Novus forlag · eISSN 2387-6670

NORINT-korpuset – et elektronisk innlærerkorpus til bruk i andrespråks-forskning

Annely Tomson, Oliwia Szymańska, Kristin Hagen
Universitetet i Oslo

Sammendrag

NORINT[1]*-korpuset* (Universitetet i Oslo, 2020) er et forholdsvis nytt innlærerkorpus utviklet ved Institutt for lingvistiske og nordiske studier (ILN) ved Universitetet i Oslo (UiO). NORINT-korpuset inneholder muntlig og skriftlig norsk innlærerspråk av voksne internasjonale studenter med norskferdigheter på eller over nivå B1 i henhold til det felles europeiske rammeverket (*Common European Framework of Reference for Languages* (CEFR)) (Council of Europe, 2001). I denne artikkelen beskriver vi datamaterialet i korpuset, hvordan det er transkribert og annotert, og hvordan man kan anvende det brukervennlige søkeprogrammet det er lagt inn i. I tillegg viser vi hvordan man kan bruke mulighetene i NORINT-korpuset i forskning. Avslutningsvis sammenlikner vi NORINT-korpuset med *ASK – Norsk andrespråks-korpus* (ASK) (Universitetet i Bergen, 2020) for å diskutere muligheter og begrensninger i NORINT-korpuset.

Nøkkelord: *innlærerkorpus, talespråkskorpus, skriftspråkskorpus, andrespråkskorpus*

[1] NORINT er en forkortelse for norsk for internasjonale studenter, brukt om norskkurstilbudet for internasjonale studenter ved Institutt for lingvistiske og nordiske studier ved Universitetet i Oslo.

Innledning

Siden engelsk er det språket som har flest innlærere i verden, har mesteparten av forskningen på andrespråkstilegnelse fokusert på innlærere som har engelsk som andrespråk (L2) (Neri, Cucchiarini & Strik, 2006, s. 358). Elektroniske innlærerkorpus, dvs. elektroniske samlinger av naturlige eller nesten-naturlige[2] språkdata produsert av fremmedspråks- eller andrespråksinnlærere og samlet inn i samsvar med eksplisitte designkriterier (Granger, 2017, s. 429), som er satt sammen for engelsk, utgjør følgelig størstedelen av innlærerkorpusene i verden (Université catholique de Louvain, 2019). I de senere årene er det også blitt laget innlærerkorpus for mindre språk som norsk, f.eks. *ASK – Norsk andrespråkskorpus* samt NORINT-korpuset, som denne artikkelen handler om.

NORINT-korpuset er så langt det eneste korpuset som inneholder norsk muntlig innlærerspråk. Det var Liv Andlem Harnæs, som underviste i norsk for internasjonale studenter ved ILN, som tok initiativet til å lage et innlærerkorpus ved UiO i 2014. Resultatet ble NORINT-korpusets to delkorpus med talespråksdata: *NORINT tale* og *NORINT opplest*. Hovedmålet med utviklingen av *NORINT tale* og *NORINT opplest* var å skape muligheter for forskning på muntlige ferdigheter, særlig uttale, hos internasjonale studenter med norskferdigheter på eller over CEFR-nivå B1. Else Ryen supplerte korpuset med et tredje delkorpus, *NORINT tekst*, som er et skriftspråkskorpus. *NORINT tekst* ble etablert for å kunne se materialet fra *NORINT tale* og *NORINT opplest* i sammenheng med et skriftspråksmateriale samlet inn fra de samme informantene. NORINT-korpuset ble utviklet i samarbeid med Annely Tomson, som underviser i norsk for internasjonale studenter, samt Janne Bondi Johannessen, Kristin Hagen og Joel Priestley fra Tekstlaboratoriet, ILN, UiO. Arbeidet har fått noe økonomisk støtte fra ILN til transkripsjon og utvikling.

Nedenfor presenterer vi først kort den historiske utviklingen og definisjonen av elektroniske innlærerkorpus. Deretter ser vi på datamaterialet i NORINT-korpuset og *Glossa*, søkeprogrammet for *NORINT tale*, *NORINT opplest* og *NORINT tekst*. Videre blir bruken av NORINT-korpuset i forskning diskutert. Til slutt ser vi nærmere på ulike egenskaper

[2] Hva som defineres som naturlige og nesten-naturlige språkdata, blir forklart nedenfor.

ved ASK og NORINT-korpuset og skriver om mulighetene og begrensningene i NORINT-korpuset ved å sammenlikne det med ASK.

Hva gjør et elektronisk innlærerkorpus til et elektronisk innlærerkorpus?

Det første generelle elektroniske korpuset, *Brown Standard Corpus of American English* (University of Essex, 1998), ble utviklet i 1960-årene, av lingvistene Henry Kučera og Winthrop Nelson Francis ved Brown University, Rhode Island, USA. Før 1960 ble språkbruksdata også innsamlet, men den gang ble de observerte språkdataene nedtegnet på papir og samlet i esker. Denne type korpus omtales som før-elektroniske korpus (Kennedy, 1998, s. 13). Mens elektroniske korpus typisk inneholder store, lett søkbare datamengder hentet fra mange informanter og er annotert, dvs. utrustet med tilleggsinformasjon som f.eks. grammatisk informasjon, omfatter før-elektroniske korpus små datamengder hentet fra et betydelig lavere antall informanter. Siden de før-elektroniske korpusene er papirbaserte, er det i tillegg tidkrevende og upraktisk å hente frem ønsket informasjon fra språkdataene i dem.

Den første versjonen av *International Corpus of Learner English* (*ICLE*) (Université catholique de Louvain, 2009), det første elektroniske innlærerkorpuset, sto ferdig i 2002 (Cock, Gilquin, Meunier & Paquot, 2011, s. 2). Arbeidet med å lage ICLE ble startet av en lingvist og pioner innenfor utviklingen av elektroniske innlærerkorpus, Sylviane Granger, ved Université catholique de Louvain i Belgia i 1990. Det var et samarbeid med flere andre universiteter, og ICLE ble satt sammen med det formål å gi andrespråksfeltet de samme mulighetene som korpuslingvister innenfor generell korpuslingvistikk hadde fått (Granger, 2012, s. 7).

ICLE inneholder skriftlige språkbruksdata, og de fleste elektroniske innlærerkorpus er skriftspråkskorpus[3] siden talespråkskorpus er mer ressurskrevende å utvikle. Blant talespråkskorpus finnes det korpus som kun består av detaljerte transkripsjoner som f.eks. *Louvain International Database of Spoken English Interlanguage* (Université catholique de Louvain, 2010) der lydopptakene ikke er tilgjengeliggjort for brukeren. En annen type talespråkskorpus er korpus der man også kan høre

[3] Det står 189 innlærerkorpus oppført i en oversikt laget av Université catholique de Louvain (2019). Bare 40 av dem står oppført som talespråkskorpus.

lydmaterialet, ofte delt opp i mindre deler, f.eks. en tur (en sekvens der en person har ordet), ved å spille av lydfilen (Ballier & Martin, 2015, s. 110). *NORINT tale* og *NORINT opplest* hører til den sistnevnte kategorien, og man kan både se transkripsjonen og høre den tilhørende talen via et enkelt og funksjonelt søkegrensesnitt på web. For å gi meningsfulle søkeresultater er talematerialet i de muntlige NORINT-korpusene delt opp i segmenter, der et segment er definert som den minste enheten en samtale skal deles opp i, for å ha enheter som er mulige å ordklassetagge og analysere syntaktisk med automatiske annoteringsverktøy.

Språkdataene i et innlærerkorpus kategoriseres enten som naturlige eller nesten-naturlige. Begrepet nesten-naturlig er brukt til å fremheve behovet for språkbruksdata som gjengir språkets naturlige bruk så virkelighetsnært som mulig, dvs. språkbruken er situasjonelt og interaksjonelt autentisk (Ellis & Barkhuizen, 2005, s. 7). Siden det er vanskelig å få tak i helt naturlige språkbruksdata, dvs. språkbruksdata produsert i autentiske kommunikasjonssituasjoner der innlæreren selv velger og har kontroll over sin språkbruk, inneholder elektroniske innlærerkorpus oftest nesten-naturlige data (Granger, 2012, s. 8). Selv om argumenterende og fortellende tekster, åpne intervjuspørsmål og samtaler som baserer seg på et bredt utvalg av forhåndsbestemte emner, legger til rette for at innlæreren i stor grad selv kan bestemme over språkbruken, forutsetter en slik språkdataproduksjon en viss kontroll fra forskerens side (Ragnhildstveit, 2018, s. 425). Språkdataene i både ASK og NORINT-korpuset regnes som nesten-naturlige – L2-innlærernes språkbruksdata i de to korpusene er produsert i sammenheng med både tester/skriftlige eksamener (i ASK og NORINT-korpuset) og som intervjuer og samtaler (kun i NORINT-korpuset). Datamaterialet i ASK og NORINT-korpuset er innsamlet med en bestemt hensikt, dvs. at språkbruksdataene i de to korpusene er satt sammen med det siktemål å bedrive forskning på L2-tilegnelse. Språkdataene utgjør følgelig ingen tilfeldig samling av tekster/lyd- og videoopptak, men datamaterialet er blitt valgt ut med en bestemt hensikt etter bestemte kriterier, og som nevnt i innledningen, er utvelgelseskriterier en av forutsetningene for at elektroniske innlærerkorpus kan omtales som det.

I det følgende diskuterer vi hvordan NORINT-korpuset er satt sammen.

Datamaterialet i NORINT-korpuset

NORINT tale, *NORINT opplest* og *NORINT tekst* består av muntlig og skriftlig L2-produksjon. Informantene i korpuset er internasjonale studenter som gikk på norskkurs for viderekomne, *ISSN0130 – Intensivt mellomkurs i norsk, trinn III*[4] *(trinn 3)*, ved UiO sommeren 2014 og 2015. Tekstene og personinformasjonen i NORINT-korpuset er anonymisert, og bruk av lyd- og videoopptakene i korpuset er godkjent av Norsk senter for forskningsdata (NSD).

NORINT tale består av intervjuer av og samtaler med 48 informanter, i alt nesten 104 000 token (ord og skilletegn). Det er gjort video- og lydopptak der studentene intervjues av en underviser som stiller spørsmål om deres bakgrunn, studier, arbeid og fremtidsplaner. Temaene er begrenset for å sikre like rammer for intervjuene. I tillegg er det gjort video- og lydopptak der informantene samtaler to og to om emner som kultur, fritid, reiser eller livet i Norge. Hva det snakkes om, velger studentene fritt fra en liste over mulige samtaleemner de fikk utdelt rett før opptakene. Intervjuene og samtalene er nesten like lange, og opptak av hver student utgjør totalt 30–40 minutter. Se tabell 1 på neste side for en fullstendig oversikt over hvilke førstespråk (L1) som finnes i *NORINT tale*, og antall informanter som har ett og samme førstespråk.

Opptakene er transkribert med transkripsjonsprogrammet ELAN (Max Planck Institute for Psycholinguistics, 2020). *NORINT tale* er dessuten annotert med grammatisk informasjon som ordklasse og andre morfologiske trekk samt med informasjon som beskriver den muntlige interaksjonen, f.eks. «sukk», «kremter», «hvisking». Den grammatiske informasjonen er automatisk lagt til ved hjelp av *NoTa-taggeren*[5], en automatisk talemålstagger.

NORINT opplest har 57 informanter: 48 av dem er de samme som har bidratt til *NORINT tale*. Informantene leser opp 60 utvalgte setninger og en kort historie.[6] De løsrevne setningene og teksten er de samme som

[4] ISSN0130 – Intensivt mellomkurs i norsk, trinn III: https://www.uio.no/studier/emner/iss/sommerskolen/ISSN0130/ (20.03.2021).

[5] NoTa-taggeren er en statistisk talemålstagger (TreeTagger) som er trent på talemålsmaterialet fra NoTa-Oslo, se under fanen «Transkripsjon» her: http://www.tekstlab.uio.no/nota/oslo/index.html (20.03.2021).

[6] Lenke til de 60 setningene og historien: http://tekstlab.uio.no/norint_opplest/innlesning_hele.pdf (20.03.2021).

Tabell 1: Informanter gruppert etter førstespråk i *NORINT tale (Ta), NORINT opplest (O)* og *NORINT tekst (Te).*

Førstespråk	Antall informanter			Førstespråk	Antall informanter			Førstespråk	Antall informanter			Førstespråk	Antall informanter		
	Ta	O	Te		Ta	O	Te		Ta	O	Te		Ta	O	Te
albansk			1	italiensk	1	2	1	persisk			2	tagalog	1	1	2
arabisk	2	2	3	japansk	1	1	1	polsk	2	2	5	tamil			1
bulgarsk	3	3	4	katalansk			1	portugisisk	1	1	1	thai	1	1	1
chichewa			1	kurdisk (sorani)	1	1		punjabi	1	1	2	tigrinja	1	1	1
engelsk	3	4	17	latvisk	1	3	1	rumensk	2	1	3	tyrkisk	3	2	3
farsi		1	1	litauisk			3	russisk	7	9	17	tysk	2	2	2
filippinsk	1	1	3	makedonsk	1	1	1	serbisk	2	2	6	ukrainsk		1	2
georgisk			2	kinesisk (mandarin)	2	2	4	singalesisk			2	ungarsk	2	2	2
hindi		1	1	marathi	1	1	1	slovakisk	1	1	1	urdu	1	2	6
ibo			1	nederlandsk	2	3	4	spansk	2	2	4	urhobo			1

ved *Språkmøterprosjektet* ved NTNU[7]. Det finnes bare lydopptak av opplesningene, og setningene i *NORINT opplest* er ikke grammatisk tagget. I tabell 1 ovenfor finnes det en fullstendig oversikt over hvilke førstespråk som finnes i *NORINT opplest*, og antall informanter som har ett og samme førstespråk.

NORINT tekst er et skriftspråkskorpus og består av 116 eksamensbesvarelser fra *ISSN0130 – Intensivt mellomkurs i norsk, trinn III* som delvis er skrevet av de samme studentene som er informanter i de muntlige delene av korpuset. Av hensyn til personvern er det imidlertid ikke synlige koplinger i korpuset, men det er mulig å kontakte Tekstlaboratoriet for å få denne informasjonen.

Tekstene foreligger i tre ulike versjoner: a) en pdf av den håndskrevne originalversjonen av den skriftlige eksamensbesvarelsen i sin helhet (lytteprøvesvar, leseforståelsesoppgaver, grammatikkoppgaver og en stiloppgave (en argumenterende/fortellende innlærertekst)), b) en innskrevet nøyaktig kopi av den argumenterende/fortellende innlærertekstens originalversjon og c) en versjon av samme tekst der alle ortografiske feil er rettet, sammen med utvalgte morfologiske og syntaktiske feil. Innskrevet kopi av den argumenterende/fortellende innlærertekstens ori-

[7] Språkmøterprosjektets prosjektbeskrivelse: https://www.yumpu.com/no/document/read/30386800/beskrivelse-ntnu (20.03.2021).

ginalversjon og den korrigerte versjonen er lenket sammen. Tekstene er automatisk annotert med en grammatisk tagger for skriftspråk, *Oslo-Bergen-taggeren* (Johannessen, Hagen, Lynum & Nøklestad, 2012).

Se tabell 1 på forrige side for en fullstendig oversikt over hvilke førstespråk som finnes i *NORINT tekst*, og antall informanter som har ett og samme førstespråk.

Utvelgelseskriterier for datamaterialet i NORINT-korpuset

Som nevnt i innledningen defineres elektroniske innlærerkorpus som elektroniske samlinger av naturlige eller nesten-naturlige språkdata produsert av fremmedspråks- eller andrespråksinnlærere og som er satt sammen etter eksplisitte designkriterier (Granger, Gilquin & Meunier, 2015, s. 1). For å unngå at NORINT-korpuset ble en lite interessant «blandet-drops»-samling (Gilquin, 2015, s. 14) av innlærerdata, ble NORINT-korpusets språkbruksdata valgt ut etter følgende kriterier: *trinn 3-kriteriet, førstespråkskriteriet, informasjonskriteriet* og *brukervennlighetskriteriet.* Disse fire kriteriene baserer seg på og er en tilpasset versjon av designkriteriene til ASK: *morsmålskriteriet, informasjonskriteriet* og *bestått-kriteriet* (Tenfjord, Hagen & Johansen, 2009, s. 55–57).

Når det gjelder *trinn-3-kriteriet*, har hovedmålet med utviklingen av *NORINT tale* og *NORINT opplest* vært å skape muligheter for forskning på muntlige ferdigheter hos internasjonale studenter med norsk-ferdigheter på eller over CEFR-nivå B1. Følgelig er NORINT-korpusets informanter studenter som hadde meldt seg opp til og fulgte *ISSN0130 – Intensivt mellomkurs i norsk, trinn III* ved UiO. For å kunne melde seg opp til trinn 3 forutsettes det at studentene har bestått *NORINT0120 – Norsk for internasjonale studenter, trinn 2*[8] eller et tilsvarende kurs med karakteren D eller bedre på både muntlig og skriftlig eksamen[9]. Så lenge det ikke er blitt utarbeidet eksplisitte kriterier for nivåplassering av karakterer fra trinn 2 ved UiO på CEFR-skalaen, kan ikke gjennomført trinn 2 garantere at studenten er på minst CEFR-nivå B1. På grunn av ressursmangel var det ikke mulig å vurdere informantenes

[8] NORINT0120 – Norsk for internasjonale studenter, trinn 2: https://www.uio.no/studier/emner/hf/iln/NORINT0120/ (20.03.2021).

[9] Info om obligatoriske forkunnskaper: https://www.uio.no/studier/emner/hf/iln/NORINT0130/index.html (20.03.2021).

norskferdigheter i henhold til CEFR, men video- og lydopptakene ble gjennomført i andre halvdel av semesteret for å øke sannsynligheten for at de potensielle informantene hadde bedret norskferdighetene sine og var på eller over CEFR-nivå B1. Tekstene i *NORINT tekst* er eksamensbesvarelser fra trinn 3 som er ment å måle norskferdigheter på eller over CEFR-nivå B1 i likhet med *Språkprøven i norsk for voksne innvandrere* og *Test i norsk – høyere nivå* i ASK (Tenfjord et. al., 2009, s. 52).

Førstespråkskriteriet sikrer et forholdsvis bredt utvalg av førstespråkene som snakkes av studenter på trinn 3. Det kommer internasjonale studenter fra hele verden til Oslo for å studere ved UiO. Ambisjonen om å sikre språkbruksdata fra så mange ulike L1-brukere som mulig er ikke realistisk av økonomiske grunner – det vil være for ressurskrevende å utvide NORINT-korpuset til å bli et så omfattende korpus at alle førstespråkene fra trinn 3 er representert i det. Derfor vil en fremtidig utvidelse av korpuset måtte basere seg på det allerede eksisterende utvalget av førstespråk.

Informasjonskriteriet sikrer at det dokumenteres selvrapporterte persondata for hver informant: opplysninger om alder, kjønn, førstespråk, hvilke andre språk hen behersker (flytende, godt, basiskunnskaper), oppholdstid i Norge, utdanning og yrke. Denne informasjonen er viktig i forskningssammenheng fordi disse variablene kan ha betydning for en informants L2-læring.

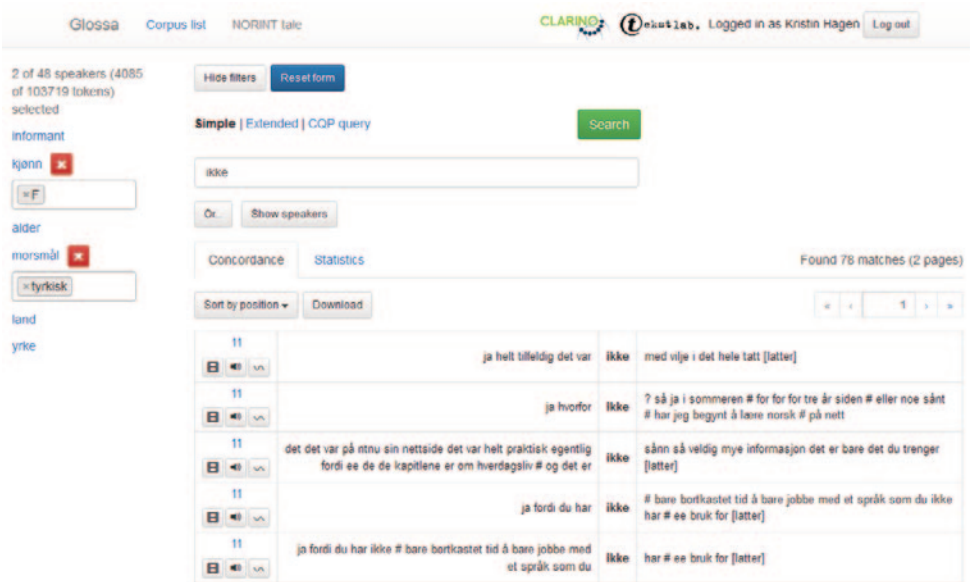
Brukervennlighetskriteriet sikrer at datamaterialet i NORINT-korpuset tilgjengeliggjøres på en effektiv og lett forståelig måte. Søkegrensesnittet som brukes i NORINT-korpuset, er et moderne, enkelt og funksjonelt søkegrensesnitt med avansert resultathåndtering for både skriftspråkskorpus og talespråkskorpus. Nedenfor gis en mer detaljert beskrivelse av hvordan man anvender søkeprogrammet NORINT-korpuset er lagt inn i.

Søkemuligheter i NORINT-korpuset

Tekster og transkripsjoner i NORINT-korpuset er lagt inn i søkesystemet *Glossa* (Kosek, Nøklestad, Priestley, Hagen & Johannessen, 2015; Nøklestad, Hagen, Johannessen, Kosek & Priestley, 2017). Glossa brukes i over tretti av korpusene som er utviklet ved Tekstlaboratoriet ved ILN, f.eks. tekstkorpuset *Leksikografisk bokmålskorpus*, parallell-

korpuset *Oslo Multilingual Corpus* og talespråkskorpusene *Nordisk dialektkorpus* og *NoTa-Oslo*. Systemet har blitt utviklet over mange år, i stor grad ut fra forskernes ønsker. Resultatet er et søkegrensesnitt som er enkelt å bruke, men som samtidig tillater avanserte søk. Søkegrensesnittet har så å si samme utseende for alle typer korpus, både tekst- og talespråkskorpus.

NORINT tekst, *tale* og *opplest* har innlogging via Feide, CLARIN og dessuten en lokal innloggingsmulighet. Det er laget en brukermanual for korpuset, men søkegrensesnittet skal være så enkelt å bruke at det ikke skal være nødvendig å gå på kurs eller lese igjennom brukermanualen på forhånd. Figur 1 viser et oversiktsbilde av *NORINT tale*. Hovedsøkesiden har en enkel søkeboks der det er mulig å søke etter ett eller flere ord. Resultatet kan filtreres ved hjelp av metadata som f.eks. alder, kjønn og førstespråk. Metadatakategoriene befinner seg til venstre på søkesiden. Søkeresultatet vises under søkeboksen. For tale-

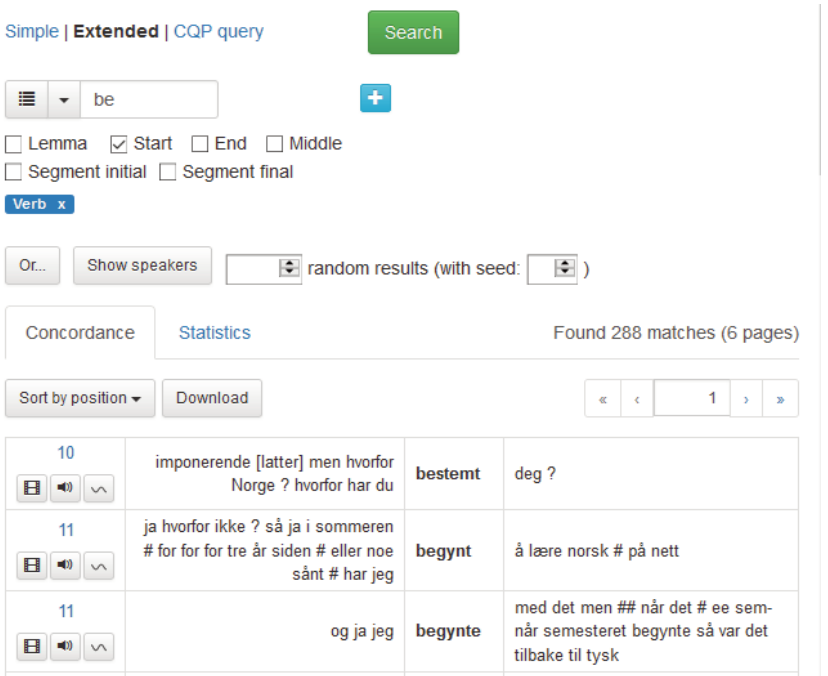


Figur 1: Oversiktsbilde av søkesiden for *NORINT tale* der det er søkt på ordet *ikke,* og der kvinnelige talere med tyrkisk som L1 er valgt i metadatamenyen til venstre. Over metadatakategoriene kan man se hvor mange informanter og token som er utgangspunkt for søket.

språkskorpusene kan man velge å vise søkeresultatet i en avspillingsboks for lyd eller video. Her kan man utvide konteksten slik at visningen omfatter mer enn ett segment.[10] Søkeresultatet/ segmentet kan også vises som bølgeformer og spektrogram.

Resultatene kan lastes ned som en tsv-, csv- eller en Excel-fil. Dermed kan man arbeide videre med resultatene, kategorisere og kommentere i ettertid. Under «Statistics» er det muligheter til å få ut ordlister med frekvenstall.

Mer avanserte søk kan gjøres ved å velge «Extended» over søkeboksen, se figur 2. Ved hjelp av menyer, bokser og klikk kan man søke



Figur 2: Et Extended-søk i *NORINT tale* der det er søkt på alle verbformer som begynner på *be-*. Ordklassen *verb* er valgt ved å trykke på menysymbolet ved siden av ordsøkeboksen.

[10] Siden fri tale normalt ikke kan deles inn i setninger slik som skriftlig tekst, brukes begrepet *segment* ofte i talespråkskorpus i stedet for setning. Segmenter kan tilsvare det vi vanligvis definerer som setninger, men det kan også være ufullstendige setninger uten subjekt og finitt verbal, eller ytringer som bare består av interjeksjoner.

på starten av ord, slutten av ord, tegn inne i et ord, begynnelsen av et segment og slutten av et segment. Et lemmasøk gir treff på alle bøyingsformer av søkeordet, ikke bare ordformen i søkeboksen. Siden korpusene er ordklassetagget, kan man også søke på ordklasse og annen morfologisk informasjon. Det er mulig å gjøre enda mer avanserte ordsøk ved å velge «CQP query» (Corpus Query Processor[11] query) over søkeboksen. I sistnevnte tilfelle kan man bruke søkespråket CQP direkte i søkeboksen dersom man behersker det.

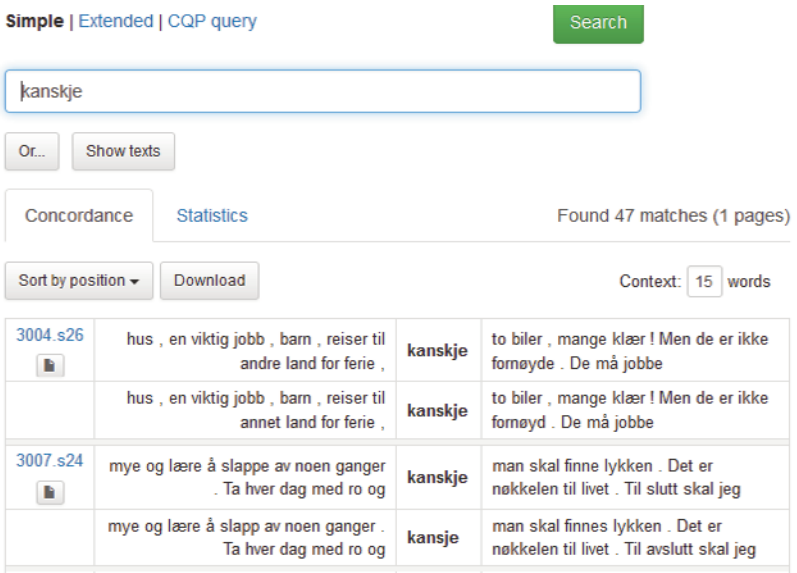
NORINT tale og *opplest* er ortografisk transkribert. Om et ord blir uttalt feil eller bøyd feil, blir dette ikke registrert i transkripsjonen, men man kan høre på resultatene og registrere og klassifisere feilene selv. Utenlandske ord eller norske ord som er brukt feil, er tagget som X-ord, og man kan søke på dem gjennom ordklassemenyen.

Søkegrensesnittet til *NORINT opplest* er nesten identisk med grensesnittet for korpuset med fri tale. Men siden alle informantene leser opp de samme setningene, gir det mindre mening å søke etter ord og grammatiske trekk. Tekstene er derfor ikke morfologisk tagget, men det er lenket til alle setningene og historien fra hovedsøkesiden. En måte å bruke korpuset på er å søke på de første ordene i en av setningene, og så høre gjennom hvordan hver informant uttaler setningen, eventuelt filtrere resultatene gjennom metadatamenyen.

NORINT tekst er et skriftspråkskorpus, og tekstene foreligger som nevnt ovenfor i tre ulike versjoner. Den innskrevne kopien av den argumenterende/fortellende innlærerteksten er koplet sammen med versjonen der det er rettet feil. Man kan søke i begge versjoner. Søker man i den ortografiske versjonen, får man et søkeresultat slik figur 3 (på neste side) viser, hvor det er enkelt å se de forskjellige variantene av eventuelle feilstavinger av ordet. Man kan også søke direkte på feilstavingene ved å søke i originaltekstene. *NORINT tekst* er ikke systematisk tagget med hensyn til feiltyper. Men siden den korrigerte versjonen er parallellstilt med originalversjonen og den korrigerte teksten skulle tagges med en automatisk, morfologisk tagger (Oslo-Bergen-taggeren), er det satt inn egne tagger for manglende hjelpeverb, infinitivsmerke og kopula i den korrigerte teksten. I tillegg er feilaktige samskrivinger løst opp. Man kan søke på taggene for eksempelvis manglende hjelpeverb i ordklassemenyen. Særskrivinger er satt sammen i den korrigerte versjonen, og

[11] Informasjon om CQP Query Language: http://cwb.sourceforge.net/temp/CQP Tutorial.pdf (20.03.2021).

den tomme plassen har fått en særskrivingstagg. Denne kan man også søke på i ordklassemenyen for å se hvilke ord som er skrevet som to ord i stedet for ett.



Figur 3: I *NORINT tekst* gjengis søkeresultatet over to linjer med den originale teksten under og den korrigerte over. Klikker man på pdf-symbolet til venstre, får man se eksamensbesvarelsen i sin helhet. Man kan søke direkte i originalversjonen ved å velge Extended-boksen og krysse av for «original».

Hvorfor bruke NORINT-korpuset?

En av fordelene ved å bruke NORINT-korpuset er at datamaterialet i det allerede er blitt innsamlet, så man trenger ikke å samle inn språkbruksdata selv. Dessuten er datamaterialet i NORINT-korpuset anonymisert og godkjent for bruk av NSD. For det tredje kan bruk av korpusbaserte metoder innen forskning på andrespråkstilegnelse gi viktig innsikt:

> Learner corpora have a lot to contribute to SLA research. They lead researchers to a better understanding of how second languages are learned and

> can help them answer questions at the heart of SLA research, such as the yet unresolved issue of the exact role of transfer in second language acquisition and the notion of avoidance. (Granger, 2008, s. 268–269)

Bruk av NORINT-korpuset i forskning på andrespråkstilegnelse kan bidra til å øke innsikten i hvordan man lærer andrespråk, men siden NORINT-korpuset inneholder relativt små datamengder fra relativt få individer, kan datamaterialet først og fremst brukes til å gjennomføre kvalitative analyser av muntlig og skriftlig innlærerspråk. Om man eksempelvis er interessert i å se nærmere på segmentale aspekter som f.eks. uttale av /e/ og /æ/ hos informanter med russisk som førstespråk, finnes det tale produsert av informanter med russisk som førstespråk i *NORINT tale* og *NORINT opplest*. Transfer forekommer også suprasegmentalt og påvirker følgelig prosodien i andrespråket. Ifølge Ringbom (2007) har fonologisk transfer en så sterk påvirkning på lydinnlæring i et andrespråk at selv innlærere med gode skriftlige og muntlige andrespråksferdigheter ikke snakker uten aksent (Ringbom, 2007, s. 62). For å belyse suprasegmentale sider, som f.eks. prosodien i andrespråket, ved de russiske informantenes uttale av norsk, kan man bruke datamaterialet fra både *NORINT tale, NORINT opplest* og *NORINT tekst* til å utføre en mindre omfattende studie.

En annen mulighet i NORINT-korpuset er å kunne gjennomføre undersøkelser der det fokuseres på å sammenligne språklige trekk i skrift og tale hos en og samme innlærer. Denne type studier har en lav grad av generaliserbarhet fordi funnene hovedsakelig vil si noe kun om andrespråkslæringen til de få informantene som er studert (Ellis, 2008, s. 8), men kvalitative analyser kan føre til interessante diskusjoner om muntlig og skriftlig andrespråkskompetanse og være et utgangspunkt for både semester-, bachelor- og masteroppgaver og mer omfattende fremtidige forskningsprosjekter.

NORINT tekst sammenliknet med ASK

Dette kapittelet gir en kort presentasjon av det norske innlærerkorpuset ASK. Deretter ser vi nærmere på ulike aspekter ved ASK og NORINT-korpuset og diskuterer mulighetene og begrensningene i NORINT-korpuset, spesielt i *NORINT tekst*.

Kort presentasjon av ASK-korpuset

ASK-korpuset er et norsk elektronisk andrespråkskorpus som består av tekster skrevet av innlærere ved to tester som måler språkferdigheter på ulikt nivå[12], der produksjonsforholdene var tilnærmet identiske for alle informantene da de befant seg i en testsituasjon (Tenfjord, Hagen & Johansen, 2009, s. 56). ASK ble utviklet ved Universitetet i Bergen i perioden 2003–2006 som en del av Norges Forskningsråds prosjekt *Parallelle korpus*. Kari Tenfjord sto i spissen som prosjektleder både for ASK-korpuset og for *ASKeladden*-prosjektet[13] konsentrert rundt korpuset. Prosjektet fokuserte på spørsmål relatert til transfer fra innlærernes førstespråk (Ragnhildstveit, 2018). ASK består av *hovedkorpuset, korrektkorpuset* og *hovedkorpuset/2015*.

I tillegg til metadata som f.eks. alder, kjønn, førstespråk, opphavsland, utdanning, yrke og lengde på norskopplæring kan språkdata i korpuset sorteres etter mer sosiologisk orientert informasjon som bl.a. motivasjon for norsklæring og sosial omgang med L1-brukere. Samtlige data i korpuset er tagget ved hjelp av Oslo-Bergen-taggeren og annotert med grammatiske kategorier. Ordklassetaggingen er også delvis redigert manuelt i etterkant. I tillegg er det tagget en rekke morfem-, leksem-, syntaks- og tegnsettingsfeil, der hver type feil har sin egen kode (f.eks. *W = galt ord*). Annotering og tillegging av feilkoder er gjort manuelt (Tenfjord, Hagen & Johansen, 2009; Johansen, 2010).

Grammatisk annotasjon i ASK muliggjør søk på forekomster av bestemte ordklasser, enkelte ord, lemma eller ordstrenger. Feilkoding brukt på innlærertekstene gir et korreksjonsforslag for ethvert avvik fra målspråksnormen; Nordanger påpeker at siden kodingen i ASK kun er deskriptiv og ikke analyserende, kan forskeren søke etter bestemte språklige trekk og samtidig beholde sin autonomi til å analysere og tolke funnene (2009, s. 14).

Et parallelt korrektkorpus gir mulighet for søk på konkrete og egendefinerte korreksjonsforslag (f.eks. *i* rettet til *på* eller *var* rettet til *har vært, føle* rettet til *kjenne* o.l.). Tekstene i kontrollkorpuset, som er skrevet av informanter med norsk som førstespråk, kan derimot brukes til å analysere over- eller underbruk av visse ord eller strukturer. Disse

[12] *Språkprøven i norsk for voksne innvandrere* og *Test i norsk – høyere nivå*, også kjent som Bergenstesten.

[13] Om Askeladden-prosjektet: https://www.uib.no/fag/andresprak/82682/forskingsprosjekt-i-norsk-som-andrespråk (20.03.2021).

egenskapene gjør ASK til et allsidig forskningsverktøy som er relativt lett i bruk og godt egnet for testing av hypoteser knyttet til andrespråksbruk og andrespråkstilegnelse (Nordanger, 2009, s. 14).

Utvalg av L1-bakgrunn i ASK er langt fra tilfeldig. Det ble sørget for at innlærernes L1 i korpuset representerte de største innvandrergruppene i Norge, samt at korpuset var typologisk variert (Johansen, 2010). I ASK finner man derfor tekster skrevet av informanter med albansk, bosnisk-kroatisk-serbisk, engelsk, nederlandsk, polsk, russisk, somali, spansk, tysk og vietnamesisk som L1. I hovedkorpus/2015 finnes det i tillegg annoterte innlærertekster skrevet av innlærere med arabisk, fransk, tamil, thai og tyrkisk som L1. Disse er foreløpig ikke lagt inn i korpuset på samme premisser som resten av dataene.

Tekstene i ASK var opprinnelig fordelt på to nivåer etter norskprøve, *Språkprøven i norsk for voksne innvandrere* og *Test i norsk – høyere nivå*. Det var først i 2010 at tekstene ble vurdert og plassert på ni ferdighetsnivåer i henhold til CEFR (Carlsen, 2010). Plassering av tekster på ulike nivåer, slik det er gjort i ASK, åpner ifølge Jarvis og Pavlenko (2008, s. 36) for kombinasjon av pseudo-longitudinelle metoder med tverrsnittsstudier. Forskeren kan undersøke samme antall innlærere i tverrsnittsstudier og se hvordan enkelte instansieringer av transfer arter seg på et lavere og på et høyere testnivå, noe som likner på en utvikling over tid, til tross for at man ikke observerer ett og samme individ (Jarvis & Pavlenko 2008, s. 36–37). Det er f.eks. mulig å undersøke mestring av inversjon på ulike ferdighetsnivåer, både innenfor en enkelt eller mellom flere L1-grupper, og kartlegge tendenser som f.eks. intragruppehomogenitet og intergruppeheterogenitet for å diagnostisere transfer i henhold til Jarvis og Pavlenkos metode for påvisning av transfer (Jarvis & Pavlenko, 2008).

Det omfattende og varierte materialet i ASK kan altså brukes til utprøving av ulike typer forskningsspørsmål, noe som bekreftes av en rekke masteroppgaver, doktorgradsavhandlinger og andre studier basert på ASK-dataene – med både kvantitativ og kvalitativ tilnærming. Se Golden, Jarvis og Tenfjord (2017) og Ragnhildstveit (2018) for en detaljert oversikt over forskning basert på data fra ASK.

Sammenlikning av ASK og NORINT-korpuset

Datateknologien gir et bredt spekter av muligheter og design som et moderne korpus kan lages etter. I korpusbasert andrespråksforskning er det

spesielt aktuelt med parallelle korpus eller spesialkorpus, der den første typen består av en kildetekst og oversettelse til et eller flere språk, mens et spesialkorpus inneholder data fra bestemte språkbrukergrupper, f.eks. L2-innlærere, og ikke er ment å være representativt for språket som helhet (Granger, 2008). Både NORINT-korpuset og ASK tilhører den andre typen, siden de er L2-korpus med data fra voksne innlærere med norsk som andrespråk.

Videre skiller man mellom åpne og lukkede korpus. Den første typen er åpen i den forstand at det til enhver tid kan tilføyes nye data. Lukkede korpus er derimot vanligvis tilknyttet et bestemt prosjekt der data samles inn i en forhåndsbestemt periode uten at man har planer om å utvide innholdet. Selv om prosjektet knyttet til ASK er avsluttet, og NORINT-korpuset ikke er blitt supplert med nye data siden 2015, kan begge regnes som åpne samlinger med mulighet for inntak av nye data. Som nevnt ovenfor er det bl.a. flere annoterte tekster fra nye innlærergrupper som venter på å bli lagt inn i hovedkorpuset til ASK, og det planlegges også en utvidelse av datamengden i NORINT-korpuset.

Med utgangspunkt i L2-korpus kan man analysere forekomster av språklige variabler på tvers av brukergrupper og i henhold til lingvistiske parametre. Både NORINT-korpuset og ASK kan kategoriseres som akademiske korpus, noe som innebærer at den språklige produksjonen de inneholder, er begrenset til enten eksamenssituasjoner (ASK og *NORINT tekst*) eller intervju- og samtalesituasjoner (*NORINT tale* og *NORINT opplest*). Videre er det vanlig å skille mellom longitudinelle korpus og tverrsnittkorpus, der forskjellen går ut på at man i det første tilfellet følger en bestemt gruppe innlærere som har utviklet L2-kunnskaper over tid, mens i det andre er det ofte data som er innsamlet samtidig, men gjerne fra flere informanter. Følgelig kan man si at ASK delvis kan kategoriseres som både et pseudo-longitudinelt og tverrsnittkorpus, ettersom det inneholder data fra ulike mestringsnivåer. Blir *NORINT tekst* supplert med flere tekster, får man også mulighet til å gjennomføre reliable kvantitative tverrsnittsstudier basert på materialet derfra.

Granger (2002) påpeker at korpus ytterligere deles inn etter pedagogiske implikasjoner. Hun skiller mellom korpus med umiddelbar bruk, der data man gjenvinner fra korpuset, kan anvendes like etter undersøkelsen, og utsatt bruk, der det trengs lengre tid før de kan benyttes. Den første typen brukes av innlærere som selv har produsert tekstene, og som på den basis trekker konklusjoner og bearbeider sin egen

språkbruk. Den andre typen brukes derimot for dyptgående studier og utvikling av mer effektive undervisningsverktøy. Både ASK og NORINT-korpuset må regnes som representanter for den andre kategorien.

Språkdataene i et andrespråkskorpus kan enten bestå av materiale som allerede eksisterer eller av materiale skapt spesielt for et korpusrelatert formål. I det andre tilfellet får man språkbruksdata som følger de samme designkriteriene, og dette er ifølge Granger (2002) en av forutsetningene for moderne korpusdesign. Felles for ASK og *NORINT tekst* er at de består av eksisterende tekster, mens språkdataene i *NORINT tale* og *NORINT opplest* er produsert med det formål å forsyne korpuset med muntlig innlærerspråk. Kriteriene datamaterialet er blitt valgt ut etter, som f.eks. innlærernes L1 og ferdighetsnivå, utgjør korpusets metadata, og metadataene kan variere fra korpus til korpus. NORINT-korpuset inneholder mindre selvrapporterte persondata enn ASK.

Ideelt sett burde et innlærerkorpus ha et mest mulig homogent materiale med hensyn til måten data samles inn på (f.eks. naturlig språkbruk vs. testsituasjon, interaksjon mellom første- og andrespråksbruker vs. samtale mellom to andrespråksbrukere, andre- vs. fremmedspråksbrukere o.l.). Innsamling av data må i praksis ofte være basert på en mer pragmatisk tilnærming, slik at dataene ikke nødvendigvis blir homogene i en strengt statistisk forstand, men et resultat av det som er praktisk mulig å anskaffe av grunndata på innsamlingstidspunktet. Om man ser nærmere på NORINT-korpusets delkorpus *NORINT tekst* og sammenligner det med ASK, som også er et tekstkorpus, finner man likheter mellom kriterier for utvelgelse av teksttype og innsamlingstidspunkt. På den annen side er *NORINT tekst* og ASK veldig ulike når det gjelder andre aspekter, som f.eks. antall informanter i enkelte innlærergrupper. *NORINT tekst* byr på et bredt utvalg av førstespråk, men i enkelte L1-grupper er det flere informanter enn i andre, noe som totalt gir ulikt antall informanter per innlærergruppe. Analyser foretatt på basis av tekstene må derfor være av en mer kvalitativ karakter og kan ikke brukes til å analysere tendenser innenfor en gitt gruppe. Følgelig muliggjør ikke delkorpuset sammenligning av utvalgte trekk på tvers av førstespråk heller, for det er for få informanter til det. Resultatene fra enkelte innlærergrupper blir med andre ord lite generaliserbare i *NORINT tekst* sammenlignet med ASK, der det er mye data fra en bestemt innlærer-

gruppe (nesten[14] 100 tekster fra hver gruppe i hver av de to prøvene). I tabellen på neste side presenterer vi en sammenstilling av hovedtrekk ved *NORINT tekst* og ASK.

Kort oppsummert er både ASK og NORINT-korpuset spesialkorpus siden de er ment å representere et språk av en spesiell type, dvs. norsk innlærerspråk. Begge korpusene kan tilføyes nye språkdata for å øke datamengden, og derfor kategoriseres de som åpne korpus. Siden den språklige produksjonen begge korpusene inneholder, er begrenset til undervisningsrelaterte situasjoner, kategoriseres de som akademiske korpus. De pedagogiske implikasjonene tatt i betraktning representerer ASK og NORINT-korpuset innlærerkorpus med utsatt bruk, dvs. at begge korpusene kan brukes for dyptgående studier og utvikling av mer effektive undervisningsverktøy. Om man ser nærmere på ulikheter, er ASK et skriftspråkskorpus som er designet spesielt for å forske på tverrspråklig innflytelse, og takket være store datamengder, tekstmateriale på ulikt nivå og tekster vurdert i henhold til CEFR kan det brukes til både tverrsnittsstudier og pseudo-longitudinelle studier. ASK egner seg altså til både kvalitative og kvantitative studier med mulighet for signifikanstesting av hypoteser. Selv om NORINT-korpuset inneholder språkbruksdata fra informanter med mange flere førstespråk enn ASK, er datamengden i NORINT-korpuset foreløpig betydelig mindre, og derfor kan NORINT-korpuset kun brukes til å utføre kvalitative undersøkelser.

Avslutning

NORINT-korpuset består av tre delkorpus: *NORINT tale*, *NORINT opplest* og *NORINT tekst*, og som det fremkommer av delkorpusenes navn, inneholder NORINT-korpuset både skriftlig og muntlig språkbruksdata. Hovedintensjonen med å etablere *NORINT tekst* var å kunne se materialet fra *NORINT tale* og *NORINT opplest* i sammenheng med et skriftspråksmateriale samlet inn fra de samme informantene. Det at en del informanter har bidratt med både muntlig og skriftlig språkbruksdata, og at det er mulig å sammenlikne språklige trekk i skrift og tale hos en og samme innlærer, gjør NORINT-korpuset unikt i norsk sammenheng. Delkorpusene kan også brukes som selvstendige korpus. Brukervenn-

[14] Det er færre besvarelser fra Bergenstesten i enkelte L1-grupper.

Tabell 2: Sammenstilling av viktigste trekk ved *NORINT tekst* og ASK.

	NORINT tekst	**ASK**
Type data	• rådata fra testsituasjon med ortografi og grammatikk som i originalen • innskrevet kopi med riktig ortografi og grammatikk • skannete testbesvarelser	• rådata fra to testsituasjoner med ortografi og grammatikk som i originalen • korrektkorpus • kontrollkorpus (200 tekster som er skrevet av informanter som har norsk som førstespråk)
Antall tekster	• 116 tekster • ingen overlapp mellom informanter i *NORINT tekst*, men delvis overlapp mellom informanter i *NORINT tekst*, *tale* og *opplest*	• nesten 2000 tekster • ingen overlapp mellom informanter (1 informant = 1 oppgave)
Kontrollkorpus	Nei	Ja (200 tekster, 100 fra hver test)
Korrektkorpus	• ortografi • utvalgte grammatiske trekk er rettet for å få tagget korpuset automatisk • rettet parallellkorpus	• korreksjonsforslag for alle feilforekomster • rettet parallellkorpus
Tagging	• ordklassetagging (automatisk, Oslo-Bergen-taggeren) • manuelt tagget for noen feiltyper (manglende hjelpeverb, infinitivsmerke og kopula, samskriving og særskriving)	• ordklassetagging (automatisk, Oslo-Bergen-taggeren) • grammatisk tagging - morfem, leksem, syntaks, tegnsetting (automatisk, Oslo-Bergen-taggeren) • manuell feilkoding • manuell korrigering
Språknivå	• på eller over B1	• A1/A2, A2, A2/B1, B1, B1/B2, B2, B2/C1, C1/C2, C2
Kontrollert for interrater-reliabilitet	Nei	Ja
Informantenes førstespråk	• 39 språk	• 10 språk (+5 i hovedkorpus/2015)
Metadata	• alder, førstespråk, kjønn, andre språk (flytende, godt, basiskunnskaper), oppholdstid i Norge, utdanning, land, yrke	• bl.a. alder, førstespråk, kjønn, land, utdanning, engelsk, opphold, kursmål, sosialt, bruk av norsk, startdato for norskopplæring, opplæringstid, m.fl.
Søkemotor	Glossa	Corpuscle
Konkordanseliste	Ja	Ja
Tilgang	Feide, CLARIN, lokal pålogging	Feide og CLARIN via søknad

ligheten og søkemulighetene i *NORINT tale*, *NORINT opplest* og *NORINT tekst* gjør at delkorpusene egner seg godt til forskning og til mer eller mindre omfattende studentoppgaver på både bachelor- og masternivå. Bruksmulighetene til korpuset i dets nåværende form er imidlertid noe begrenset. Det er planer om å supplere NORINT-korpuset med språkdatamengder fra et utvalg av innlærergrupper med samme L1 og norskferdigheter på eller over CEFR-nivå B1 for å sikre at reliable kvantitative tverrsnittsstudier kan baseres på NORINT-korpuset. Følgelig har NORINT-korpuset potensial til å videreutvikles til en enda nyttigere ressurs i korpusbasert forskning på andrespråkstilegnelse.

Litteraturliste

Ballier, N. & Martin, P. (2015). Speech annotation of learner corpora. I S. Granger, G. Guilqin & F. Meunier (Red.), *The Cambridge Handbook of Learner Corpus Research* (s. 107–134). Cambridge: Cambridge University Press.

Carlsen, C. (2010). Å knytte ASK til Rammeverket – hvorfor og hvordan. I H. Johansen, A. Golden, J. E. Hagen & A.-K. Helland (Red.), *Systematisk, variert, men ikke tilfeldig: Antologi om norsk som andrespråk i anledning Kari Tenfjords 60-årsdag* (s. 133–147). Oslo: Novus forlag.

Cock, S. D., Gilquin, G., Meunier, F. & Paquot, M. (2011). Putting corpora to good uses: A guided tour. I F. Meunier, S. D. Cock, G. Gilquin & M. Paquot (Red.), *A Taste for Corpora: In honour of Sylviane Granger* (s. 1–6). Amsterdam & Philadelphia: John Benjamins.

Council of Europe (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment.* Cambridge: Cambridge University Press.

Ellis, R. (2008). *The Study of Second Language Acquisition.* Second Edition. Oxford: Oxford University Press.

Ellis, R. & Barkhuizen, G. (2005). *Analysing learner language*. Oxford: Oxford University Press.

Gilquin, G. (2015). From design to collection of learner corpora. I S. Granger, G. Gilquin & F. Meunier (Red.), *The Cambridge Handbook*

of Learner Corpus Research (s. 9–34). Cambridge: Cambridge University Press.

Golden, A., Jarvis, S. & Tenfjord, K. (2017). *Crosslinguistic Influence and Distinctive Patterns of Language Learning*. Bristol: Multilingual Matters.

Granger, S. (2002). A bird's-eye view of learner corpus research. I S. Granger, J. Hung & S. Petch-Tyson (Red.), C*omputer Learner Corpora, Second Language Acquisition and Foreign Language Teaching* (s. 3–33). Amsterdam & Philadelphia: John Benjamins.

Granger, S. (2008). Learner corpora. I A. Lüdeling & M. Kytö (Red.), *Corpus Linguistics: An International Handbook. Volume 1* (s. 259–275). Berlin & New York: Walter de Gruyter.

Granger, S. (2012). How to use foreign and second language learner corpora? I A. Mackey & S. M. Gass (Red.), *Research Methods in Second Language Acquisition: A Practical Guide* (s. 7–29). Malden: Blackwell.

Granger, S. (2017). Learner Corpora in Foreign Language Education. I S. Thorne & S. May (Red.), *Language, Education and Technology: Encyclopedia of Language and Education* (3rd ed.), 427–440. Cham: Springer. https://doi.org/10.1007/978-3-319-02237-6_33

Granger, S., Gilquin, G. & Meunier, F. (2015). Introduction: Learner corpus research – past, present and future. I S. Granger, G. Gilquin & F. Meunier (Red.), *The Cambridge Handbook of Learner Corpus Research* (s. 1–5). Cambridge: Cambridge University Press.

Jarvis, S. & Pavlenko, A. (2008). *Crosslinguistic Influence in Language and Cognition.* New York: Taylor & Francis.

Johannessen, J. B., Hagen, K., Lynum, A. & Nøklestad, A. (2012). OBT+stat: A combined rule-based and statistical tagger. I G. Andersen (Red.), *Exploring Newspaper Language: Using the web to create and investigate a large corpus of modern Norwegian* (s. 51–66). Amsterdam: John Benjamins.

Johansen, H. (2010). Kontroll, Analyse, Reliabilitet, Innlærerkorpus og Transfer – Slik Egner seg Korpus til Studier av Transfer hos Innlærere. I H. Johansen, A. Golden, J. E. Hagen & A.-K. Helland (Red.), *Systematisk, variert, men ikke tilfeldig: Antologi om norsk som andrespråk i anledning Kari Tenfjords 60-årsdag* (s. 160–174). Oslo: Novus forlag.

Kennedy, G. (1998). *An introduction to Corpus Linguistics*. London & New York: Longman.

Kosek, M., Nøklestad, A., Priestley, J., Hagen, K. & Johannessen, J. B. (2015). NEALT Proceedings Series 25. I G. Grigonytė, S. Clematide, A. Utka & M. Volk (Red.), *Visualisation in speech corpora: Maps and waves in the Glossa system, Proceedings of the Workshop on Innovative Corpus Query and Visualization Tools at NoDaLiDa 2015* (s. 23–31). Vilnius, Litauen, 11.–13. mai 2015.

Max Planck Institute for Psycholinguistics. (2020). ELAN 6.0. Hentet fra https://archive.mpi.nl/tla/elan

Neri, A., Cucchiarini, K. & Strik, H. (2006). Selecting segmental errors in non-native Dutch for optimal pronunciation training. *IRAL – International Review of Applied Linguistics in Language Teaching, 44,* 357–404. doi: 10.1515/IRAL.2006.016

Nordanger, M. (2009). *Keiserens nye klær? Lingvistisk og konseptuell transfer i markeringen av grammatikalisert definitt referanse i russisktalende og engelsktalendes norske mellomspråk: en studie basert på ASK* [Masteroppgave]. Universitetet i Bergen.

Nøklestad, A., Hagen, K., Johannessen, J. B., Kosek, M. & Priestley, J. (2017). A modernized version of the Glossa corpus search system. I J. Tiedemann (Red.), *Proceedings of the 21st Nordic Conference on Computational Linguistics (NoDaLiDa)* (s. 251–254).

Ragnhildstveit, S. (2018). ASK – et elektronisk innlærerkorpus til bruk i andrespråksforskning. I A.-K. H. Gujord & G. T. Randen (Red.), *Norsk som andrespråk – perspektiver på læring og utvikling* (s. 422–448). Oslo: Cappelen Damm.

Ringbom, H. (2007). *Cross-linguistic Similarity in Foreign Language Learning*. Clevedon: Multilingual matters.

Tenfjord, K., Hagen, J. E. & Johansen, H. (2009). Norsk andrespråkskorpus (ASK) – design og metodiske forutsetninger. *NOA – Norsk som andrespråk, 25*(1), 52–81.

Université catholique de Louvain. (2009). The International Corpus of Learner English. Hentet fra https://uclouvain.be/en/research-institutes/ilc/cecl/icle.html

Université catholique de Louvain. (2010). LINDSEI CD-ROM and handbook. Hentet fra https://uclouvain.be/en/research-institutes/ilc/cecl/lindsei-cd-rom-and-handbook.html

Université catholique de Louvain. (2019). Learner corpora around the world. Hentet fra https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html
Universitetet i Bergen. (2020, 25. april). *ASK – Norsk andrespråkskorpus*. Hentet fra https://clarino.uib.no/korpuskel/corpus-list
Universitetet i Oslo. (2020, 14. juni). *NORINT-korpuset*. Hentet fra https://www.hf.uio.no/iln/om/organisasjon/tekstlab/prosjekter/norint/
University of Essex. (1998, 5. februar). *The Brown Corpus*. Hentet fra https://www1.essex.ac.uk/linguistics/external/clmt/w3c/corpus_ling/content/corpora/list/private/brown/brown.html

Abstract

The *NORINT*[15] *Corpus* (University of Oslo, 2020) is a relatively new learner corpus compiled at the Department of Linguistics and Scandinavian Studies at the University of Oslo. The NORINT Corpus consists of spoken and written data elicited from adult learners of Norwegian (international students) with an intermediate command of the target language i.e., the B1 level or higher in accordance with the Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2001). In this article we outline how the data is transcribed and annotated and how the easy-to-use search program of the NORINT Corpus can be applied. Furthermore, we suggest how to employ the possibilities the NORINT Corpus offer in research. Finally, we compare the NORINT Corpus with *ASK – Norsk andrespråkskorpus* (University of Bergen, 2020) to discuss possibilities and limitations in the NORINT Corpus.

Keywords: *learner corpus, speech corpus, text corpus, second language learner corpus*

[15] NORINT is an abbreviation for Norwegian for International Students, and it refers to Norwegian courses offered for international students at the Department of Linguistics and Scandinavian Studies, University of Oslo.