens 38 % er substantiv. Forskjellen mellom LIA og NDT i frekvensen av pronomen og substantiver som vi finner i de nominale indirekte objektene, kan sannsynligvis knyttes til de generelle kjennetrekkene ved talemål, som gjør at vi også typisk finner flere pronominelle subjekt i LIA enn i NDT, som vi diskuterte i 3.1.

3.7 Rekkefølgen finitt verb–negasjon i leddsetninger

I norsk kan man finne variasjon i rekkefølge på finitt verb og negasjon i leddsetninger. Eksemplet i (16) viser den vanlige rekkefølgen hvor negasjonen *ikke* kommer før det finitte verbet, mens eksempelet i (17) viser motsatt rekkefølge: Negasjonen kommer etter det finitte verbet.

(16) Da mente han [at han ikke kunne være gift] (Neg-V)
(17) Da mente han [at han kunne ikke være gift] (V-Neg)

Denne variasjonen i rekkefølgen på verb og negasjon i leddsetninger har blitt grundig diskutert blant annet i Julien (2007, 2009) og Wiklund et al. (2009).

For å undersøke om det er noen forskjell mellom disse to leddstillingsmønstrene i skrift og tale søkte vi etter mønstre i trebankene der finitt verb ikke er en rotnode, men likevel har både en subjunksjon og *ikke* som dependenter.

Tabell 3 viser at rekkefølgen hvor *ikke* kommer før det finitte verbet, er helt klart mest frekvent i både NDT og LIA. Dette er ikke overraskende fordi dette er default ordstilling i leddsetninger, mens rekkefølgen hvor negasjonen kommer etter det finitte verbet, har visse restriksjoner. Som vi kan se i tabellen, forekommer denne rekkefølgen nesten ikke i NDT. Overraskende nok har en fjerdedel av

leddsetningene ordstillingen der *ikke* kommer etter det finitte verbet i LIA. Dessverre er ikke LIA-korpuset stort nok til å finne veldig mange eksempler, men vi konkluderer likevel med at det er en klar tendens til at talemålet har flere tilfeller av rekkefølgen V-neg enn i skrift. Våre resultater støttes av funnene i Ringstad (2019). Hun har undersøkt andre talemålskorpus som Scandiasyn, Nota og Big Brother og finner den samme fordelingen mellom rekkefølgen Neg-V (66 %) og V-neg (34 %) som i LIA.

	NDT (N=457)	*LIA (N=43)*
Neg-V	98,9 %	72,1 %
V-neg	1,1 %	27,9 %

Tabell 3: Rekkefølgen av det finitte verbet og ikke i leddsetninger

3.8 Så-*konstruksjon*

I norsk kan en ha setninger som (18):

(18) Da vi hadde gått i fleire timar, **så** kom vi fram til ei hytte (Faarlund et al. 1997, 817)

Faarlund et al. (1997, 817) beskriver denne konstruksjonen med at *så* står i forfeltet etter et ledd (*Da vi hadde gått i fleire timar*) i ekstraposisjon i et løst forfelt. Det vanligste er at det ekstraponerete leddet er alle typer fritt adverbial, ikke bare tidsadverbial slik som i eksempelet. Faarlund et al. (1997, 817) påpeker også at *så* i skrift er sjelden når adverbialet foran bare er en frase, og er mer vanlig i bruk i skrift når det er etter en leddsetning. Salvesen (2017) møtte mye motstand hos informantene som skulle vurdere skriftlige setninger med denne bruken av *så*. Informantene kommenterte ofte at setningene var bedre uten *så* og at setningen ble barnslige med *så.* Dette kan tyde på at en vurderer konstruksjonen som noe som ikke hører hjemme i skrift, men kanskje heller tilhører talemålet.

For å finne denne konstruksjonen i korpusene søkte vi etter mønster der *så* er dependent på rotverbet og har en apposisjons-dependent. Vi finner at av alle setninger i LIA har 2,5 % av dem *så*-konstruksjo-

nen, mens bare 0,2 % av alle setningene i NDT har denne konstruksjonen.

Våre resultater støtter påstanden om at *så*-konstruksjonen forbindes med talemålet, og at den er mindre frekvent i skrift enn i tale. Vi finner at *så* kan stå etter flere ulike typer ekstraponerte ledd i LIA enn det gjør i NDT. I eksemplene (19)–(22) er det ekstraponerte leddet en adverbiell leddsetning, en adverbiell preposisjonsfrase og et enkelt adverb. Likevel fant vi *så* etter fraser i NDT som Faarlund et al. (1997, 817) påpeker er sjeldent i skrift slik som i (22) der det ekstraponerte leddet er en preposisjonsfrase.

(19) Når vi hadde fått krøttera oppover til til ved Auenhaugen **så** kunne vi leike på Auenhagen nesten heile tida (LIA)
(20) I rettig gammal tid **så** var det nå byggmjøl (LIA)
(21) Seinare **så** trudde vel kanskje ikkje så mykje på det. (LIA)
(22) For oss sunnmøringar **så** liknar det mest ein formastelse å servere ball som noko søtt (NDT)

4 Konklusjon

Denne studien har vi presentert en sammenlignende studie av et utvalg av syntaktiske fenomener i talemål og skriftspråk. Vi har vist at søk i syntaktiske strukturer lar oss gjøre denne studien på en presis måte gjennom kvantitative sammenligninger. Den syntaktiske annotasjonen kan fortelle oss en god del om den grammatiske strukturen i språket vårt, og dette gir gode muligheter for videre studier av både skriftspråket og talemålet. Annotasjonen av korpus gir ikke bare muligheter til å gjøre kvantitative studier slik som denne, men en kan også gjøre mer kvalitative studier ved at det gjennom annotasjonen er mulig å finne sjeldne konstruksjoner i språket. Denne typen annotasjon lar oss også bekrefte eller avkrefte tidligere antakelser som ofte ikke var basert på faktisk språkbruk.

Det er visse utfordringer ved å bruke syntaktisk annotasjon. Det er foreløpig en teknisk barriere som det kan være vanskelig å komme over. Et ønske for fremtiden vil være at det kommer på plass en tek-

nisk løsning som gjør det mer brukervennlig å gjøre søk i de syntaktiske annotasjonene. Dessuten er også materialet foreløpig ganske lite. Likevel er de to trebankene, og spesielt LIA-trebanken med det unike dialektmaterialet, en verdifull ressurs for forskningen på norsk språk.

Referanser

Borthen, Kaja. 2018. Pronominal høyredislokering i norsk, det er et interessant fenomen, det. *Norsk lingvistisk tidsskrift* 36, 403–450.

Faarlund, Jan Terje, Svein Lie og Kjell Ivar Vannebo. 1997. *Norsk referansegrammatikk.* Oslo: Universitetsforlaget.

Hagen, Kristin, Live Håberg, Eirik Olsen og Åshild Søfteland. 2018. Transkripsjonsrettleiing for LIA. http://www.tekstlab.uio.no/LIA/pdf/transkripsjonsrettleiing_lia.pdf.

Hanssen, Eskil, Thomas Hoel, Ernst Håkon Jahr, Olaug Rekdal og Geirr Wiggen. 1978. *Oslomål. Prosjektbeskrivelse og syntaktisk analyse av oslomål med henblikk på sosiale skilnader*. Oslo: Novus.

Johannessen, Janne Bondi og Fredrik Jørgensen. 2006. Annotating and parsing spoken language. I *Treebanking for Discourse and Speech: Proceedings from NODALIDA 2005 Special Session on Treebanks for Spoken language and Discourse,* redigert av Peter Juel Henrichsen og Peter Rossen Skadhauge, 83–104. København: Samfundslitteratur.

Julien, Marit. 2007. Embedded V2 in Norwegian and Swedish. *Working Papers in Scandinavian Syntax* 80, 103–161.

Julien, Marit. 2009. The force of the argument. *Working Papers in Scandinavian Syntax* 84, 225–232.

Kinn, Kari, Per Erik Solberg og Pål Kristian Eriksen. 2014. *NDT Guidelines for Morphological Annotation*. Oslo: Språkbanken, Nasjonalbiblioteket.

Kulbrandstad, Lars Anders og Torodd. Kinn. 2016. *Språkets mønstre*. Oslo: Universitetsforlaget.

Marcus, Mitchell, Beatrice Santorini og Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics* 19 (2), 313–330

Mel'čuk, Igor Aleksandrovič. 1988. *Dependency syntax: theory and practice.* Albany, NY: SUNY press.

Nivre, Joakim, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Yoav, Jan Hajič, Jan, Christopher Manning, Ryan McDonald, Ryan, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty og Daniel Zeman.. 2016. Universal dependencies v1: A multilingual treebank collection. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* Portoroz, Slovenia: European Language Resources Association (ELRA).

Nygård, Mari. 2018. *Norwegian Discourse Ellipsis. Clausal architecture and licensing conditions.* Amsterdam: John Benjamins.

Papazian, Eric. 1984. *Talemål og skriftspråk. Forskjellen og sam svaret.* Skriftserie nr. 16, serie B. Bergen*:* Institutt for fonetikk og lingvistikk, Universitetet i Bergen.

Salvesen, Christine Meklenborg. 2017. I norsk så bruker vi så. Foredrag, MONS 17, Bergen

Solberg, Per Erik, Arne Skjærholt, Lilja Øvrelid, Kristin Hagen og Janne Bondi Johannessen. 2014. The Norwegian Dependency Treebank. *Proceedings the Ninth International Conference on Language Resources and Evaluation (LREC'14),* Reykjavik: European Language Resources Association (ELRA).

Stjernholm, Karine. 2008. Subjektsellipser: fins pro i norsk talespråk? Masteroppgave, Universitetet i Oslo.

Ringstad, Tina Louise. 2019. Distribution and function of embedded V–Neg in Norwegian: A corpus study. *Nordic Journal of Linguistics* 42 (3), 1–35.

Rosén, Victoria. 2008. Mot en trebank for talespråk. I *Språk i Oslo. Ny forskning omkring talespråk*, redigert av Janne Bondi Johannessen og Kristin Hagen, 214–225. Oslo: Novus

Teleman, Ulf. 1974. *Manual för grammatisk beskrivning av talad och skriven svenska.* Lundastudier i språkvetenskap. Lund: Studentlitteratur.

Tesnière, Lucien. 1965. *Éléments de syntaxe structurale.* Paris: Librairie C. Klincksieck.

Velldal, Erik, Lilja Øvrelid, og Petter Hohle. 2017. Joint UD parsing of Norwegian bokmål and nynorsk. *Proceedings of the Nordic Conference on Computational Linguistics (NoDaLiDa'17).* Linköping: Linköping University Electronic Press.

Western, August. 1921. *Norsk riksmåls-grammatikk for studerende og lærere.* Kristiania: H. Aschehoug & Co.

Wiklund, Anna-Lena, Kristine Bentzen, Gunnar Hrafn Hrafnbjargarson og Thorbjörg Hróarsdóttir. 2009. On the distribution and illocution of V2 in Scandinavian that-clauses. *Lingua* 119 (12), 1914– 1938.

Øvrelid, Lilja og Petter Hohle. 2016. Universal dependencies for Norwegian. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16).* Portoroz, Slovenia: European Language Resources Association (ELRA).

Øvrelid, Lilja, Andre Kåsen, Kristin Hagen, Anders Nøklestad, Per Erik Solberg og Janne Bondi Johannessen. 2018. The LIA treebank of spoken Norwegian dialects. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC'18)* Miyazaki, Japan: European Language Resources Association (ELRA).

English summary

This article presents a comparative study of Norwegian spoken and written language that addresses a number of syntactic differences between written and spoken language that have been proposed in previous studies. The newly developed LIA treebank contains transcribed spoken language and is annotated in accordance with the written language Norwegian Dependency treebank, with certain additions to cover spoken language-specific phenomena. The syntactic analysis in these two treebanks facilitate precise syntactic searches that allow us to investigate a number of syntactic phenomena, such as co-ordination vs. subordination, omission of constituents and word

order, and compare these phenomena in writing and speaking in Norwegian.

Signe Laake
Institutt for grunnskole- og faglærerutdanning
Oslomet – storbyuniversitetet
signe.laake@oslomet.no

Lilja Øvrelid
Institutt for informatikk
Universitetet i Oslo
liljao@ifi.uio.no