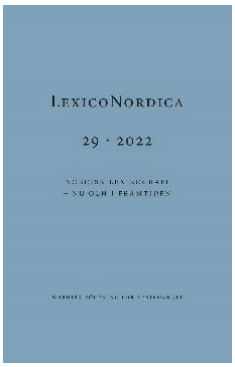
LexicoNordica

Titel:	Om å bygge en leksikalsk ressurs for diakron skriftspråksvariasjon
Forfatter:	Magnus Breder Birkenes, Lars G. Johnsen & Andre Kåsen
Kilde:	LexicoNordica 29, 2022, s. 15-31
URL:	http://ojs.statsbiblioteket.dk/index.php/lexn/issue/archive



Om å bygge en leksikalsk ressurs for diakron skriftspråksvariasjon

Magnus Breder Birkenes, Lars G. Johnsen & Andre Kåsen

The present article sketches the process of constructing a pairing of modern word forms with their historical counterparts. We describe a particular pipeline for inducing such a lexical mapping, which results in a digital lexicographic resource. This resource can be used to amend existing digital dictionaries and build historical dictionaries, and it may form an essential part in applications and fields that work with textual data from a wide timespan.

1. Innledning

Moderne språkteknologi og digital tekstanalyse antar som regel at språket den opererer på, er standardisert og har få innslag av variasjon. For eksempel forutsettes det ofte at en gitt stamme- eller bøyningsform skrives på bare én måte innenfor et leksem. Dette er imidlertid ikke tilfellet. I norsk kan for eksempel en grammatisk funksjon med en viss morfologisk koding, for eksempel bestemt form entall, i noen tilfeller uttrykkes med ulike bøyingsaffikser og dermed anta ulike skriftformer. For eksempel kan bestemt form entall av *sol* i bokmål skrives enten *solen* eller *sola*.

Her er det altså snakk om synkron variasjon – som er ganske omfattende i norsk sammenheng (både i bokmål og i nynorsk). Men formvariasjon skaper også problemer i diakrone sammenhenger, for eksempel når man ønsker å se på historisk språk eller sammenligne eldre og nyere tekster, ettersom ordenes skrivemåte kan ha endret seg. For å kunne håndtere de forskjellige og noen ganger mange variantene rent praktisk er man imidlertid også nødt til å etablere én form som alle varianter kan kobles til, slik at for eksempel formerne *qvinde* og *kvinne* automatisk analyse-

res som varianter av samme leksem. I det følgende omtales denne samleformen som «grunnform».

Problemet med formvariasjon skyldes at koblingen mellom formene ikke er eksplisitt kodet i eksisterende ordbøker, men likevel underforstått. Mens det synkrone aspektet ved slik variasjon er godt ivaretatt i moderne ordbøker, både for menneskelig og maskinell tolkning, er det diakrone ikke tatt hensyn til i samme grad. Beskrivelse av det diakrone aspektet er ofte begrenset til etymologiske forhold eller henvisninger til eldre belegg (implisitt informasjon), hvor ordet gjerne har en helt annen skrivemåte. Ved å slå opp på leksemet *kvinne* i NAOB vil man i belegglisten finne formene *qvinde* (for eksempel hos Henrik Wergeland), *kvinde* (for eksempel hos Henrik Ibsen) og *kvinne* (for eksempel hos Dag Solstad) i både entall og flertall. Det er imidlertid ikke angitt eksplisitt at disse formene skal eksemplifisere *kvinne*. Den kunnskapen forutsettes det at leseren har selv.

Fra et leksikografisk ståsted tilhører historiske stavingsvarianter som *kvinne* og *kvinde* samme leksem. Nærmere bestemt er *kvinne* og *kvinde* varianter av den grammatiske formen ubestemt form entall, og *kvinne* representerer den moderne skrivemåten. Når eldre stavingsvarianter som *kvinde* innlemmes i tekstgrunnlaget til en ordbok, kan vi tenke oss at leksemet utvides med historiske former. I digital tekstanalyse er det likevel fullt mulig å tenke seg at det opprettes et nytt leksem med utgangspunkt i formen *kvinde* (bøyningsparadigmet er også grafemisk forskjellig fra det moderne).

Det at nye former introduseres gjennom modernisering og språkreformer, omtales i den datalingvistiske litteraturen som *lexical replacement*, altså leksikalsk utskiftning eller erstatning. Det vil si at en ny form erstatter den gamle. Formene fra ulike tidsepoker refereres til som historiske kognater. Den moderne formen er en liten endring av den historiske. I denne artikkelen tar vi bare for oss kognater og utskiftning med utgangspunkt i språkets gra-

femiske representasjon og holder talespråket utenfor. Skriftspråket kan endre seg uten at talen gjør det, og omvendt, og førstnevnte har i Norge endret seg mye gjennom rettskrivingsvedtak som har kommet på løpende bånd fra 1860-tallet og frem til i dag. Hvorvidt disse endringene reflekterer fonologiske endringer, skal vi ikke forfølge videre.

I denne artikkelen skal vi beskrive en metode for å knytte grafemiske ordformer fra ulike tidsperioder (som altså også kan kalles historiske kognater) sammen uten å se på leksemtilhørighet, og deretter skal vi presentere en måte å organisere formene videre i leksemer på, det vil si historisk fulle leksemer. Nærmere bestemt går metoden ut på at de historiske ordformene, enten de er grunnformer eller andre morfologiske varianter, blir koblet til en av de tilsvarende variantene i moderne norsk, som så kobles til et leksem.

Med denne metoden planlegger vi å lage en historisk ordliste. Ordlisten vil bli en digital ressurs som blant annet kan fungere som en utvidelse av eksisterende ordbøker ved at ordbøkene inneholder de historiske skrivemåtene. For å konstruere ordlisten benytter vi oss av det digitaliserte materialet ved Nasjonalbibliotekets samling samt metadata for gruppering og organisering av tekster.

I dette arbeidet vil vi av praktiske hensyn begrense oss til bokmål, som har en veldokumentert utviklingsbane fra dansk og frem til sin moderne form. Bokmål har dessuten et tilstrekkelig datagrunnlag. Nynorsk har en til dels uavhengig utvikling samt et betydelig magrere datagrunnlag.

2. Problemet

Fra et datamaskinelt perspektiv kan vi ta utgangspunkt i hva som møter algoritmene i digital tekstanalyse. For datamaskinen vil alt

som skrives likt, være likt, og alt som skrives forskjellig, være forskjellig. For å gruppere sammen distinkte forekomster etter andre kriterier enn at de er like på overflaten, kreves en representasjon i form av for eksempel en digital ordbok. Ved hjelp av en digital ordbok kan algoritmer gruppere ulike grunnformer og/eller bøyde former under det samme leksemet, for eksempel fastslå at forekomstene av *spise* og *spiste* kan grupperes under leksemet *spise*.

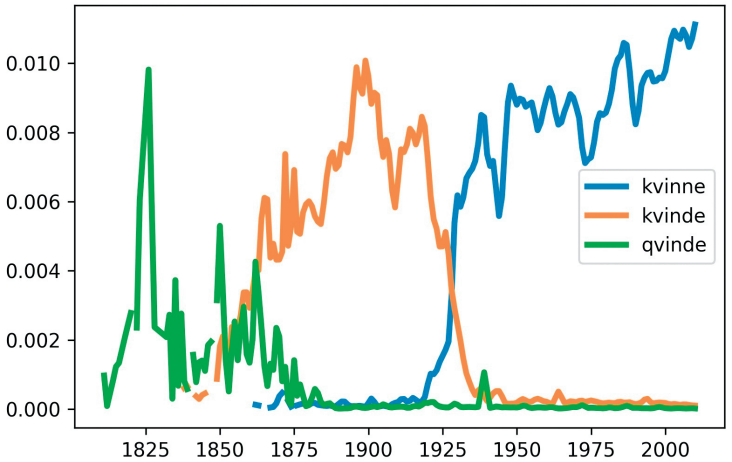
Vi føyer til at homografer ikke blir behandlet her. En og samme ordform kan tilhøre forskjellige leksemer, som *legger*, som kan være presens av verbet *legge* eller flertall av substantivet *legg*. Den grafemiske skrivemåten gir ingen hint om hvilket leksem som er i sving. For en slik disambiguering må man se på konteksten ordet står i. I det historiske tilfellet vil skriftkonvensjoner begrense omfanget av homografi, da substantiver på 1800-tallet typisk ble skrevet med stor forbokstav. Man hadde altså en kontrast mellom *Lægger* (substantiv) og *lægger* (verb).

I arbeidet vi presenterer her, er det forskjellige skrivemåter av en ordform som skal grupperes, og disse må derfor kodes med leksemtilhørighet og eventuelt grammatiske funksjoner. Men det må påpekes at denne kodingen i første omgang vil vise seg ved at for eksempel *legger* ses på som en utvikling av *lægger*. Grammatiske egenskaper blir tilordnet den historiske formen basert på en beskrivelse av den moderne. Se under for en beskrivelse av hvordan det kan gjøres.

2.1. Noen eksempler på historisk formvariasjon

Det er den nakne ordformen slik den står i teksten, som er utgangspunktet for arbeidet vårt med å utvikle en historisk ordliste. For eksempel ville en naiv analyse av 1800- og 1900-tallets tekster se på *kvinde* og *kvinne* som to forskjellige ord. I dag er det ingen digitale ordbøker for norsk som kobler de to formene sammen. Begge ordformene refererer til det samme og burde derfor kunne

relateres til hverandre. Vi skal senere beskrive en løsning på dette problemet, men først skal vi se litt nærmere på den historiske utviklingen i bruk av kjente skrivemåter av ordet *kvinne* ved hjelp av såkalte n-gram-trendlinjer[1] (Birkenes et al. 2015), som vist i figur 1.



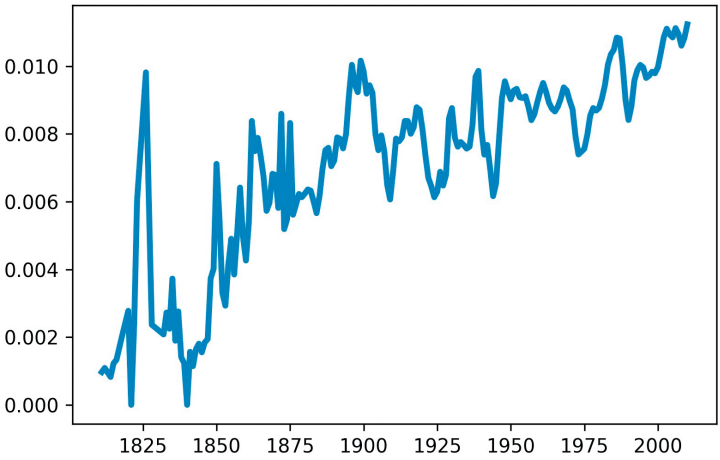
Figur 1: Trendlinjer for ulike skrivemåter av ordet *kvinne*.

Her har vi i tillegg med varianten *qvinde*. Grafen viser den relative frekvensen av de ulike stavingsvariantene av *kvinne* fra ca. 1800 og frem til år 2000.

I en praktisk søkekontekst der noen ønsker å se utviklingen i den relative frekvensen av ordet *kvinne* i norske bøker, kunne man tenke seg at de historiske skriftvariantene (de historiske kognatene) står for det samme, slik at trendlinjen for *kvinne* kan illustreres ved å slå sammen trendlinjene for de ulike skrivemåtene, som vist i figur 2.

En annen søkesituasjon kan være at noen er ute etter forskjellige kontekster for ordet *kvinne* i et historisk korpus og ønsker å

1 Trendlinjen viser den relative (prosentvise) frekvensen av en ordform i et korpus, det vil si hvor mange ganger ordformen forekommer sammenlignet med det totale antallet ordformer i korpuset.



Figur 2: Sammenslåtte trendlinjer fra figur 1.

få opp alle dokumenter som inneholder ordet, med tilhørende eksempler (konkordanser).

Digital tekstanalyse gir ofte en liste med ordformer som resultat, gjerne sammen med kvantitativ informasjon. Resultatet av slike lister gir grunnlag for aggregering, for eksempel at ordformene *kvinne* og *kvinnen* skal summeres opp som et mål på forekomsten av entall *kvinne*. Selv når ordformene kommer fra forskjellige perioder med forskjellige skrivemåter, skulle de kunne la seg gruppere.[2] For eksempel vil ekstraksjon av kollokasjoner gi en vektet liste av ordformer som statistisk er nær knyttet til hverandre. Om kollokatene stammer fra en tekstsamling som spenner over et tidsrom med språkendringer, vil det oppstå slike kognat-koblinger som ikke fanges inn med dagens digitale ressurser. Om vi tar ordet *kaffe*, for eksempel, og prøver å hente ut de viktigste foranstilte ordene (kollokatene) fra perioden 1900 til 1930, finner vi blant topp

2 Merk at det her er snakk om en annen type gruppering enn den som gis ved lemmatisering eller semantisk likhet.

ti kognatsett som *kop*/*kopp* og *sekker*/*sække*/*sækker*.[3] Her er det tenkelig at man i et gitt forskningsprosjekt ville foretrekke at hele ordlisten moderniseres slik at kollokaten *kopp* innbefatter begge formene *kop* og *kopp*.

2.2. Skjematisk fremstilling av prosessen

Målet vårt er å lage en ressurs som inneholder koblinger mellom den moderne formen av et ord og ordets historiske kognater. Vi viser hvordan koblinger lages mellom enkeltord, det vil si slik de opptrer grafemisk, og hvordan de koblingene på et senere stadium kan benyttes til å utvide leksemer. Skjematisk ser det slik ut for grunnformen *kvinne* og flertallsformen *kvinnene*:

kvinne → {kvinde, qvinde}
kvinnene → {kvinderne, qvinderne}

Slike koblinger legger grunnlaget for å utvide spørringer (såkalt *query expansion*) i søkesammenheng, i tillegg til at de danner grunnlaget for å gruppere former som nevnt ovenfor. Nedenfor ser vi også på hvordan koblingene kan benyttes til å konstruere historiske leksemer.

I en søkesituasjon der man leter etter bøker som tar for seg begrepet *kvinne*, vil man gjerne få treff på alle bøkene som inneholder en eller annen form av ordet *kvinne*. Det vil si *kvinne*, *kvinnen*/*kvinna*, *kvinner* og *kvinnene* i tillegg til alle de historiske kognatene, som for eksempel *qvinde*, *qvinden*, *qvinder* og *qvinderne*, og de tilsvarende formene basert på grunnformen *kvinde*. I en leksikografisk sammenheng vil en slik utvidelse også kunne benyttes til å finne historiske belegg for varianter av ord/leksemer. Det vil si at

3 Analysen ble laget med kollokasjonsappen fra DH-LAB (Nasjonalbibliotekets laboratorium for digital humaniora, se litteraturlisten). Perioden 1900–1930 inneholder de viktige 1907- og 1917-reformene i norsk normering.

man kan oppgi formen *kvinne* som søkeuttrykk og få treff på alle formene av ordet, også de historiske. Hvor mye en søking skal utvides, vil være opp til den som søker, for eksempel om det bare skal lages varianter for en grammatisk form som bestemt form entall.

Ressursen vil derfor kunne anvendes til generelt begrepssøk av brukere uten kjennskap til leksemkategorier og historiske former og også til mer spissede leksikografiske søk av brukere som har oversikt over formene og grammatikken til ordet de er ute etter.

3. Data og metode

I det følgende skal vi presentere datagrunnlaget og metodene vi bruker for å lage en historisk ordliste på basis av Nasjonalbibliotekets samlinger. Vi skal benytte teknikker fra datautvinning og statistisk maskinoversettelse. For å ha et tilstrekkelig datagrunnlag trenger vi et korpus bestående av bøker som foreligger både i en historisk og i en moderne variant. Med slike par av bøker på plass kan vi starte prosessen med å finne ord som korresponderer med hverandre, og benytte den korrespondansen til å konstruere par av kognater.

3.1. Datagrunnlag

Vi skal benytte oss av Nasjonalbibliotekets tekstsamlinger, slik de er eksponert gjennom bibliotekets DH-LAB med tilhørende API (*application programming interface*, norsk: *programmeringsgrensesnitt*), i tillegg til en såkalt spesialbibliografi, nemlig *Nasjonalt autoritetsregister for verk* (Verksregisteret). DH-LAB tilbyr en datamaskinell inngang til bibliotekets samlinger som lar en studere kvantitative aspekter ved samlingene på en programmatisk måte. DH-LAB består av flere komponenter som kan være av interesse for forskere med ulik grad av datateknisk kyndighet.

Formålet med Verksregisteret er å gruppere de forskjellige utgavene av et verk slik at alle utgavene som faller inn under det, skal kunne gjenfinnes og brukes. Verksregisteret ble utviklet i prosjektet SHARE-VDE, hvor flere amerikanske universitetsbiblioteker samt nasjonalbibliotekene i Finland og Norge deltar.

Ved hjelp av Verksregisteret finner vi altså alle utgaver av et verk, men for å kunne lage en historisk ordliste trenger vi i tillegg informasjon om utgavens språkform. En ny utgave av et verk er ikke nødvendigvis moderne i språkformen. Det finnes for eksempel både faksimiler og diplomatiske utgaver som gjengir den opprinnelige formen. Derfor må vi dele inn utgavene fra Verksregisteret i bolker basert både på utgivelsestidspunkt og på språkform. Dette skal vi se nærmere på i neste kapittel.

3.2. Parallelltekst

En viktig teknikk i det automatiserte arbeidet med å lage en historisk ordliste er såkalt *alignering*, det vil si å parallellstille tekster eller deler av tekster. I datalingvistikk er det vanlig å bruke termen *bitext* om én og samme tekst som er oversatt til ett eller flere språk (Tiedemann 2011), eller om flere versjoner av samme grunnlagstekst. Termen *parallel text* (norsk: *parallelltekst*) er også vanlig i samme betydning. På samme måte som om vi skulle alignert tekster på to ulike språk, velger vi i denne sammenhengen å alignere varianter av den samme teksten fra ulike språkstadier. Her er vi til dels begrenset av Verksregisteret og hva det inneholder. Det betyr at tilfanget av verker vi ser på, i prinsippet kan være mindre enn det faktiske antallet verker, men vi legger til grunn av varianter innad i et verk er komplett for alle praktiske formål.

For å identifisere tekster som kan representere ulike språkstadier, ser vi på frekvensen av de ulike formene som et høyfrekvent ord opptrer i. Høyfrekvente ord er med i de fleste tekster, og skrivemåten avslører tidsepoken. Hvis for eksempel ordformen *paa*,

og ikke *på*, går igjen i en tekst, gir det et hint om at teksten tilhører tiden før rettskrivingsreformen fra 1917. Et sett med slike høyfrekvente ord ble benyttet til å avgjøre om teksten er historisk eller moderne.[4] Fra et tilfeldig utvalg av fem hundre tekster fra de to periodene fant vi følgende (illustrert i tabell 1) blant de ordene som 1) har høy frekvens og 2) har en tilstrekkelig skillende effekt. Det siste vil si at formene har vesentlig forskjellig frekvensfordeling i de to periodene.

Ordform	**Historiske tekster**	**Moderne tekster**	**Kognat**
lese	0	1637	*læse*
fat	2712	0	*fatt*, men også *fat* substantiv
vide	5146	0	*vite*, men også *vid* adjektiv
paa	216 235	10 071	*på*
inn	0	27521	*ind*
ind	22 370	0	*inn*
på	9143	244 801	*paa*
lægge	2527	0	*legge*

Tabell 1: Et lite utvalg av ordformer med frekvenser. Frekvensene er for de utvalgte ordformene i henholdsvis historiske og moderne tekster, sammen med en indikasjon av ordformenes historiske kognater.

Ved å sammenligne frekvensen av moderne og historiske former i en variant av et verk med frekvensen av de samme formene i tekstutvalget kan vi avgjøre hvorvidt varianten tilhører et tidligere språkstadium eller ikke. Hver tekst får en vektet score[5] basert på hvor mange «moderne» ordformer den har, og hvor mange som tilhører den «historiske» perioden.

4 Her er det gjort et utvalg av historiske tekster i perioden 1700–1917, mens de moderne er hentet fra perioden 1920–2010.

5 For eksempel kan man bare telle opp forventet antall av de forskjellige ordformene.

En mulig utfordring med metoden er eventuelle forskjeller i måter å modernisere enkelte forfatterskap på. Man kunne se for seg at det er ulike tradisjoner for modernisering av for eksempel Henrik Ibsen og Bjørnstjerne Bjørnson. Dette problemet skal vi ikke forfølge videre her.

En annen utfordring er anakronismer: Selv i moderniserte utgaver vil visse former holde seg lenger, for eksempel riksmålsformer av hyppige ord, som *nu* og *efter*, eller pronominaladverb som *dertil* og *hvorpå*, som må kunne betraktes som gammeldagse i moderne bokmål. For å omgå effekten av enkeltord tar vi med en forholdsvis stor liste[6] med skillende ord i vurderingen av om en tekst er moderne eller historisk.

3.3. Setningsalignering

Når dokumentene er alignert, må setningene og ordene innad i dokumentene aligneres. Dette er to separate automatiserte prosesser der setningene sammenstilles først, så ordene.

En gjengs og mye brukt setningsalignerer er beskrevet i Varga et al. (2005): *hunalign*. Denne tar utgangspunkt i en maskinell ord-for-ord-oversettelse mellom en kildetekst og en måltekst. I vårt arbeid benytter vi nyere modeller for alignering, som sBERT (Reimers & Gurevych 2020). Disse modellene gjør ikke noen oversettelse *per se*, men omdanner tekst til maskinlesbare trekkstrukturer, og sammenligner de ulike trekkene for å finne de setningene som ligner mest på hverandre.

Setningsaligneringen utføres så utgave for utgave ved at utgavene i den historiske delen av korpuset sammenlignes med utgavene i den moderne. Det genereres altså kombinasjoner av gammel og ny tekst hvor setningene fra hvert par sidestilles. Ligger sannsynligheten for at setningene i setningsparet faktisk er like, under et visst nivå, forkastes setningsparet. Ikke alle setninger er paral-

6 I testinger har vi brukt en liste med 1500 ord.

lelle (det kan være tilføyelser eller andre endringer, eller det kan være forskjeller i tekstkvalitet forbundet med digitaliseringen av teksten, for eksempel automatisk bokstavgjenkjenning).

3.4. Ordalignering

Når setningene er alignert, kan vi så alignere på ordnivå, og det er slik vi får den tidligere illustrerte koblingen fra moderne til historiske ordformer:

kvinne → qvinde

For denne typen alignering har vi benyttet verktøyet SimAlign, som til dels bygger på samme teknologi som sBERT. Språkmodellen vi har brukt i SimAlign, er den som beskrives i Kummervold et al. (2021), som er trent på et stort materiale fra Nasjonalbibliotekets digitale samlinger. Giza++, beskrevet i Och & Ney (2003), er et alternativ til SimAlign, men Sabet et al. (2020) viser at sistnevnte er mer treffsikker.

4. Historisk ordliste som leksikografisk ressurs

Resultatet av aligneringen er en ordliste som kan brukes til forskjellige formål. Ett er å gi en beskrivelse og en kategorisering av historiske former, et annet er å oversette eldre tekster automatisk.

Om man i en moderne ordbok prøver å finne informasjon om ordformene *sygdom* eller *kvinder*, får man ikke noen treff,[7] men man får treff på *sykdom* og *kvinner*. Når historiske tekster er digitalt tilgjengelige, er det ønskelig å ha en fyldig leksikografisk beskrivelse av dem. I kapittel 4.1 ser vi på hvordan kognatparene kan brukes til å konstruere fulle historiske leksemer. I kapittel 4.2 tar vi

7 De aktuelle ordene er sjekket med NAOB og *Bokmålsordboka*.

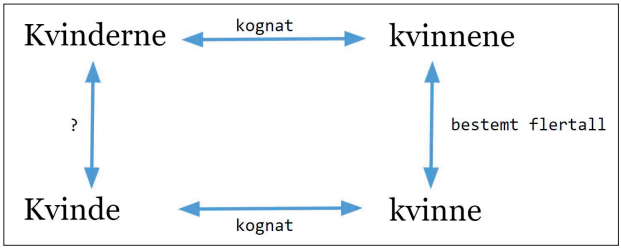
for oss muligheter for å modernisere tekster ut fra det perspektivet at modernisering er en form for oversettelse. Kapittel 4.3 tar for seg et problem som oppstår i forbindelse med automatisk modernisering: tvetydighet ved funksjonell splitting.

4.1. Fra ordformer til leksemer

Når koblingen mellom historiske og moderne ordformer er etablert, kan vi prøve å overføre informasjon fra den moderne beskrivelsen til den historiske. Så langt har vi kun sett på kobling mellom ordformer, som at *kvinne* korresponderer med *kvinde*, og *kvinner* med *kvinder*. Men vi kan også stille spørsmålet om det er mulig å koble relasjonene mellom ordformene sammen til et helt leksem, slik at alle de historiske formene blir gruppert sammen med de respektive moderne formene til ett leksem.

I figur 3 er det illustrert to relasjonstyper, én som viser at en ordform er kognat av en annen (heretter kognatrelasjonen), og én som angir ordformenes plass i leksemet (heretter leksemrelasjonen). Figuren viser med andre ord at *kvinne* og *kvinnene* er kognater av henholdsvis *Kvinde* og *Kvinderne* og samtidig at *kvinnene* er bestemt flertall av *kvinne*. Den ukjente relasjonen – som ikke er direkte kodet i tilgjengelige ordbøker – er den mellom *Kvinde* og *Kvinderne*.

Leksemrelasjonen og kognatrelasjonen antas å være kommutative. Det betyr at om vi starter i hjørnet som inneholder ordformen *Kvinde*, og så følger pilene for kognaten og deretter pilene for leksemrelasjonen bestemt flertall, vil vi få samme resultat som om vi først velger leksemrelasjonen bestemt flertall og så kognaten. Vi får *kvinnene* som resultat i begge tilfeller.



Figur 3: Kommutativt diagram over relasjonene *kognat* og *flertallsform*.

I praksis betyr det at informasjon om de historisk relaterte kognatene kan berikes med informasjon fra beskrivelser av det moderne språket. Resultatet er at de historiske kognatparene kan danne et utgangspunkt for å konstruere en digital historisk ordbok.

4.2. Automatisk modernisering

Både hver for seg og i kombinasjon kan ordaligneringen og setningsaligneringen komme til nytte i automatisk modernisering av tekst. Resultatet av ordaligneringen er en ordliste (som beskrevet over) som kobler to ordformer sammen. Formålet med modernisering kan være å tilgjengeliggjøre eldre tekster for de som ikke er kjent med historiske ordformer, eller det kan være forskjellige formål innen tekstanalyse – for eksempel å koble tekstene til moderne ordbøker.

En automatisk modernisering kan med de beskrevne ressursene ta to former: 1) Den konstruerte ordlisten kan fungere som grunnlag for substitusjon, det vil si at de historiske formene byttes ut med de moderne. 2) Ved hjelp av det setningsalignerte materialet (det vil si listen med par av setninger) kan vi trene maskinoversettere (for eksempel slike som er beskrevet i Raffel et al. 2020). Alternativ 2 vil kunne gi mer idiomatisk moderne norsk enn det som oppnås med alternativ 1.

4.3. Funksjonell splitting

Alignering vil gi flere koblinger mellom ord enn de som forbinder kognater. Disse kan oppstå for eksempel som følge av OCR-feil[8] (feil ved automatisk tekstgjenkjenning), som i *kjcerlighed*/*kjærlighet*, ved at ordet er oversatt, som i *kvinde*/*woman*, eller ved at ordet er byttet ut med en semantisk ekvivalent, som i *glad*/*lykkelig*.

Særlig utfordrende i den sammenhengen er tilfeller der et ord får en oppsplitting av funksjoner. Den historiske formen beholder et bruksområde i det moderne språkstadiet, samtidig som en nyere form erstatter den historiske innenfor andre områder. To eksempler er eldre *at* vs. nyere *at*, *aa* og *å*, og eldre *der* vs. nyere *der* og *det*.

Oppsplitting i funksjon er en utfordring i prosessen med å lage en modernisert tekst på grunnlag av en eldre. Om vi benytter alternativ 1 i forrige kapittel for automatisk modernisering av tekst (der eldre ordformer blir byttet ut med korresponderende moderne former), vil *at* byttes ut med *å* betingelsesløst, også der det ikke er riktig å gjøre det. I setningene *han prøvede at gaa* og *han troede at han var syg* vil utskifting av *at* med *å* i den første gi riktig resultat, mens i den siste vil *å* for *at* resultere i en ugrammatisk (og meningsløs) setning. For de andre ordene kan man bare velge en moderne form uten å se på kontekst. Samme situasjon gjelder for *der* vs. *det*, for eksempel i *der staar en kat i haven* vs. *han stod der i haven*. I den første setningen går det fint å bytte ut *der* med *det*, men i den siste setningen vil det resultere i en ugrammatisk setning.

5. Oppsummering

Vi har beskrevet en metode for å automatisk utvide eksisterende ordbøker til å omfatte historisk materiale, både på ordformnivå

8 OCR = Optical Character Recognition (optisk tegngjenkjenning).

og på leksemnivå. Vårt fokus har vært på leksikalske ressurser for digital prosessering av tekst, og vi har sett både på hvordan slike ressurser kan benyttes i beskrivelsen av språket selv (det rent leksikografiske), og på hvordan slike ressurser kan forbedre søk i litteratur og ellers gjøre det mulig å sammenligne tekster fra ulike tidsperioder i digital tekstanalyse.

Litteratur

Digitale ressurser

Bokmålsordboka. Språkrådet og Universitetet i Bergen. <ordbokene.no> (juli 2022).

DH-LAB = Nasjonalbibliotekets laboratorium for digital humaniora. <nb.no/dh-lab> (juli 2022).

NAOB = *Det Norske Akademis ordbok*. <naob.no> (juli 2022).

Verksregisteret = Nasjonalt autoritetsregister for verk. <bibliotekutvikling.no/kunnskapsorganisering/kunnskapsorganisering/nasjonalt-autoritetsregister-for-verk/> (juli 2022).

Annen litteratur

Birkenes, Magnus Breder, Lars G. Johnsen, Arne M. Lindstad & Johanne Ostad (2015): From digital library to n-grams: NB N-gram. I: *Proceedings of the 20th Nordic Conference of Computational Linguistics*. Linköping: Linköping University Electronic Press, 293–295.

Kummervold, Per E., Javier De la Rosa, Freddy Wetjen & Svein Arne Brygfjeld (2021): Operationalizing a National Digital Library: The Case for a Norwegian Transformer Model. I: *Proceedings of the 23rd Nordic Conference on Computational*

Linguistics (NoDaLiDa), 20–29. <aclanthology.org/2021.nodalida-main.3> (september 2022).

Och, Franz Josef & Hermann Ney (2003): A Systematic Comparison of Various Statistical Alignment Models. I: *Computational Linguistics* 29(1), 19–51.

Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li & Peter J. Liu (2020): Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. I: *Journal of Machine Learning Research* 21, 1–67.

Reimers, Nils & Irina Gurevych (2020): Making monolingual sentence embeddings multilingual using knowledge distillation. I: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. <arxiv.org/abs/2004.09813> (september 2022).

Sabet, Masoud Jalili, Philipp Dufter, François Yvon & Hinrich Schütze (2020): SimAlign: High Quality Word Alignments without Parallel Training Data using Static and Contextualized Embeddings. I: *Findings of the Association for Computational Linguistics: EMNLP* 2020, 1627–1643. <dx.doi.org/10.18653/v1/2020.findings-emnlp.147> (september 2022).

Tiedemann, Jörg (2011): Bitext alignment. I: *Synthesis Lectures on Human Language Technologies* 4(2), 1–165.

Varga, Daniel, Péter Halácsy, András Kornai, Viktor Nagy, László Németh & Viktor Trón (2005): Parallel corpora for medium density languages. I: *Proceedings of the RANLP* 2005, 590–596.

Magnus Breder Birkenes
Forskningsbibliotekar
dr.phil.
Nasjonalbiblioteket
Henrik Ibsens gt. 110
NO-0255 Oslo
magnus.birkenes@nb.no

Lars G. Bagøien Johnsen
Forskningsbibliotekar
dr.art.
Nasjonalbiblioteket
Henrik Ibsens gt. 110
NO-0255 Oslo
lars.johnsen@nb.no

Andre Kåsen
Forskningsbibliotekar
M.Sc.
Nasjonalbiblioteket
Henrik Ibsens gt. 110
NO-0255 Oslo
andre.kasen@nb.no