STUDIA UBB PHILOLOGIA, LXV, 3, 2020, p. 107 - 124
(RECOMMENDED CITATION)
DOI:10.24193/subbphilo.2020.3.08

AUTOMATISERT KLASSIFIKASJON AV NORSKE MÅLFORMER VHA. DATAUTVINNING AV UANNOTERT TEKST

FARTEIN TH. ØVERLAND[1]

ABSTRACT. ***Automated Classification of Variants of Norwegian by Means of Text Mining of Unannotated Text.*** This article presents a model for automatically classifying different variants of modern Norwegian Language (*bokmål* and *nynorsk* ranging from 1930 to 2011) by means of data mining unannotated text. The model is built in the Orange visual programming interface, and is based on a modification of an example model presented by the project which had the original purpose of semantical classification of fairy tale types in the Aarne-Thompson-Uther Index. The core modules of the model are Bag-of-Words and Logistic Regression. The model is trained with four different translations of the Gospel of John, and cross validated with various random texts. The model is proven to be very sound for classification of Norwegian language variation, and yields correct classification in 100% of the realistic tests.

Keywords: *Language Variation, Text mining, Orange Data Mining, Text Clustering, Text Classification, Bag-of-Words, Logistic Regression, Predictive Model, Norwegian Language, Nynorsk, Bokmål*

REZUMAT. ***Clasificare automatizată a diferitelor variante de norvegiană utilizând extragerea digitalizată a textelor neanotate.*** Acest articol prezintă un model pentru clasificarea automată a diferitelor variante ale limbii norvegiene moderne (*bokmål* şi *nynorsk*, între 1930 şi 2011) cu ajutorul extragerii automatizate a textului neanotat. Modelul este construit în interfaţa de programare vizuală Orange şi se bazează pe modificarea unui model-exemplu prezentat de proiect, care a avut ca scop iniţial clasificarea semantică a tipurilor de poveşti din indexul Aarne-Thompson-Uther. Modulele de bază ale modelului sunt Bag-of-Words şi Regresie logistică. Modelul este axat pe patru traduceri diferite ale Evangheliei lui Ioan şi este validat de alegerea aleatorie a fragmentelor. Modelul s-a dovedit a fi foarte solid pentru clasificarea variaţiei limbii norvegiene şi obţine o clasificare corectă în 100% din testări.

Cuvinte cheie: *variaţia limbii, extragerea digitalizată, interfaţa de programare Orange, clasificarea textelor, Bag-of-Words, regresie logistică, model predictibil, limbă norvegiană, nynorsk, bokmål*

[1] Visiting Norwegian Lecturer at the Babeş-Bolyai University, Department of Scandinavian Languages and Literatures; has published articles on corpus linguistics, Old Norse language, skaldic poetry and deep learning. E-mail: farthegn@gmail.com

1. Innleiing

Føremålet med denne studien er å utvikla og testa ein sjølvverkande modell for klassifikasjon av ulike variantar av norsk språk. Modellen vil verta bygd opp og køyrt i programvaren Orange; ei verktøykasse for visuell programmering av datautvinning (eng. *data mining*) utvikla ved Universitetet i Ljubljana (jf. Demar J et al. 2013 og nettsida til prosjektet[2]) og stør seg i stor mon på dette prosjektets retningslinjer for klassifikasjon av tekst med visse brigde for å tilpassa modellen til mitt føremål. Kjernen av klassifikasjonsmetoden er pose-med-ord-modellen (etter eng. *Bag-of-Words*, mi omsetjing) (jf. Zellig 1954) og logistisk regresjon. Denne modellen fungerer med tekst utan metadata og både korpuset som vil verta nytta for å trena og testa modellen er såleis rein, uannotert tekst. Norsk språk etter unionstida skil seg frå mange andre språk ved å ha stor variasjon, mest openbert med omsyn til dei offisielle skriftspråka, men også pga. hyppige språkreformer og stor valfridom mellom ulike former innanfor kvar av målformene. Artikkelen vil også kort drøfta kva fylgjer denne variasjonen får for bruk av pose-med-ord-modellen på norsk jamført med språk med mindre variasjon i rettskrivinga, som t.d. moderne engelsk. Denne drøftinga knyter seg til hovudproblemstillinga fordi føremålet til klassifikasjonsmodelen min modell byggjer på er klassifikasjon av genrar, ikkje språkvariasjon. Det vil også verta gjeve ei vurdering av bruken av Orange frå eit brukarperspektiv, og skisser til vidareutvikling av metoden som vert presentert her for andre forskingsmål innan nordiskfaget.

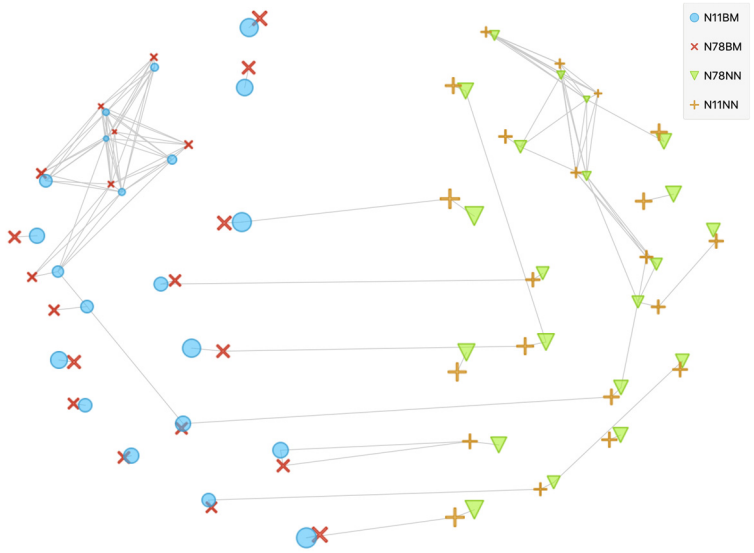
2. Val av korpus til trening av modellen

For denne studien er det naudsynt med eit korpus med tekst som finst i parallelle versjonar med ulike variantar av bokmål og nynorsk med tilstrekkeleg lengd til å byggja ein solid modell og med liknande tekstar til å testa modellen med. Alternativ kunne vore offisielle dokument som offentlege vedtekter og lovtekstar eller lærebøker, men eg ynskte helst eit korpus med narrative trekk (som på einskildordplanet ovrar seg som høg frekvens av m.a. personlege pronomen og verb for handlingar, t.d. *gå og* talehandlingar, t.d. *seia*). Eit godt alternativ har vist seg å vera norske bibelomsetjingar, særleg sidan Bibelen har eit velutvikla system for tekstinndeling og kryssreferansar og ulike delar med liknande innhald. Ulike omsetjingar av Johannesevangeliet (som utgjer 21 kapittel og om lag 6000 ord) har vorte nytta for å trena modellen. Av praktiske omsyn har eg nytta dei omsetjingane som er tilgjengelege digitalt frå Det Norske Bibelselskap[3]. Eg tolkar deira *Retningslinjer for bruk av Bibelselskapets*

[2] https://orange.biolab.si (Sett den 29. juni 2020)
[3] På http://www.bibel.no/Nettbibelen (Sett den 29. juni 2020)

bibeloversettelser[4] som at denne bruksmåten av dataa er gangbar med tanke på opphavsrett. Dei aktuelle utgåvene er bokmålsutgåvene frå 1930, 1978/85 og 2011 (heretter: N1930BM, N78BM og N11BM) og nynorskutgåvene frå 1938 (også kjent som Indrebøbiblen) 1978/85 og 2011 (heretter: N38NN, N78NN og N11NN). Det er ikkje plass her til å gå i djupna når det gjeld utgjevingshistoria til desse utgåvene og dei tidlegare utgåvene dei byggjer på, her viser eg vil Bibelselskapets eigne oversyn *Oversettelser 1814-1938*[5] og *De nyeste bibeloversettelsene til norsk*[6].

Når eg visualiserte modellen si klassifisering av utgåvene, viste det seg, som ein kunne venta, at N78NN/N11NN og N78BM/N11BM ligg så nær kvarandre språkleg at dei ikkje kan brukast til å skilje ulike variantar av målformene. Ein kan nytta multidimensjonal skalering for å framstilla desse forholda visuelt.



Figur 1. Fråstanden mellom N78NN, N11NN, N78BM og N11BM visualisert med multidimensjonal skalering (skjermdump frå Orange)

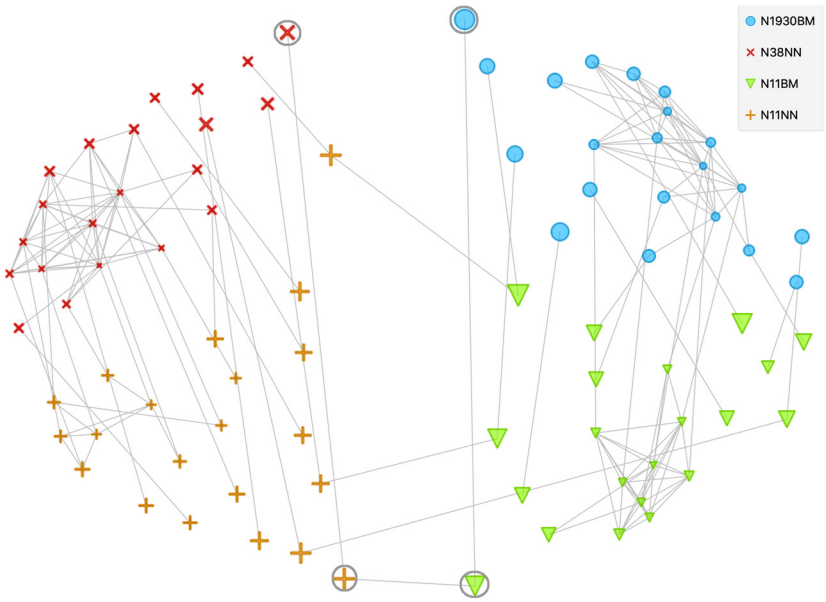
Punkta i figuren utgjer dei 21 kapitla i Joh og linene liknande par. Som ein kan sjå er skilnadene i språkføringa mellom 1978/85- og 2011-utgåvene så liten at identiske ordformer til ord frå bestemte kapittel ligg nærmare kvarandre mellom utgåvene enn ordformene til andre kapittel i same utgåve. Jamføring av språkføringa i N78NN og N11NN stadfester dette; N78NN bruker fleire tradisjonelle

[4] http://www.bibel.no/Nettbibelen/Opphavsrett_2 (Sett den 29. juni 2020)

[5] https://www.bibel.no/OversettelseSprakLitteratur/Bibeloversettelser/Oversettelser-Norge/Oversettelser1814-1938 (Sett den 29. juni 2020)

[6] https://www.bibel.no/OversettelseSprakLitteratur/Bibeloversettelser/Oversettelser-Norge/Etter1938 (Sett den 29. juni 2020)

former som t.d. *fylgja*[7] (Joh 3: 21), *ljos* og (Joh 12: 35) og *hjå* (Joh 12: 8) medan N11NN nyttar *følgja, lys* og *hos*. Større variasjon kunne ein heller ikkje venta seg med tanke på normeringshistoria i den perioden det er snakk om. Bortsett frå det er skilnaden mellom utgåvene fyrst og fremst omsetjingsmetodane; 1978/85-utgåvene er i stor grad konkordante (grunntekstnære) og 2011-utgåvene idiomatiske (meiningsnære) (jf. *Ulike måter å oversette Bibelen på*[8]). For å maksimalisera den språkleg kontrasten til dei eldre utgåvene har eg difor vald å ikkje bruka 1978/85-utgåvene. Då står N1930BM, N38NN, N11BM og N11NN att, og desse dannar eit korpus med fin symmetri mellom eldre riksmål og nynorsk og yngre bokmål og nynorsk. Det er fleire variantar av norsk som ikkje er inkludert i modellen (1800-tals riksmål, klassisk landsmål, midlandsmål og moderne nynorsk og bokmål med samnorskformer), men det er god grunn til å rekna med at dei ville fungert på same måte i modellen som dei variantane som finst i korpuset, og, som me skal sjå i kap. 5.3 nedanfor, kan modellen også estimera nærskylde målformer som han ikkje har vorte trent i.



Figur 2. Fråstanden mellom N38NN, N11NN, N1930BM og N11BM visualisert med multidimensjonal skalering (skjermdump frå Orange, med nokre punkt manuelt markert av forfattaren)

[7] Dette og alle fylgjande døme på einskilde ord er ført opp i lemma-form, ikkje den bøyingsforma dei har i teksten, men mindre noko anna er oppgjeve.

[8] https://www.bibel.no/OversettelseSprakLitteratur/Bibeloversettelser/Oversettelsesmetoder (Sett den 29. juni 2020)

Språket i N1930BM kan ein karakterisera som eldre riksmål med ein del danske former som t.d. *I* (Joh 12: 8), *op* (Joh 12: 10) *mig* (Joh 12: 14), *øie* (Joh 12, 37) og *nogen* (Joh 12: 47) der N11BM har *dere*, *opp, meg* øye *og noen*. Språkføringa i N11BM gjer seg elles nytte av moderate og riksmålsnære variantar innanfor offisiell bokmålsnormering. Sjølv om N38NN kom ut same året som den store rettskrivingsforma i 1938, fylgjer omsetjinga reforma frå 1917 med i-mål med skilnad mellom sterke og linne femininum, t.d. bund. f. sg. *grav**i*** (Joh 20: 1), men *kon**a*** (Joh 8: 4) og bund. f. pl. *synagog**one*** (Joh 16: 2), men pil**er** (1. Sam 20: 20) og tradisjonelle former som t.d. *fyrr* (Joh 8: 54) *skjota* (1. Mos 40: 10) og *ganga* (2. Mos 3: 3) osb. der N11NN har *grav**a,*** (*kon**a***), *synagog**ene,*** (*pil**er***), *før, skyta og gå*.

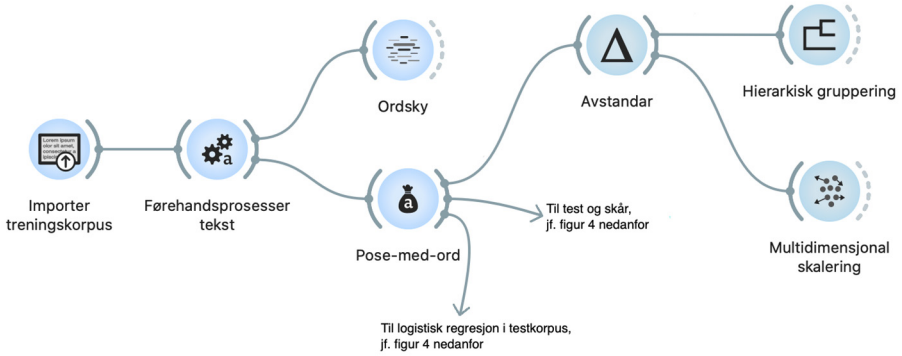
Ein kan leggja merke til at MDS-visualiseringa også viser semantiske mønster som kan vera av interesse for semantisk analyse. I fig. 2 kap. 2 i dei fire utgåvene med gråe sirklar, og ein kan sjå at dei har stor avstand (dvs. høg frekvens avvikande ordformer) frå andre kapittel i same utgåve. Dette mønsteret kan visa oss at kapittelet er tematisk avvikande frå resten av evangeliet, noko som nærlesing vil stadfesta (adjektivet *drukken* (Joh 2: 10) opptrer t.d. berre i framstillinga av bryllaupet i Kana i heile Joh). Om ein bruker eit korpus med einsarta språkform, vil slike mønster verta tydlegare.

3. Strukturen til treningsdelen av modellen

I tillegg til skriftleg dokumentasjon, har Orange-prosjektet dokumentert ei innføring i bruk av programvaren i videoformat. Dette gjeld også tilleggspakken (eng. *add-on*) for datautvinning av tekst (eng. *text mining*) (Orange3-Text[9]) Mi røynsle har vore at dette fungerer utmerkt som eit didaktisk hjelpemiddel, men er ikkje like lett å visa til i skriftleg form. Men i og med at mitt prosjekt byggjer direkte på ein modifikasjon av dømeprosjekta i instruksjonsvideoane, vil det vera lettare å fylgja og etterprøva framstellinga med å visa til desse. Sidan audiovisuelle læremiddel ikkje har sidetal eller laupande tekst, vil sitat vera tufta på transkripsjon av undertekster og tidsstempel. Føremålet til dømeprosjektet eg byggjer på – presentert *Getting Started with Orange 18: Text Classification*[10] – er å trena ein model med eit korpus av tekstar frå *Brørne Grimms eventyr* som er klassifiserte etter Aarne-Thompson klassifikasjonssystemet som *dyreeventyr* eller *eigenlege eventyr* (jf. Antti Aarne og Stith Thompson 1961) og deretter føreseia om andre uklassifiserte eventyr høyrer til i den eine eller andre kategorien.

[9] Dokumentert på https://orange3-text.readthedocs.io/en/latest/ (Sett den 29. juni 2020)
[10] https://youtu.be/zO_zwKZCULo (Sett den 29. juni 2020)

Sidan modellen er bygd opp med visuell programmering, vil det vera lettare å fylgja den vidare framstellinga ved fyrst å sjå på den grafiske visninga av modellen og deretter omtala funksjonen og innstillingane til kvar einskild modul.



Figur 3.: Strukturen til treningsdelen av modellen (skjermdump frå Orange med omsetjing til norske namn på modulane (eng. *widgets*) og digital biletbehandling av forfattaren).

3.1. Import av treningskorpus

Bruken av modulen *importer treningskorpus* (orig. modulnamn *Import Documents*[11]) vert forklart i instruksjonsvideoen *Getting Started with Orange 19: How to Import Text Documents*[12]. Grunnfunksjonen er å importera filer med rein tekst (.txt) for å byggja opp eigne korpus (andre val for i Orange er spesielle filformat for tekstkorpus eller direkte lesing av diverse nettressursar). Filene vert klassifiserte etter struktur til mappene dei ligg i utan andre metadata.

3.2. Førehandsprosessering av tekst

Bruken av modulen *førehandprosesser tekst* (orig. modulnamn *preprocess text*[13]) er forklart i instruksjonsvideoen *Getting Started with Orange 16: Text Preprocessing*[14]. Grunnfunksjonen til modulen er å fjerna teiknsetjing, tal osb. for å klårgjera teksten til vidare maskinlæring med ord som einskilde datapunkt, typisk vha. pose-med-ord-modellen. I datautvinning av tekst der siktemålet er

[11] Jf. dokumentasjon på https://orange.biolab.si/widget-catalog/text-mining/importdocuments/ (Sett den 29. juni 2020)
[12] https://youtu.be/faIqvWxFGRc (Sett den 29. juni 2020)
[13] Jf. dokumentasjon på https://orange.biolab.si/widget-catalog/text-mining/preprocesstext/ (Sett den 29. juni 2020)
[14] https://youtu.be/V70UwJZWkZ8 (Sett den 29. juni 2020)

å finna semantiske mønster, vil ein ofte også filtrera ut ord med den tekniske termen *stoppord* (eng. *stop words*), dvs. ord som ikkje ber mykje semantisk innhald (subjunksjonar, konjunksjonar, preposisjonar osb.). For vårt siktemål er nettopp stopporda av ekstra stor verdi, for sidan dei opptrer så frekvent i alle former for tekstar vil ein kunna byggja opp ein solid modell utan å trena han på eit enormt stort korpus. Den sentrale rolla til stopporda i måten modellen predikerer målform på er påvist i nomogrammet i 4.2 nedanfor.

3.3. Pose-med-ord og avstandar

I instruksjonsvideoen *Getting Started with Orange 17: Text Clustering*[15] vert funksjonen til modulen *pose-med-ord* (orig. modulnamn *Bag-of-Words*[16]) forklart slik: "*For machine learning we need to transform text into numerical representation, and a simple way to do it is to count how many times each word appears in the text. This approach is called bag-of-words."* (ibid. [0:26-:0:40]). Modulen lagar altså ein matrise over ordformene i korpuset der kvar ordform dannar ei kolonne med ein vektor som tel kor mange gonger ordforma ovrar seg.

Som nemnd ovanfor, vart analyse (med ein modell konstruert av logistisk regresjon, jf. 4.2 nedanfor) av ordform-matrisen brukt for å klassifisera eventyrtypar i Oranges dømeprosjekt, som min modell byggjer på. Min fyrste tanke når eg såg at dette var mogleg på engelsk, var at liknande ikkje ville fungera like godt for ei korpus med ueinsarta norske tekstar, og enno dårlegare for tekstar med større variasjon, t.d. diplomatarisk transkripsjon av norrøne eller mellomnorske handskrifter eller fonetisk transkripsjon av talte norske dialektar. Som eit *argumentum ad absurdum* kan ein tenkja seg korleis pose-med-ord ville handsama eit korpus bygd opp av lister av variantformer frå ordbøker. Som døme kunne ein tenkja seg lista over variantar av det ubest. pron. *nǫkkur* 'noen' i Ordbog over det norrøne prosasprog[17] – som har identifisert over 600 distinkte former! Pose-med-ord ville vekta kvar av desse med 1, og det klart at ein treng enormt store korpus eller ein modell som tek omsyn til syllabisk struktur for å handsama slik variasjon. Eit meir relevant døme frå moderne norsk kunne vera at pose-med-ord vil klassifisera dei sju ordformene *gren, grenen, grener, grenene, grein, greina og greini* som distinkte oppslag og ikkje ha noko grunnlag for å kopla dei til eitt lemma. Som jamføring vil eit engelsk korpus berre ha to distinkte ordformer for ordet *branch* 'grein'*: branch* og

[15] https://youtu.be/rH_vQxQL6oM (Sett den 29. juni 2020)

[16] Jf. dokumenasjon på https://orange.biolab.si/widget-catalog/text-mining/bagofwords-widget/ (Sett den 29. juni 2020)

[17] I digital utgåve på https://onp.ku.dk/onp/onp.php?o57831# (>Comp., Gloss., Litt &c.) (Sett den 29. juni 2020)

branches. Men for denne studien er det nettopp å bruka modellen til å identifisera variasjonen som interesserer oss.

Som dokumentasjonen til modulen forklarer bereknar *avstandar* (orig. modulnamn *Distances*[18]) avstanden mellom rader og kolonnar i eit datasett og sender ut ein matrise over avstandane mellom dei. For at eit menneske skal kunna tolka denne matrisen, kan ein visualisera han på fleire vis. I *Text Clustering*-videoen nyttar dei hierarkisk gruppering (orig. modulnamn *Hierarchical Clustering*[19]) som fungerer utmerkt for semantiske grupperingar. Denne visualiseringsmetoden viste seg å fungera dårleg for mitt korpus pga. av likskapen mellom ordforrådet i dei ulike kapitla på tvers av utgåvene. Ein meir formålstenleg visualiseringsmetode viste seg å vera *multidimensjonal skalering* (orig. modulnamn *MDS*[20]) som framstiller datapunkta – i vårt tilfelle kapitla – som punkter på ein to-dimensjonal flate, og knyter liknande par saman med liner (jf. fig. 1 og 2 i kap. 2 ovanfor).

4. Testdelen av modellen

Som det snart vil verta vist, kan ein bruka logistisk regresjon som utgangspunkt for ein modell for å predikera kva for ei målform dei ulike datapunkta – altså kapitla – høyrer til basert på utdata frå pose-med-ord-modulen i treningsdelen av modellen som vart drøfta i kap. 3 ovanfor. Før me ser på det, vil eg kort drøfta alternative modellar som kunne vorte nytta til same føremål og deira føremon og ulemper jamført med modellen som vert presentert i denne studien.

4.1. Alternative modellar

Den mest openberre alternative metoden er manuell klassifisering. Ein person med kjennskap til norsk rettskrivingshistorie ville utan nærlesing kunna identifisera målforma i dei fire ulike bibelutgåvene (jf. kap. 2 ovanfor) som har vorte nytta som korpus for å trena denne modellen (jf. kap. 3 ovanfor) med å sjå på eit avsnitt i fugleperspektiv. Eg vil nedanfor (i 4.1) argumentera for at tankeprosessen personen då går gjennom svarer til modellen den logistiske regresjonen i modellen denne studien genererer (heretter omtalt som LR-modellen). Prosessen baserer seg på eit hierarki av minimale par av frekvente

[18] Jf. dokumentasjon på https://orange.biolab.si/widget-catalog/unsupervised/distances/ (Sett den 29. juni 2020)

[19] Jf. dokumentasjon på https://orange.biolab.si/widget-catalog/unsupervised/hierarchicalclustering/ (Sett den 29. juni 2020)

[20] Jf. dokumentasjon på https://orange.biolab.si/widget-catalog/unsupervised/mds/ (Sett den 29. juni 2020)

ord- og bøyingsformer som skil dei ulike variantane frå kvarandre. T.d. vil det minimale paret *ikke*/*ikkje* vera nok til å avgjera om ein tekst er skriven på bokmål eller nynorsk, *efter*/*etter* om han er skriven på tradisjonelt riksmål eller moderne bokmål og *um/om* om han er skrive på tradisjonelt landsmål/nynorsk eller offisiell nynorsk etter 1938-reformen. Ulempa med denne metoden er sjølvsagt at det vil innebera mykje arbeid å klassifisera store datamengder.

Ein annan modell kunne vore manuell programmering av logiske reglar etter same logikk som er skissert i avsnittet ovanfor. I pseudo-kode kunne ein tenkja seg dette som reglar som *VISS ordform = "ikkje": målform = nynorsk, ELLES målform = bokmål* og så bortetter. Ulempa med denne metoden er igjen arbeidsmengda. For å kunna fungera på kortare tekstutdrag vil modellen trenga mange reglar, og der ein med manuell klassifisering vil kjenna att målforma utan å måtta tenkja medvite på kva slags logikk ein bruker for å resonnera, vil ein for denne metoden tenkja gjennom kva for nokre minimale par ein reknar med er mest frekvente.

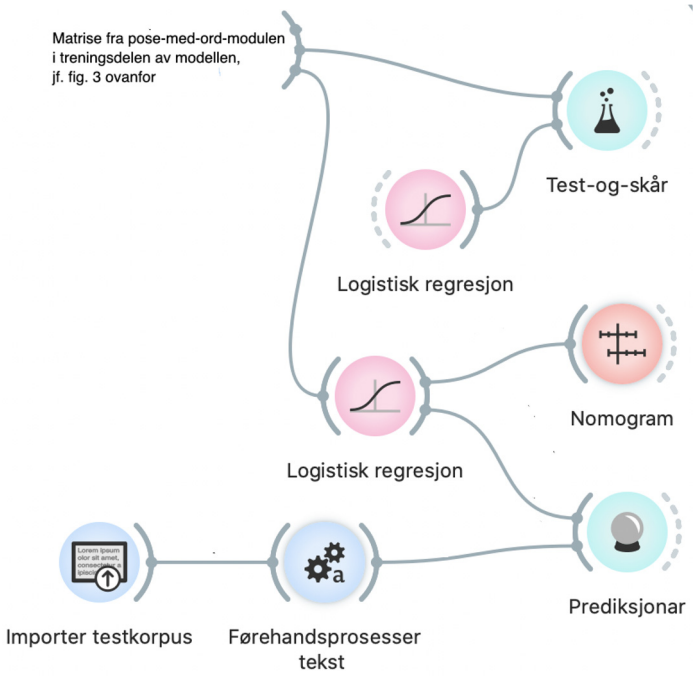
Om ein har korpus med morfologisk annotasjon, kan enklare logiske reglar skilja ulike variantar frå kvarandre. Ta midlandsmål som eit døme. Berre i denne normaliserte varianten av moderne norsk skriftspråk finst det fleirtals former på -ir og -ine, som t.d. *synir* der alle andre variantar vil ha *syner*. Éin einskild regel som *viss eit substantiv i ubund. form pl. ender på -ir, er målforma lik midlandsmål* vil då vera nok til å klassifisera målforma automatisk, og ein kan lett konstruera liknande morfologiske kriterium for alle normaliserte norske målformer med éin eller nokre heilt få reglar. Ulempa her er at metoden berre vil fungera for morfologisk annoterte korpus og ikkje kan nyttast på rein tekst.

Eit siste alternativ kunne vore å bruka djup læring (eng. *deep learning*) til å trena modellen. Antakeleg vil det laga den sterkaste modellen for å identifisera språk, men ulempa er at det er vanskeleg, eller umogleg, for eit menneske å forstå korleis modellen fungerer (jf. Øverland 2017, s. 223).

4.2. Strukturen til testdelen av modellen

Grunnlaget for strukturen til testdelen av modellen er dømemodellen som vert lagt fram i instruksjonsvideoen *Getting Started with Orange 18: Text Classification*[21]. Denne delen har den statistiske metoden logisk regresjon som kjernefunksjon og to hovudelement; eit for å testa og gje ein skår til prestasjonsnivået til modellen og ein for å predikera målforma til uklassifiserte tekstar. Grafisk sett ser modellen slik ut.

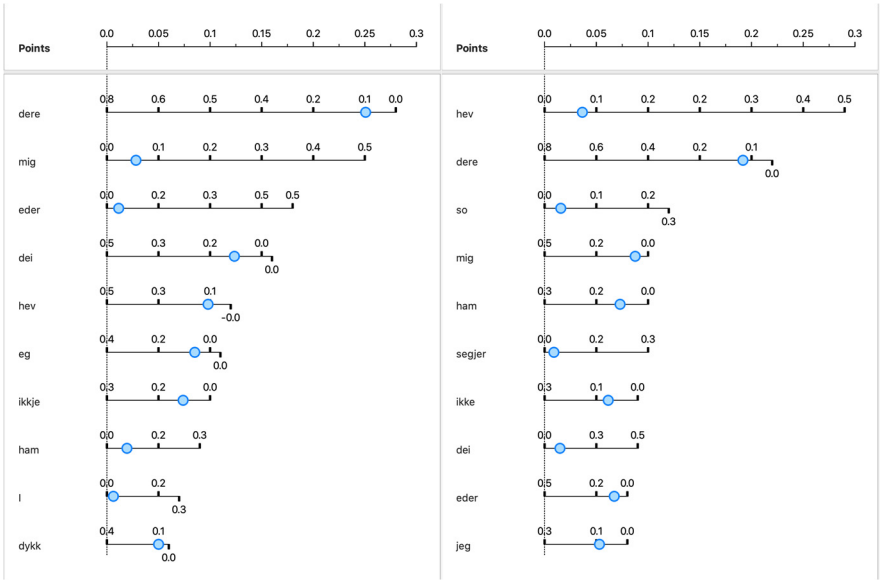
[21] https://youtu.be/zO_zwKZCULo (Sett den 29. juni 2020)



Figur 4. Strukturen til testdelen av modellen. (skjermdump frå Orange med omsetjing til norske namn på modulane og digital biletbehandling av forfattaren).

Matrisen med ordformer med vektorar som tel kor mange gonger dei opptrer i tekstane frå pose-med-ord-modulen i treningsdelen av modellen (jf. kap. 3.3 ovanfor) vert nytta som inndata for *logistisk regresjon* (orig. modulnamn *Logistic Regression*[22]), som genererer ein modell for å predikera kva for ei målform eit kapittel høyrer til. Det matematiske grunnlaget for logistisk regresjon er for omfattande til å gå inn på her, men det er ganske intuitivt å korleis LR-modellen fungerer om me visualiserer modellen i eit *nomogram* (orig. modulnamn *Nomogram*[23]).

[22] Jf. dokumentasjon på https://orange.biolab.si/widget-catalog/model/logisticregression/ (Sett den 29. juni 2020)
[23] Jf. dokumentasjon på https://orange.biolab.si/widget-catalog/visualize/nomogram/ (Sett den 29. juni 2020)



Figur 5. Nomogram for N1930BM (venstre) og N38NN (høgre) (skjermdump frå Orange).

Nomogrammet til venstre viser dei ti høgst rangerte orda LR-modellen for å predikera om eit kapittel er frå N1930BM. Me kan leggja merke til at alle orda er svært frekvente ord, og dei fleste eller alle ord som ville vorte rekna som *stoppord* om me såg etter semantiske mønster (jf. 3.2 ovanfor). Orda er det eit av to i minimale par av den typen som vart drøfta som mentale kriterier for manuell klassifisering i kap. 4.1. Det høgst rangerte ordet *dere* er ein del av det minimale paret *I*/*dere* som vart drøfta i kap. 2. Modellen seier altså at om ordet *dere* finst i ein tekst (dette er forma som vert nytta i N11BM), er det svært liten sjanse for at han har same målform som N1930BM medan om ordet *I* (her er det sjølvsagt naudsynt å skilja mellom små og store bokstavar) finst i teksten er det svært høgt sannsyn for at han har same målform som N1930BM. I andre tekstar med liknande innhald som Joh, vil dette kriteriet vera nok for å skilja eldre riksmål og moderne bokmål. Ordet *hev* (pres. av *hava* i målforma brukt N38NN) vs. *har* som vert nytta i N1930BM og *eg*, som vert nytta i både N38NN og N11NN, vs. *jeg* er vil vera nok til å skilja målform i N1930BM frå nynorsk i tekstar der desse orda finst. Nomogrammet kan gje eit falskt inntrykk av modellen berre vil fungera på tekstar med høg frekvens av personlege pronomen, sidan dei fleste av dei ti mest rangerte orda høyrer til denne ordklasse, men også ord med langt lægre rangering kan vera utslagsgjevande kriterium, men det er ikkje plass til å visa hundrevis av rangeringar her.

Modellen for kor sannsynleg det er at ei ordform har same målform som ei av utgåvene kan verta testa med av modulen *test-og-skår* (orig. modulnamn *Test and Score*[24]). I prosjektet i *Getting Started with Orange 18: Text Classification*[25], som, som nemnd i kap. 3 ovanfor, hadde som føremål å klassifisera eventyr frå *Brørne Grimms eventyr* i Aarne-Thompson klassifikasjonssystemet fekk LR-modellen ein AUC-skår (*Area under ROC curve*) på 0.91 (ibid. [2:03 – 2:34]). Det vil seia at modellen kan skilja mellom eit *dyreeventyr* og eit *eigenleg eventyr* i 91% av tilfella. Nomogrammet viser at utslagsgjevande ord er ord som *fox* 'rev' (høg sjanse for *dyreeventyr*) og *king* 'konge' (låg sjanse for *dyreeventyr*). Men desse semantiske kategoriane er ikkje like mekaniske som rettskrivingsreglar – t.d. er det ikkje umogleg at ordet *king* ovrar seg i eit dyreeventyr (eller at eit eventyr ikkje rettar seg etter Aarne-Thompson klassifikasjonssystemet i det heile, men blandar element frå *eigenlege eventyr* og *dyreeventyr* – nettopp dette vert oppdaga og drøfta i videoen), medan det på hi side *er* utenkjeleg at ordet *ikkje* opptrer i ein bokmålstekst (med mindre det er eit sitat frå ein nynorsktekst inni teksten, og det er for det fyrste ikkje aktuelt for det korpuset me har nytta og vil for det andre i nesten alle tilfelle ha så låg frekvens at det ikkje vil påverka prediksjonen). Såleis det er det ikkje uventa at modellen vår har ein AUC-skår på 1.0 og vil predikera kva utgåve eit kapittel høyrer til i korrekt i 100% av tilfella. Men som det vert påpeika i videoen: "*We don't want to predict something we already know*" (ibid. [2:35-2:42]).

Ved å setja inn ein korpusmodul til kan me testa andre uklassifiserte tekstar imot den modellen som har vorte trent med Johannesevangelie-utgåvene. Både testkorpuset og prediksjonane frå LR-modellen vert kopla til modulen *prediksjonar* (orig. modulnamn *Predictions*[26]) som vil freista å føreseia kva for ei av utgåvene eit anna tekstutdrag ligg nærmast.

5. Utprøving av testkorpus

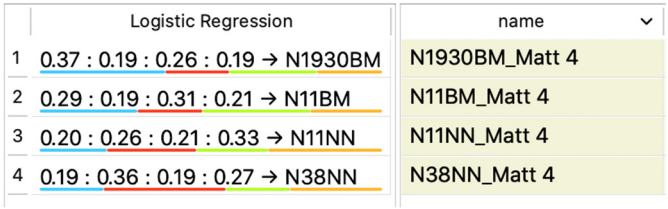
I det neste kapittelet vil me testa modellen med tekstar av aukande estimert vanskegrad. Det speler i prinsippet inga rolle kva slags tekstar som vert nytta for det er sannsynleg at alle norske tekstar av ei viss lengd inneheld nokre ord som vil ha ei sannsynsrangering i testdelen av modellen som vart drøfta i førre kapittel. Det er naturleg å rekna med at ein annan del av NT i dei same utgåvene av Bibelen som modellen har vorte trent på vil ha lægst vanskegrad, ikkje berre har dei nett same språkform, men omtalar også i stor grad om dei same

[24] Jf. dokumentasjon på https://orange.biolab.si/widget-catalog/evaluate/testandscore/ (Sett den 29. juni 2020)
[25] https://youtu.be/zO_zwKZCULo (Sett den 29. juni 2020)
[26] Jf. dokumentasjon på https://orange.biolab.si/widget-catalog/evaluate/predictions/ (Sett den 29. juni 2020)

hendingane og vil såleis ha mange parallelle ordformer for mindre frekvente ord også, eit døme kan vera ordforma *vinranke* (Joh 19: 29 og Luk 23: 36).

5.1. Matt 4 som testkorpus

Eit testkorpus bygd opp av kap. 4 av evangeliet etter Matteus frå dei same utgåvene modellen har vorte trent opp på gjev fylgjande resultat i prediksjonar-modulen.



	Logistic Regression	name
1	0.37 : 0.19 : 0.26 : 0.19 → N1930BM	N1930BM_Matt 4
2	0.29 : 0.19 : 0.31 : 0.21 → N11BM	N11BM_Matt 4
3	0.20 : 0.26 : 0.21 : 0.33 → N11NN	N11NN_Matt 4
4	0.19 : 0.36 : 0.19 : 0.27 → N38NN	N38NN_Matt 4

Figur 6. Prediksjonar for Matt 4. (Skjermdump frå Orange)

I fig. 6 indikerer blått sannsynet for at tekstutdraget har same målform som N1930BM, raudt for N38NN, grønt N11NN og oransje N11NN. Modellen har predikert riktig målform for alle dei fire versjonane av Matt 4. Den relativt høge rangeringa av dei andre utgåvene kan verka merkeleg, men til tross for skilnadene mellom dei, vil det vera mange ordformer som er identiske på tvers av i alle utgåvene, t.d. *du*.

5.2. Eit vilkårleg utval tekstutdrag som testkorpus

Neste nivå er å av vanskegrad er å testa modellen mot andre tekstar. Eg har vald eit vilkårleg utval av tekstar der med liknande målformer som testkorpuset. Novelletten *Sletten* av Sigbjørn Obstfelder[27] (språkforma liknar N1930BM, men har nokre trekk som ligg nærmare dansk) som døme på eldre riksmål, den nyleg publiserte forskning.no-artikkelen *Hvor mange tanker kan hjernen tenke samtidig?*[28] som døme på moderne bokmål, kap. 1 av Matias Skards omsetjing frå 1930 av *Soga um Håvard Isfjording*[29] som døme på nynorsk etter 1917-normalen og den nyleg publiserte nrk.no-artikkelen *No er brua over Noregs best besøkte*

[27] I Dokumentasjonsprosjektet ved UiOs utgåve på https://www.dokpro.uio.no/litteratur/obstfelder/ (Sett den 29. juni 2020)

[28] https://forskning.no/hjernen-menneskekroppen-psykologi/hvor-mange-tanker-kan-hjernen-tenke-samtidig/1705172 (Sett den 29. juni 2020)

[29] I digital utgåve frå Heimskringla.no på https://heimskringla.no/wiki/Soga_um_Håvard_Isfjording

naturattraksjon på plass[30] som døme moderne nynorsk. Modellen predikerte riktig målform i for alle fire utdraga med omlag same p-nummer som for Matt 4 i førre kapittel.

	Logistic Regression	name
1	0.28 : 0.20 : 0.30 : 0.22 → N11BM	Hvor mange tanker kan hjernen tenke samtidig?
2	0.23 : 0.27 : 0.23 : 0.27 → N38NN	Soga um Håvard Isfjording, kap. 1
3	0.22 : 0.25 : 0.22 : 0.31 → N11NN	No er brua over Noregs best besøkte naturattraksjon på plass
4	0.34 : 0.19 : 0.27 : 0.20 → N1930BM	Obstfelder - Sletten

Figur 7. Prediksjonar for vilkårleg valde tekstutdrag (Skjermdump frå Orange)

5.3. Testkorpus med nærskylde målformer

I førre kapittel vart det påvist at modellen kan predikera riktig målform for testkorpus med tekstar med tilsvarande språkdrakt. Som påpeikt i kap. 2 ovanfor, er det fleire variantar av norsk språk modellen ikkje har vorte trent på, inkl. midlandsmål og klassisk landsmål (som er mest nærskyld til N38NN) og 1800-tals riksmål (som er mest nærskyld til N1930BM). Utprøving av kap. 1 av bondeforteljinga *En glad gut*[31] av Bjørnstjerne Bjørnson som eit døme på 1800-tals riksmål, kap. 1 av Arne Garborgs roman *Mannfolk*[32] som døme på midlandsmål og Ivar Aasens omsetjing av *Den burtkomne sonen*[33] (d.e. Luk 15: 11-32). Modellen plasserte dei tre døma i dei mest nærskylde målformene; *En glad Gut* → eldre riksmål (N1930BM) og *Mannfolk* og *Den burtkomne sonen* → eldre nynorsk (N38NN).

5.4. Testkorpus med svært stutte tekstar

Dess lenger eit tekstutdrag er dess større sjanse er det for at det har ordformer som er vekta av modellen. Det kan såleis vera interessant å testa modellen på svært korte tekstar, og eg har nytta Olav H. Hauges tingdikt *Sagi* som testkorpus. Hauges målform fylgde med visse små avvik 1917-normalen, og riktig klassifikasjon vil altså vera → eldre nynorsk (N38NN). Diktet har berre elleve ord, og i i-målsforma i *sagi* kan ikkje vera til hjelp for klassifikasjonen

[30] https://www.nrk.no/vestland/den-kontroversielle-gangbrua-over-voringsfossen-er-kome-pa-plass-1.15076627 (Sett den 29. juni 2020)

[31] I Dokumentasjonsprosjektet ved UiOs utgåve på https://www.dokpro.uio.no/litteratur/bjoernson/ (Sett den 29. juni 2020)

[32] I Dokumentasjonsprosjektet ved UiOs utgåve på https://www.dokpro.uio.no/litteratur/garborg/ (Sett den 29. juni 2020)

[33] I digital utgåve frå Wikikilden på https://no.wikisource.org/wiki/Den_burtkomne_Sonen (Sett den 29. juni 2020)

fordi dette ordet ikkje finst i treningskorpuset. Diktet inneheld heller ingen av dei 10 høgast rangerte orda i modellen. Men likevel vekta modellen → eldre nynorsk (N38NN) med 30% sannsyn og predikerte at det var den mest sannsynlege målkategorien. Det er ordforma *segjer* som er utslagsgjevande, denne forma finst berre i N38NN (medan N11NN nyttar *seier*). Dei hine orda i diktet, *stød*, *ved*, *ho*, *kva tykkjer* har same form i N1NN (medan ljodordet *skrat,* som ein kunne venta, korkje finst i treningskorpuset eller ordbøker), så det einaste ordforma *segjer* modellen kan bruka til å klassifisera målforma riktig.

Det vil kanskje vera mogleg å finna korte tekstar som ikkje vil verta riktig klassifiserte, men om ein aukar storleiken til anten trenings- eller testkorpuset til over 20 ord, trur eg det er umogleg.

5.5. Forsøk på å lura modellen med homonymi er moglege…

Det er mogleg å lura modellen ved å medvite velja korte tekstutdrag som inneheld homonym som svarer til ord med høg vekt i modellen. Om me t.d. nyttar lina 1 frå strofe 4 av Arne Gardborgs dikt *Mot soleglad*[34]: "*Naar Dagen sig som Eld og Blod*" som testkorpus vil det verta feilklassifisert som → eldre riksmål (N1039BM) pga. den høge rangeringa av ordforma *sig,* som der er det frekvent brukte refl. pron. 3. pers og ei unik form for denne varianten, medan det i diktet er homonymet *sig* pres. av *å siga.*

Men slike freistnader når ikkje så langt. Om me nyttar heile strofa (18 ord) vert den riktig klassifisert som → eldre nynorsk, rett nok med berre litt høgare vekt for N38NN enn N1930BM.

6. Konklusjon og vegen vidare

Modellen som har vorte prøvd ut i denne studien har vist seg å vera svært solid for det føremål som vart presentert i innleiinga – å automatisk klassifisera kva for ein variant av norsk skriftspråk ein tekst er skriven i. Modellen vart trent med fire kategoriar (eldre riksmål, moderne bokmål, eldre nynorsk og moderne nynorsk), og ein kan rekna med at han ville fungert på same måte for andre variantar han ikkje vart trent på (t.d. 1800-tals riksmål, klassisk landsmål og midlandsmål). Når andre variantar vart nytta som testkorpus, i kap. 5.3, vart dei riktig klassifisert i den mest nærskylde varianten i treningskorpuset. Fyrst når me med vilje prøvde å lura modellen med homonym i svært korte tekstutdrag (<10 ord), i kap. 5.5., oppnådde me feilklassifisering.

[34] I Dokumentasjonsprosjektet ved UiOs utgåve på https://www.dokpro.uio.no/litteratur/garborg/ (Sett den 29. juni 2020)

Det er sannsynleg at modellen vil fungera like godt for å klassifisera variasjon i mange andre skriftspråk, eller identifisera kva for eit av fleire språk ein tekst er skriven på. Modellen er likevel ikkje den sterkaste til dette føremålet, og eg peikte på andre modellar i kap. 4.1. Føremona til modellen er at han er enkel å forstå og implementera og ikkje krev manuell programmering av reglar. Drøftinga her gjev neppe nye innsikter i norsk språkvitskap, men modellen er eit startpunkt som kan verta utvida, t.d. med eldre tekstar der det vanskelegare å nytta intuisjon for å sjå variasjonsmønster.

Modellen har fylgt døma i introduksjonsvideoane til Orange-prosjektet tett. Frå eit læringsperspektiv har kombinasjonen av eit visuelt programmeringsrammeverk og audiovisuelle læremiddel fungert svært godt. For nokre år sidan ville det vore naudsynt å ha EDB-fagleg bakgrunn for å vera i stand til å byggja og køyra ein slik modell. Orange skal ha honnør for å gjera denne typen databehandling meir tilgjengeleg, og eg tilrår alle som er interesserte i å bruka datautvinning i si forsking å prøva ut programmet.

I ein framtidig studie planlegg eg å testa ut den same modellen på det mellomnorske materialet i Diplomatarium Norvegicum. I kap. 3.3 peikte eg på utfordringar pose-med-ord-metoden har med korpus med stor variasjon. Likevel er min hypotese at det er mogleg at dette materialet er stort nok til å byggja opp ein modell som t.d. kan klassifisera typar av diplom eller dialektbakgrunnen/skriftnorma til skrivaren.

LITTERATUR

Primærkjelder

Bibel, bokmål 2011 (N11BM). Bibelselskapets omsetjing. Henta frå www.bibel.no/Nettbibelen (Sett den 29. juni 2020)

Bibel, nynorsk 2011 (N11NN). Bibelselskapets omsetjing. Henta frå www.bibel.no/Nettbibelen (Sett den 29. juni 2020)

Bibel, bokmål 1930 (N1930BM). Bibelselskapets omsetjing. Henta frå www.bibel.no/Nettbibelen (Sett den 29. juni 2020)

Bibel, nynorsk 1938 (N38NN). Bibelselskapets omsetjing. Henta frå www.bibel.no/Nettbibelen (Sett den 29. juni 2020)

Bibel, bokmål 1978/85 (N78BM). Bibelselskapets omsetjing. Henta frå www.bibel.no/Nettbibelen (Sett den 29. juni 2020)

Bibel, nynorsk 1978/85 (N78NN). Bibelselskapets omsetjing. Henta frå www.bibel.no/Nettbibelen (Sett den 29. juni 2020)

Bjørnson, Bjørnstjerne, *Samlede Digter-Verker*, Kristiania, Gyldendal, 1919-1920. Henta frå https://www.dokpro.uio.no/litteratur/bjoernson/ (Sett den 29. juni 2020)

Christensen, Mie Haugaard, *Hvor mange tanker kan hjernen tenke samtidig?* Publisert på https://forskning.no/hjernen-menneskekroppen-psykologi/hvor-mange-tanker-kan-hjernen-tenke-samtidig/1705172 2. juli 2020.

Den burtkomne sonen [d.e. Luk 15:11-32], til landsmål ved Ivar Aasen, ukjend årstal (1869 eller tidlegare). Henta frå https://no.wikisource.org/wiki/Den_burtkomne_Sonen (Sett den 29. juni 2020)

Dolve, Sjur Mikal og Hauso, Tale, *No er brua over Noregs best besøkte naturattraksjon på plass*. Publisert på https://www.nrk.no/vestland/den-kontroversielle-gangbrua-over-voringsfossen-er-kome-pa-plass-1.15076627 2. juli 2020.

Hauge, Olav H., *Dikt i samling*, Oslo, Samlaget, 2000.

Garborg, Arne, *Skriftir i samling*, -Jubilæumsutg. Kristiania, Aschehoug, 1921-1922. Henta frå https://www.dokpro.uio.no/litteratur/garborg/ (Sett den 29. juni 2020)

Obstfelder, Sigbjørn, *Skrifter*, København, Gyldendal, 1917. Henta frå https://www.dokpro.uio.no/litteratur/obstfelder/ (Sett den 29. juni 2020)

Soga um Håvard Isfjording, frå gamalnorsk ved Matias Skard, Oslo, Det norske samlaget, 1930. Henta frå https://heimskringla.no/wiki/Soga_um_Håvard_Isfjording (Sett den 29. juni 2020)

Sekundærlitteratur

Antti Aarne og Stith Thompson, "The Types of the Folktale: A classification and Bibliography", Helsinki, *FF Communications #184*, 1961.

De nyeste bibeloversettelsene til norsk, Det Norske Bibelskap. Henta frå https://www.bibel.no/OversettelseSprakLitteratur/Bibeloversettelser/Oversettelser-Norge/Etter1938 (Sett den 29. juni 2020))

Demsar J, Curk T, Erjavec A, Gorup C, Hocevar T, Milutinovic M, Mozina M, Polajnar M, Toplak M, Staric A, Stajdohar M, Umek L, Zagar L, Zbontar J, Zitnik M, Zupan B "Orange: Data Mining Toolbox in Python", *Journal of Machine Learning Research 14*, s. 2349–2353, 2013.

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

Getting Started with Orange 16: Text Preprocessing, Universitet i Ljubljana. Sett på https://youtu.be/V70UwJZWkZ8 (Sett den 29. juni 2020))

Getting Started with Orange 17: Text Clustering, Universitet i Ljubljana. Sett på https://youtu.be/rH_vQxQL6oM (Sett den 29. juni 2020))

Getting Started with Orange 18: Text Classification, Universitet i Ljubljana. Sett på https://youtu.be/zO_zwKZCULo (Sett den 29. juni 2020))

Getting Started with Orange 19: How to Import Text Documents, Universitet i Ljubljana. Sett på https://youtu.be/faIqvWxFGRc (Sett den 29. juni 2020))

Orange3 Text Mining Documentation, Universitet i Ljubljana. Henta frå https://orange3-text.readthedocs.io/en/latest/ (Sett den 29. juni 2020))

Orange Data Mining, Universitet i Ljubljana. Henta frå https://orange.biolab.si (Sett den 29. juni 2020, ulike delar av heimesida som har vorte nytta er nemnd fortlaupande med fotnotar)

Ordbog over det norrøne prosasprog [ONP], Københavns universitet. Henta frå https://onp.ku.dk/ (Sett den 29. juni 2020)

Oversettelser 1814-1938, Det Norske Bibelskap. Henta frå https://www.bibel.no/OversettelseSprakLitteratur/Bibeloversettelser/Oversettelser-Norge/Oversettelser1814-1938 (Sett den 29. juni 2020)

Retningslinjer for bruk av Bibelselskapets bibeloversettelser, Det Norske Bibelskap. Henta frå http://www.bibel.no/Nettbibelen/Opphavsrett_2 (Sett den 29. juni 2020)

Torp, Arne og Vikør, Lars S., *Hovuddrag i norsk språkhistorie*, Oslo, Gyldendal akademisk, 2014.

Ulike måter å oversette Bibelen på, Det Norske Bibelskap. Henta frå https://www.bibel.no/OversettelseSprakLitteratur/Bibeloversettelser/Oversettelsesmetoder (Sett den 29. juni 2020)

Øverland, Fartein Th. "Dróttkvætt and Deep Learning", Cluj-Napoca, *Dynamics of Specialized Languages: Innovative Approaches and Strategies*, s. 221-234, 2017.

Zellig S. Harris, "Distributional Structure", *WORD*, 10:2-3, s. 146-162, 1954.