Genustilordning i nynorsk:
Ei datamaskinell etterprøving

Gro Egset Halse

9 februar 2004

Hovudoppgåve i datalingvistikk
Seksjon for lingvistiske fag
Universitetet i Bergen

Samandrag.

Genus i norsk har vorte sett på som meir eller mindre arbitrært. Trosterud (2001) går ut frå at norsk har systematisk genustilordning, og presenterer eit regelsett bestående av semantiske, morfologiske og fonologiske genustilordningsreglar i nynorsk. For å kvantitativt etterprøve og optimalisere dette regelsettet, har datamaskinelle læringsmetodar vorte nytta. Eit datasett med 13384 nynorske substantiv har vorte konstruert, og til substantiva er det hovudsakleg manuelt lagt til 13 attributt som er naudsynte med omsyn til Trosterud sitt forslag til regelsett. Ein minnebasert maskinlæringsalgoritme er nytta, i tillegg til to regelbaserte algoritmar. Reglane genererte av dei regelbaserte metodane vart samanlikna med reglane til Trosterud, med det føremål å verifisere hypotesen hans om eit systematisk genustilordningssystem med maskulin som defaultgenus.

Abstract.

Some linguists claim that Norwegian gender is arbitrary. Trosterud (2001) assumes that gender assignment in Norwegian is systematic, and presents a set of semantic, morphological and phonological assignment rules for New Norwegian. Computational learning methods have been applied to quantitatively test and optimize this rule set. A database consisting of 13384 New Norwegian nouns has been constructed, and 13 multiple valued features necessary for the rules proposed by Trosterud, have mainly manually been added to the nouns. A memory-based learning approach has been applied, in addition to two rule-based algorithms. The rules generated by the rule-based methods were compared to Trosteruds rule set, with the aim of verifying his hypothesis of a gender assignment system of New Norwegian, with masculine as default gender.

Forord.

Dette er ei avsluttande hovudoppgåve for mitt studium i datalingvistikk ved Universitetet i Bergen. Mitt personlege mål med denne oppgåva var å lære meir om korleis maskinlæringsmetodar kan nyttast innan det lingvistiske feltet genustilordning. Det har så vidt eg veit ikkje vorte gjort liknande datamaskinelle undersøkingar for genustilordning i nynorsk.

Datasettet er ikkje vedlagt i papirversjon, men er tilgjengeleg i elektronisk versjon på seksjonen for lingvistiske fag ved universitetet i Bergen.

Eg vil rette ein stor takk til Trond Trosterud for bruk av hans materiale. Vidare vil eg takke rettleiaren min, Professor Koenraad de Smedt. Gjennom hans faglige kunnskap, råd og rettleiing har eg fått uvurderlig hjelp gjennom dette prosjektet. Han har også bidrege med program som vart nytta ved oppbygging av databasen.

Til slutt ønskjer eg å takke vener og familie for all hjelp, oppmuntring og støtte.

Innhald.

Appendiks A: Reglar med døme frå Trosterud (2001)
Appendiks B: Oversikt over tilgjengeleg materiale

1. Introduksjon.

Ein går ut i frå at norsk genus, eller ordkjønn, opprinneleg har vore relatert til biologisk kjønn. Talrike endringar gjennom den naturlege språkutviklinga har ført til at denne relasjonen no eksisterer i liten grad. Graden av regelbundenheit ved genustilordning (korleis morsmålstalarar tildelar substantiv ulike genus), er varierande frå språk til språk. Mange indo-europeiske språk syner lita regelbundenheit med omsyn til genus, og det dominerande synet i tilknytning til denne språkgruppa har vore at genustilordning er arbitrært.

Målet med denne hovudoppgåva er å etterprøve ein hypotese om at genustilordning i nynorsk er regelstyrt. Trosterud (2001) sitt framlegg til eit regelsett for genustilordning i nynorsk er nytta som grunnlag. Ved bruk av datamaskinelle læringsmetodar ønskjer ein å kvantifisere i kor stor grad dette regelsettet dekkjer nynorske substantiv, og ein er interessert i om regelsettet kan optimaliserast eller forbetrast. Etterprøvinga av Trosterud sitt regelsett kan i tillegg syne relevansen av dei ulike elementa semantikk, morfologi og fonologi ved genustilordning. Gjennom etterprøving av dette spesifikke regelsettet håpar ein å komme nærare eit svar på spørsmålet om genustilordning i nynorsk, og i ein større kontekst, indo-europeisk, er regelstyrt.

Resten av denne hovudoppgåva er organisert som følgjer: Kapittel 2 skildrar genus i nynorsk på bakgrunn av lingvistisk teori om genus. Kapittel 3 tek for seg Trosterud sitt framlegg til tilordningsreglar og ulike prinsipp som desse reglane er bygde på. I kapittel 4 vert dei datamaskinelle metodane som er tekne i bruk i denne oppgåva, omtala. Ulike typar maskinlæringsalgoritmer vert skildra, i tillegg til dei spesifikke systema som er nytta. Kapittel 5 inneheld ei skildring av datasettet med substantiv og attributtverdiar, og korleis dette er oppbygd i forhold til reglane til Trosterud. I kapittel 6 vert resultata frå dei ulike eksperimenta lagt fram og diskutert. Til slutt, i kapittel 7, vert relevansen av semantikk, morfologi og fonologi ved genustilordning diskutert på bakgrunn av feilratane frå ulike eksperiment. I tillegg vert det trekt ein del slutningar.

2. Lingvistisk bakgrunn.

2.1. Innleiing.

Dette kapittelet vil ta for seg genus, først og fremst genustilordning. For å kunne plassere genustilordning i nynorsk i forhold til genustilordning i andre språk, må ein vite skilnaden på ulike typar genustilordningsreglar, og korleis desse stiller seg i forhold til kvarandre. Her vil dei tre hovudtypane bli forklart, og døme vil bli gitt. I tillegg vil genustilordning i nynorsk verte omtala.

2.2. Den grammatiske kategorien genus.

Den grammatiske kategorien genus, eller ordkjønn, kan definerast som "klasser av substantiv som vert spegla av i korleis andre ord i samsvar med desse oppfører seg" (Hockett sin definisjon sitert i Corbett 119:1) Det er i orda i den syntaktiske omgjevnaden til eit substantiv at genus vert realisert, og det er på grunnlag av korleis desse orda oppfører seg at ein kan skilje ei genusklasse frå ei anna. Kva språklege element som syner genussamsvar, er språkavhengig. Vanlege element er mellom anna adjektiv og determinativ. Genussamsvar (kongruens) kjem til syne i desse elementa si form. I spansk til dømes, der artiklar er blant elementa som syner genussamsvar, skil den bestemte artikkelen mellom dei to genusa maskulinum og femininum ved at dei får henholdsvis formene *el* og *la*. Namn på genus er ikkje viktig i denne samanhengen. Genuset femininum i eitt språk kan omfatte andre substantiv enn femininum i eit anna språk gjer.

Ordet *genus* vert nytta både om den grammatiske kategorien og om ei klasse av substantiv. Det kan seiast at eit språk har den grammatiske kategorien genus, og at det til dømes har dei tre genusa maskulinum, nøytrum og femininum. Genus står i somme språk svært sentralt, medan andre språk ikkje har genus i det heile. (Corbett, 1991)

2.3. Genustilordning.

Genus har ofte delvis samanheng med naturleg kjønn, men dette er ikkje alltid tilfelle. I dei ulike språka i verda som har den grammatiske kategorien genus, kan substantiva vere delte inn i genusklasser etter ulike system (jf. 2.3.1-2.3.3 for døme).

Ein morsmålstalar må vite kva genusklasse eit substantiv tilhøyrer for å kunne produsere dei rette samsvarande elementa i den syntaktiske omgjevnaden til substantivet. Genustilordning, det vil seie korleis morsmålstalarar tildeler substantiv ulike genus, har vore mykje diskutert. Eit mogleg svar på korleis dette skjer, er at ein morsmålstalar hugsar genuset til kvart ord individuelt. Corbett (1991) framfører tre argument mot dette: Det første er at morsmålstalarar gjer få eller ingen feil ved genusbruk. Om genus for kvart ord vart hugsa individuelt, hadde ein forventa mange fleire feil. I tillegg argumenterer Corbett med at lånord i eit språk krev visse genus, noko som tyder på ei slags regelbundenheit ved genustilordning. Det tredje argumentet er at morsmålstalarar som vert presenterte for *nonsense-ord* (ord utan meining), har ein tendens til å tilordne desse orda genus på ein konsekvent måte. Desse tre argumenta peikar mot at morsmålstalarar tilordnar genus til substantiv på ein systematisk måte. Modellar på korleis dette skjer kallar vi tilordningssystem. For somme språk finst det ei etablert lingvistisk skildring av språket sitt genustilordningssystem, medan det i tilknytning til andre språk vert arbeidd med å etablere ei slik skildring.

Genustilordning skjer på grunnlag av ulike typar informasjon om substantivet. Corbett (1991) gir ei systematisk framstilling av genustilordning på grunnlag av ei slik inndeling: Informasjon om

substantivet kan delast opp i to hovudtypar: tyding (semantikk) og form. Form kan igjen delast inn i ordstruktur (morfologi), og lydstruktur (fonologi). Språk kan kombinere desse ulike typane informasjon på fleire måtar ved tilordning av genus. Corbett skil mellom tre typar tilordningssystem: *reine semantiske system, hovudsaklig semantiske system, og formelle system.*

2.3.1. Reine semantiske system.

Eit tilordningssystem har alltid ei semantisk kjerne, men semantikken er av ulik viktigheitsgrad. I reine semantiske system vert eit substantiv tilordna genus berre på grunnlag av semantikk. Forma til ordet vert ikkje teke omsyn til. Det finst nokre få unntak, men i prinsippet kan ein på grunnlag av tydinga til eit substantiv slutte seg til substantivet sitt genus. Slike system er ikkje særleg vanlege, men finst i til dømes dei fleste dravidiske språk og i ein del aust-kaukasiske språk (Corbett, 1991). Ulike semantiske system deler inn substantiv i semantiske klasser etter ulike kriterium. Det finst likevel kriterium som går igjen i mange språk. I det dravidiske språket tamil, som vert snakka hovudsakleg i Tamil Nadu i sør-aust-India, er substantiva delte inn i rasjonelle versus ikkje-rasjonelle substantiv (Corbett: 8). Dei rasjonelle er igjen delte inn i maskuline og feminine substantiv. Substantiv som denoterer biologisk hankjønn eller gudar, er maskuline, medan substantiv som denoterer biologisk hokjønn eller gudinner, er feminine. Alle andre, det vil seie ikkje-rasjonelle substantiv, er nøytrum. Guddommelege vesen sine roller i mytologien spelar i tamil inn ved genustilordning, og dette går igjen i mange av språka i verda. Eit skilje mellom rasjonelle og ikkje-rasjonelle substantiv og mellom biologisk hokjønn og biologisk hankjønn, er også vanleg i ei rekkje språk. Elles finn ein i språk med semantiske genustilordningssystem, inndeling av substantiv i grupper etter svært varierande semantiske kriterium.

2.3.2. Hovudsakleg semantiske system.

I motsetnad til reine semantiske system, som i prinsippet ikkje tillet unntak, finst det system som hovudsakleg tilordnar genus på grunnlag av semantikk, men som tillet sett av unntak i sine tilordningsreglar. Desse unntaka utgjer det som Corbett kallar *den semantiske resten*. Den semantiske resten er dei substantiva som ikkje får tilordna genus på grunnlag av eit positivt semantisk kriterium. I tamil (jf. 2.3.1) får maskuline og feminine substantiv genus etter positive kriterium, medan nøytrumssubstantiva utgjer den semantiske resten. I reine semantiske system får alle substantiva i den semantiske resten same genus. I hovudsakleg semantiske system derimot, får substantiva i den semantiske resten ulike genus. Slike system tillet *"lekkasjar"* mellom ulike genus. Genustilordning i det kaukasiske språket lak har fire genus:

Kriterium	**Genus**	**Døme**	**Tyding**
Mannleg rasjonell	I	las	ektemann
Kvinneleg rasjonell	II	ninu	mor
Andre animatar (men: somme kvinnelege menneske og mange inanimatar	III	nic	okse
Semantisk rest	IV	nex	elv

Tabell 1: Genustilordning i lak (Corbett, 1991:25)

I genus I og II finst berre substantiv som denoterer menneske og åndelege vesen. Det finst ingen lekkasjar frå andre genus og til desse. Genus III inneheld hovudsakleg inanimate substantiv, og genus IV den semantiske resten, men mellom desse to genusa finst det lekkasjar. Genus III og IV

inneheld ei rekkje unntak, i form av substantiv i genus III som ein forventar er i genus IV, og substantiv i genus IV som ein forventar er i genus III.

Regelsettet som Trosterud legg fram (jf. Kap. 3), peikar mot at norsk er eit hovudsakleg semantisk system. Heile 28 av reglane hans er semantiske tilordningsreglar, medan 9 er morfologiske og 3 fonologiske. Kor viktige desse semantiske reglane er for genustilordning i nynorsk, håper ein å få svar på gjennom den datamaskinelle etterprøvinga av Trosterud sitt regelsett.

2.3.3. Formelle system.

Når det gjeld formelle tilordningsreglar, skil ein i prinsippet mellom morfologiske og fonologiske reglar, sjølv om skiljet ikkje alltid er så klart. Ein fonologisk regel refererer til ei form av substantivet, medan ein morfologisk regel refererer til meir enn ei form.

Morfologiske system.

Det finst ingen reine morfologiske tilordningssystem. System som inneheld morfologiske tilordningsreglar, har alltid ei semantisk kjerne, og dei morfologiske reglane tilordnar genus til substantiv som ikkje får genus etter semantiske reglar, altså substantiv i den semantiske resten. Russisk har eit slikt tilordningssystem. I russisk, som har tre genus, får substantiv som denoterer biologisk hankjønn, genuset maskulinum, og substantiv som denoterer biologisk hokjønn, får femininum. Substantiv som ikkje får genus etter desse to semantiske tilordningsreglane, altså substantiva i den semantiske resten, får tilordna genusa maskulinum, femininum eller nøytrum etter morfologiske reglar, nærare bestemt reglar som tilordnar genus etter bøyingstype for kvart (bøyelege) substantiv:

1. Substantiv av bøyingstype I er maskuline
2. Substantiv av bøyingstype II og III er feminine
3. Substantiv av bøyingstype IV er nøytrum.

Russisk har fire hovudbøyingsparadigme, som gjer greie for dei aller fleste russiske substantiv. Substantiva vert bøygde i numerus (singularis og pluralis), og i seks ulike kasus. Ein morfologisk tilordningsregel refererer som nemnt til meir enn berre ei form av substantivet, i tilfellet for russisk til eit heilt bøyingsparadigme.

Russisk inneheld ei rekkje tilfelle av overlapping av semantiske og morfologiske tilordningsreglar. Substantivet *otec* (’far’) vil bli tilordna maskulint genus fordi det er av bøyingstype I. Slike tilfelle peikar mot at semantiske tilordningsreglar er overflødige i russisk. Men det finst også tilfelle der semantiske og morfologiske reglar motseier kvarandre, som tilfellet er for *djadja* (’onkel’), som er av bøyingstype II. Etter dei morfologiske tilordningsreglane skulle *djadja* få feminint genus, men får maskulint, fordi semantiske reglar overstyrer morfologiske. (Corbett, 1991)

Med omsyn til russisk kan ein spørje seg om det heller er slik at bøying rettar seg etter genus, det vil seie at maskuline substantiv er av bøyingstype I, osb. Dersom det er stort samanfall mellom morfologi og genus, løyser ikkje morfologi spørsmålet om korleis genus vert tilordna, fordi spørsmålet forskyv seg til korleis bøyingstype vert tilordna.

Fonologiske system.

Som det vart påpeika i 2.3.3, tilordnar ein fonologisk tilordningsregel genus til eit substantiv på grunnlag av ei enkel form av substantivet. Fonologiske kriterium for genustilordning kan vere til dømes den siste/ dei to siste fonene i eit substantiv, tal på stavingar, plassering av aksent med fleire.

Fransk har eit fonologisk tilordningssystem. Det vart lenge hevda frå mange hald at genus i fransk var tilfeldig, fram til Tucker, Lambert og Rigault i 1977 la fram data om distribusjonen av finale foner i substantiv (Corbett, 1991). Der dei finale fonene ikkje predikerte genus i tilstrekkelig grad, vart den nest siste og stundom også den tredje siste fonen teken omsyn til. Resultata frå desse studia var eit fonologisk tilordningssystem som klassifiserer majoriteten av franske substantiv med omsyn til genus. Døme på fonologiske reglar i fransk er:

1. Substantiv som endar på /ɛzɔ̃/, /sjɔ̃/, zjɔ̃/, / ʒjɔ̃/ og /tjɔ̃/ er feminine
2. Andre substantiv som endar på /ɔ̃/ er maskuline.

Fransk har i tillegg semantiske reglar og ein morfologisk regel. Dei semantiske reglane overstyrer både dei fonologiske og den morfologiske regelen, men dei aller fleste substantiva får tilordna genus etter fonologiske reglar.

Det finst som nemnt reine semantiske tilordningssystem, men ingen reine formelle system. Alle genustilordningssystem har ei semantisk kjerne med i alle fall eit minimum av semantiske reglar. I formelle tilordningssystem er det alltid ei viss overlapping av semantiske og formelle kriterium. Følgjeleg vil det finnast tilfelle av substantiv der fonologi og/ eller morfologi, i tillegg til semantikk, peikar mot same genus. Overlapping av kriterium kan ha ulike årsaker, ei av dei er avleiingsmorfologi. Om eit avleiingsuffiks med ei særskilt tyding er svært produktivt, og avleidde substantiv vert tilordna genus på grunnlag av dette suffikset, er konsekvensen mange substantiv med lik fonologi, morfologi og semantikk, og same genus. Eit døme på overlapping av kriterium for genustilordning i nynorsk, er substantiv som endar på *–eri*. Mange av desse er ord for samfunnsinstitusjonar og kulturelt skapte stader (*bakeri, fargeri* osb.) Trosterud inkluderer i sitt regelsett både ein regel *STR 1: Ord for kulturelt skapte stader og for samfunnsinstitusjonar, er n*, og ein regel *MTR 8: Ord avleidd med suffiksa –eri, -ment, -skop, er n.* Desse to reglane tilordnar begge nøytrum, og overlappar for ein del ord som endar på *–eri*. Slike tilfelle av overlapping kan gjere det vanskeleg å avgjere kva kriterium som eigentleg tilordnar genus, og dermed kva type tilordningssystem ein har med å gjere.

2.4. Føremålet med systematisering av genustilordning i reglar.

Eit vanleg syn på genus innan lingvistikken, har vore at genus i eit språk ikkje vert tilordna ved hjelp av eit sett av tilordningsreglar, men at genus derimot er arbitrært. Først etter at Corbett si *Gender* kom ut i 1991, har forsking på genustilordning vorte sett i fokus (Trosterud, 2001).

Det teoretiske føremålet med å setje opp eit sett av genustilordningsreglar for eit språk, vil vere å forstå leksikonet sin struktur betre (Corbett, 1991). Med dette meiner ein å forstå genus i seg sjølv, og i tillegg å forstå utviklinga av genus i eit diakront perspektiv.

Eit viktig praktisk føremål med genustilordningsreglar er å redusere problem ved læring av framandspråk. Om eit språk har få eller ingen klare reglar for tilordning av genus, vert innlæring av genus eit problematisk område for framandspråklege. Ved å forstå på kva grunnlag ein tilordnar genus til substantiv, og å systematisere kriterium i eksplisitte reglar, kan ein hjelpe framandspråklige til å lære inn genus på ein meir systematisk måte.

2.5. Genustilordning i norsk.

Både bokmål og nynorsk har tre genus: maskulinum, femininum og nøytrum. Riksmålsvarianten av bokmål har berre to genus, då maskulinum og femininum har smelta saman i eit felleskjønn.

Det dominerande synet på genustilordning innan nordistikken har til no vore at genus i norsk (bokmål og nynorsk) i prinsippet er arbitrært. Dette synet dominerer framleis, sjølv om ein i nokre norske grammatikkar finn ei rekkje genustilordningsreglar (Trosterud, 2001). Både i Beito (1986 [1970]), og i *Norsk Referansegrammatikk* (Faarlund, Lie og Vanneboe, 1997), finst det systematiske framstillingar av genustilordningsreglar i norsk. Begge framstillingane inneheld ei rekkje semantiske tilordningsreglar, i tillegg til nokre reglar som tilordnar genus etter form. Faarlund, Lie og Vanneboe (1997) hevdar trass i reglane at genus i dei aller fleste tilfelle korkje samsvarar med form eller tyding av substantivet.

Innan fagfeltet norsk for framandspråklege dominerer ei anna oppfatning, nemleg at det er mogleg å setje opp reglar for genustilordning (Trosterud, 2001). Føremålet med dette fagfeltet er å leggje norskundervisninga for framandspråklege til rette for best mogleg læring. Læring av genus i norsk er på grunn av arbitrariteten eit problemområde for framandspråklege, og reglar for tilordning av genus kan minske desse problema. I *Norsk på grunnlag av samisk* (1999), ei lærebok for elevar med nordsamisk som førstespråk i den vidaregåande skulen, legg Leirvaag fram ei rekkje genusreglar. 21 reglar vert framstilt, danna på grunnlag av enten semantikk eller form, og det vert ikkje stilt spørsmål ved reglane si gyldigheit. Husby (1990) legg i si bok om ordlaging i norsk, fram ei rekkje suffiks som dannar substantiv i norsk. Heile 58 suffiks med tilhøyrande genus er inkluderte i denne framstillinga.

Også i lånordforskinga har kriterium for genustilordning vorte drøfta, og genus har dermed vorte sett på som ein del av grammatikken (Trosterud, 2001). Graedler (1998), som omhandlar engelske lånord i norsk, inneheld ein hypotese for tilordning av genus til engelske lånord i norsk, bestående av semantiske og morfologiske tilordningsreglar. Graedler hevdar at den systematiske måten lånord vert tilordna genus på, kan nyttast som bevis på at det finst ein mekanisme for genustilordning.

Denne hovudoppgåva tek for seg nynorsk først og fremst fordi Trosterud sitt framlegg til genustilordningsreglar i nynorsk er eit einestående utgangspunkt for etterprøving ved hjelp av maskinlæring. Det finst ingen framlegg til like omfattande regelsett for bokmål.

3. Trosterud sitt forslag til tilordningsreglar.

Trond Trosterud er ein av få nordistar som har gått ut i frå at genustilordning ikkje er arbitrært, men ein del av grammatikken. Han gjer i sin artikkel *Genustilordning i nynorsk er regelstyrt* (2001) greie for eit forslag til eit sett av tilordningsreglar for nynorsk. Som korpus har han nytta 31500 usamansette ord frå Nynorskordboka, utanom ord med genusvariasjon for same tyding, men inkludert elles identiske ord med ulikt genus (Trosterud, 2001:30). Han hevdar at reglane hans tilordnar korrekt genus til om lag 94% av substantiva i korpuset.

Trosterud presenterer tre generelle, overgripande reglar, og i tillegg ei rekkje spesifikke reglar. Av dei spesifikke reglane er 28 semantiske, 9 morfologiske og 3 fonologiske. Dei spesifikke reglane overstyrer dei generelle. Nokre av reglane er inspirerte av eksisterande tilordningsreglar, men i tillegg er ei rekkje nye reglar danna ut i frå Trosterud sine eigne prinsipp for genustilordning. Om fleire reglar veg eit substantiv mot ulike genus, vil substantivet få det genus som flest eller mest tungtvegande reglar talar for. For å gi eit innblikk i korleis dei ulike typane reglar fungerer i forhold til kvarandre, følgjer ein gjennomgang av ein del av dei reglane som Trosterud presenterer. (Heile regelsettet inkludert døme finst i appendiks A)

3.1. Generelle reglar.

Trosterud tek utgangspunkt i den hypotesen at defaultgenus i nynorsk er maskulinum, med andre ord at dersom ingen regel tilseier noko anna, så er eit substantiv maskulint:

Regel 1: Default: Alle norske ord er m

Fordi Trosterud går ut i frå at maskulinum er defaultgenus, inneheld regelsettet relativt få andre reglar for tilordning av maskulinum. I tillegg til defaultregelen legg han fram to overgripande fonologiske reglar:

Regel 2: Tostava ord på trykklett *–e* er f

Regel 3: Einstava ord på vokal er f

Unntak til desse reglane vert tilordna genus av meir spesifikke semantiske eller fonologiske reglar som overstyrer desse.

3.2. Genusekstensjon og genusinversjon.

Trosterud presenterer to semantiske genustilordningsprinsipp som han kallar *genusinversjon* og *genusekstensjon*. Desse prinsippa går igjen som grunnlag for fleire av tilordningsreglane hans. Med genusinversjon meiner han at eit semantisk eller morfologisk felt kan bli etablert ved at grupper av ord i dette feltet får eit anna genus enn dei skulle ha fått etter overgripande reglar. Desse orda skil seg på ein systematisk måte frå ord som vert tilordna genus etter dei overgripande reglane, og eigne tilordningsreglar for slike grupper av ord kan etablerast. STR (*semantisk tilordningsregel*) 5 og 6 er døme på slike reglar:

STR 5: Genusinversjon for kroppsdelar: Ord på *–C* for kroppsdelar er n.

STR 6: Genusinversjon for kroppsdelar: Ord på *–e* for kroppsdelar er m.

Ved hjelp av genusinversjon vert eit semantisk felt for ytre organ på menneskekroppen etablert. Når det gjeld kroppsdelar, får ord som endar på konsonant, nøytrum (jf. STR 5), og ord som endar på trykklett *–e,* maskulinum (jf. STR 6). Ord for kroppsdelar får då eit anna genus enn dei skulle ha fått etter dei overgripande fonologiske reglane. Trosterud presenterer i tillegg ein regel som gjeld unntak til denne inversjonsregelen:

STR 7: Ord for sentrale kroppsdelar kjem ikkje inn under genusinversjon.

Denne regelen uttrykkjer at ord for sentrale kroppsdelar, av Trosterud definert som 'ord knytte til dei viktigaste sanseapparata for kognitiv verksemd' (Trosterud: 37), ikkje følgjer inversjonsregelen, men derimot dei overgripande fonologiske reglane.

Det andre tilordningsprinsippet som Trosterud presenterer, det han kallar genusekstensjon, tilordnar genus ved at det først vert etablert eit genus for ei gruppe av ord etter semantiske kriterium, og at dette genuset så vert utvida til ord av same morfologiske eller fonologiske form. Dette synes i to av dei morfologiske tilordningsreglane til Trosterud:

MTR 8: Ord avleidd med suffikset *–eri, -ment, -skop*, er n.

MTR 9: Ord på *–ine, -inne, -enne, -ette, -øse, -ette*, er f.

Når det gjeld MTR (*morfologisk tilordningsregel*) 9, er mange av substantiva som denne regelen dekkjer, ord som refererer til kvinner (*prinsesse, blondine* osb). Det semantiske kriteriet *hokjønn* er grunnlag for tilordning av feminint genus: (jf. 3.3: STR 1), og andre substantiv med identiske suffiks har ved hjelp av genusekstensjon fått same genus (*delikatesse, mitraljøse* osb). Det same har skjedd når det gjeld MTR 8. Mange av substantiva som dette gjeld, har tydingane 'stad/ kulturelt skapt stad/ samfunnsinstitusjon' (*bakeri, drogeri* osb), og vert tilordna nøytrum, i følgje STR 10 og STR 15:

STR 10: Ord for stader er n.

STR 15: Ord for kulturelt skapte stader og for samfunnsinstitusjonar, er n.

Genusekstensjon har ført til at også andre substantiv med identiske suffiks (*broderi, argument* osb.) vert tilordna nøytrum.

3.3. Semantiske tilordningsreglar.

STR 1: Der det er ulike ord for referentane med ulikt biologisk kjønn, har orda tilsvarande genus

STR 1 uttrykkjer den semantiske kjernen i tilordningssystemet, og mange av dei andre semantiske reglane står i eit metaforisk forhold til denne. Ein metafor som går att i fleire av reglane er *form*. Dette synes i to av reglane som tek for seg det semantiske feltet *ytre organ på menneskekroppen*:

STR 8: Metaforisk genustilordning: Ord for mannlege kjønnsorgan og andre avlange organ er m.

STR 9: Metaforisk genustilordning: Ord for kvinnelege kjønnsorgan og hol er f.

Metaforen *form* tilordnar ifølgje Trosterud, også genus til ord for ulike terrengformasjonar (STR 11 og STR 12) og for avlange objekt (STR 16):

STR 11: Metaforisk genustilordning: Ord for terrengtoppar og avlange terrengformasjonar er m.

STR 12: Metaforisk genustilordning: Ord for terrengfordjupingar er f.

STR 16: Ord som refererer til avlange objekt, er m.

Eit anna metaforisk tilhøve til den semantiske kjerna uttrykt i STR 1, har med fruktbarheit å gjere. Dette er uttrykt i to reglar:

STR 17: Ord for stein og mineralar er m.

STR 18: Ord for jord er f.

Trosterud presenterer ei rekkje semantiske felt som han meiner er konstituerte på grunnlag av metaforiske tilhøve til den semantiske kjernen i STR 1, eller ved genusinversjon. For ein presentasjon av alle dei semantiske felta og alle tilordningsreglane kan det refererast til Trosterud (2001).

3.4. Morfologiske tilordningsreglar.

Dei morfologiske tilordningsreglane som Trosterud legg fram, er delte opp i bøyingsreglar og avleiingsreglar. Trosterud påpeikar at genus i norsk stort sett ikkje er avhengig av bøying, slik som for eksempel i russisk, der nesten alle substantiv sitt genus kan determinerast ut i frå bøyingsklasse. Han har likevel funne regelmessigheiter i norsk fleirtalsbøying, noko som har resultert i to bøyingsreglar:

MTR 1: Omlydssubstantiv er f

MTR 2: Substantiv utan segmental formativ i ub.pl, er n.

Når det gjeld omlydssubstantiv, er nesten alle feminine, og dei få unntaka som finst, får andre genus fordi MTR 1 vert overstyrt av semantiske reglar. MTR 2 er i normalisert nynorsk og enkelte dialektar tautologisk, sidan alle nøytrumsord vert bøygde utan segmental formativ i fleirtal. Ein risikerer å forskyve spørsmålet om genus til kvifor orda har ei viss bøying, og MTR 2 kan av den grunn vanskeleg forsvarast som genustilordningsregel.

Trosterud inkluderer sju avleiingsreglar. Nokre av dei meiner han er danna ved genusinversjon: Regel 2 og 3 dekkjer mange av substantiva som endar på vokal, og defaultregelen dekkjer difor mange substantiv på konsonant. Verbalsubstantiv av verbstammen endar oftast på konsonant, og MTR 3 kan difor sjåast på som ein inversjon av Regel 1:

MTR 3: verbalsubstantiv av verbstammen er n

Substantiv danna av adjektivstamme pluss trykklett *–e*, er maskuline (jf. MTR 4), og er difor ein inversjon av Regel 2. MTR 5-7 er reglar for tilordning av avleiingar med *–ing* og *–heit*. Tilsvarande reglar er inkludert i Faarlund, Lie og Vanneboe (1997). MTR 8 og MTR 9 er som nemnt danna ved genusekstensjon.

3.5. Fonologiske tilordningsreglar.

Trosterud legg fram tre fonologiske tilordningsreglar på grunnlag av fonologiske endingar; to for substantiv som endar på konsonant og likevel ikkje er maskuline, men feminine (jf. FTR 1) eller nøytrum (jf. FTR 2), og ein for substantiv som endar på trykklett *-e* og som likevel ikkje er feminine, men maskuline (jf. FTR 3). Desse reglane overstyrer dei overgripande fonologiske reglane fordi dei er meir spesifikke:

FTR 1: Ord på *–idd, -emd, -erd, -Cn, -rg, -ft, -o:d, -vd, -pt, -kt, -V[+høg]:l,* og einstava ord som endar på *–gd* og *–V:,* er f.

FTR 2: Ord på *–V:d, -om, -e:m, -um, -ym, -a:r, -ie:r, -iv, a:t*, er n.

FTR 3: Ord på *–Rbe* og *-V:be* er m.

3.6. Tilordningsreglar for norske fornamn.

I tillegg til tilordningsreglar for substantiv generelt i nynorsk, legg Trosterud fram eit sett av fonologiske og morfologiske reglar som skal tilordne genus til norske fornamn, altså på grunnlag av desse reglane skilje mellom mannsnamn og kvinnenamn. I dette hovudfagsprojektet er det sett bort i frå namnereglane. Særnamn er ikkje inkluderte i datasettet som skal nyttast til å etterprøve Trosterud sine reglar. Namnereglane er spesifikke reglar som gjeld berre for mannsnamn og kvinnenamn, og ei etterprøving av desse krev eit eige datasett med særnamn og tilhøyrande attributtverdiar.

4. Metodar.

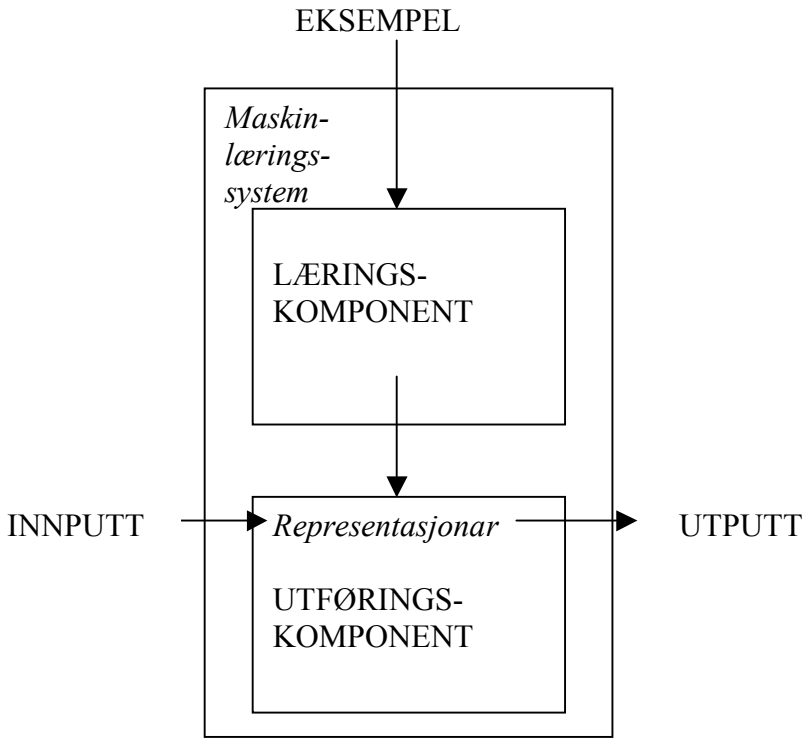
4.1. Maskinlæring.

Maskinlæring er ei grein av KI (kunstig intelligens) som nyttar algoritmar som automatisk lærer av erfaring til ulike klassifikasjonsføremål. Slike algoritmar står i kontrast til klassifikasjonsmodellar som er laga av menneskelege ekspertar (Quinlan, 1992), til dømes regelbaserte algoritmar som inneheld dersom-så-reglar. Erfaring vert gitt til eit maskinlæringssystem ved hjelp av eit sett med eksempel som vert nytta til å trene systemet. Etter trening av systemet er målet å kunne klassifisere nye ukjente eksempel. Det finst ulike typar maskinlæringsmetodar som på ulike måtar nyttar eksempel til trening av ein klassifikator. Nevrale nettverk nyttar eksempla til å trene eit nettverk av prossesseringselement med koplingar i mellom. Slike nettverk er inspirerte av korleis biologiske nevronar i hjernen prossesserer informasjon, og dei lærer ved å bli presentert for same informasjon i form av eksempel, mange gongar. I motsetnad til andre typar maskinlæringssystem, lærer altså eit nevralt nettverk av seg sjølv, og ikkje ut i frå ein ytre faktor. Andre læringsmetodar krev ein ytre faktor i form av ein spesifikk algoritme. To typar algoritmar, minnebasert læring og induksjon av reglar og beslutningstre, vert omtala i henholdsvis 4.2 og 4.3.

Det går eit skilje mellom såkalla overvaka og uovervaka læring, det vil seie læring med og utan ”fasit”. Ved uovervaka læring inneheld ikkje eksempla i treningssettet si korresponderande klasse. Her må systemet sjølv finne likskapar mellom eksempla under trening på ein slik måte at det kan klassifisere nye eksempel (Daelemans og Durieux, 2000). Ved overvaka læring derimot, er kvart eksempel knytta til si riktige klasse. For vårt føremål vert det nytta overvaka læring, fordi vi har tilgang på kvart substantiv sitt genus, og desse skal nyttast ved trening av ein klassifikator.

Ved overvaka læring er kvart eksempel representert av ein vektor av *n* attributtverdiar, i tillegg til eksempelet si korresponderande klasse. (Daelemans og Durieux, 2000). Målet er at systemet etter trening på desse eksempla, ved hjelp av ein læringsalgoritme, skal kunne klassifisere nye ukjente eksempel. For å kunne gjennomføre eit maskinlæringseksperiment trengst eit treningssett og eit testsett. Treningssettet inneheld eit sett av eksempel, og predefinert klasse for kvart eksempel. Kvart eksempel har *n* attributt som kan få ulike verdiar. Desse verdiane kan vere *binære* (kvar klasse har to moglege verdiar, til dømes sann/usann, T/NIL), *symbolske* (til dømes bokstavar, ord), eller *numeriske* (til dømes oppteljingar, signalmålingar) (van den Bosch, 2002). Testsettet inneheld ukjente eksempel, med andre ord eksempel bestående av attributtverdiar, men utan predefinert klasse. Systemet skal etter trening på treningssettet klassifisere eksempla i testsettet. I vårt tilfelle består eit eksempel av ein vektor av attributtverdiar som inneheld semantisk, morfologisk og fonologisk informasjon for det aktuelle substantivet, nærare bestemt informasjon som er nødvendig med omsyn til Trosterud sitt regelsett (jf. Kap. 5). Treningssettet inneheld i tillegg genus for kvart substantiv, medan genus i testsettet er ukjent.

Meir systematisk sett kan ein seie at ein maskinlæringsalgoritme består av to komponentar, ein læringskomponent og ein utføringskomponent. Utføringskomponenten produserer utputt gitt eit visst innputt. Læringskomponenten modifiserer utføringskomponenten på grunnlag av erfaring, slik at systemet forbetrar prestasjonen (Daelemans og Durieux, 2000):



Figur 1: Arkitekturen til eit maskinlæringssystem (basert på figur 1 i Daelemans og Durieux:5)

4.2. Minnebasert eller lat læring.

Ved *lat læring* (også kalla *minnebasert læring* eller *eksempelbasert læring*) vert eksempla i treningsdata lagra i minnet, og ved klassifikasjon vert dei ukjente eksempla samanlikna med dei lagra eksempla. Her finst det ingen abstraksjon eller rekonstruksjon av data ved læring, derav namnet *lat læring*.

Minnebasert læring er basert på hypotesen om at kognitive oppgåver vert utførte på grunnlag av likskap mellom nye situasjonar og lagra representasjonar av allereie erfarte situasjonar. Dei lagra representasjonane tek ved minnebasert læring form av eksempel i eit treningssett. Læringskomponenten er minnebasert: Læring skjer ved at eksempla i treningssettet vert lagra i minnet. Utføringskomponenten er likskapsbasert: Klassifikasjon av eksempla i testsettet skjer ved at kvart testeksempel X vert samanlikna med alle eksempla i minnet Y. Likskapen mellom X og alle Y vert rekna ut ved hjelp av ein likskapsfunskjon, og den mest frekvente klassa blant dei *k* mest like eksempla i minnet (*k-nn: k nearest neighbours*) vert gitt som kategori til det nye eksempelet. (Daelemans, Zavrel, van der Sloot og van den Bosch, 2001)

Ein type likskapsfunksjon er ein såkalla *overlappingsfunksjon*. Ved bruk av denne vert likskap definert som talet på like attributtverdiar hos to eksempel som vert samanlikna. Denne funskjonen ser på alle attributtverdiar i ein vektor som like relevante ved klassifikasjon. Men dette er ikkje alltid tilfelle, og det finst ulike metodar for vekting og selektering av verdiar (jf. 4.4.1). (Daelemans og Durieux, 2000)

4.3. Beslutningstre som læringsmodell.

I motsetnad til minnebaserte læringsmetodar, som baserer seg på ein hypotese om læring på grunnlag av likskap mellom nye og lagra eksempel, er *grådig læring* grunnlagt på ein hypotese om at kognitive oppgåver vert utført ved at mentale reglar vert abstraherte frå tidlegare erfaringar og applikerte på nye situasjonar (Daelemans, Zavrel, van der Sloot og van den Bosch, 2001). Trening av ein grådig læringsalgoritme skjer ved at ein abstrakt modell, til dømes eit beslutningstre eller eit regelsett, vert konstruert på grunnlag av likskapar og forskjellar mellom eksempel i treningssettet. Desse abstrakte modellane vert nytta ved klassifikasjon av ukjente eksempel. Reglar og beslutningstre er to modellar som har ulik utsjånad, men som er ekvivalente.

Eit beslutningstre er ei ordning av testar, med ein egna test for kvart steg i ein analyse (*Overview of Decision Trees*). Målet er å, på grunnlag av testar om attributtverdiane, generere eit beslutningstre som forutseier kategoriar riktig. I *Building Classification Models: ID3 and C4.5* er det gitt eit døme på korleis eit beslutningstre vert bygd opp og nytta som klassifikasjonsmodell. Føremålet med klassifikatoren er på grunnlag av ulike vêrforhold å predikere om ein kan spele golf eller ikkje. Vêrforholda er attributtverdiar:

Attributt	**Moglege verdiar**
Vêrutsikter?	Sol, overskya, regn
Temperatur	Kontinuerlig
Fuktigheit	Kontinuerlig
Vind	Sann, usann

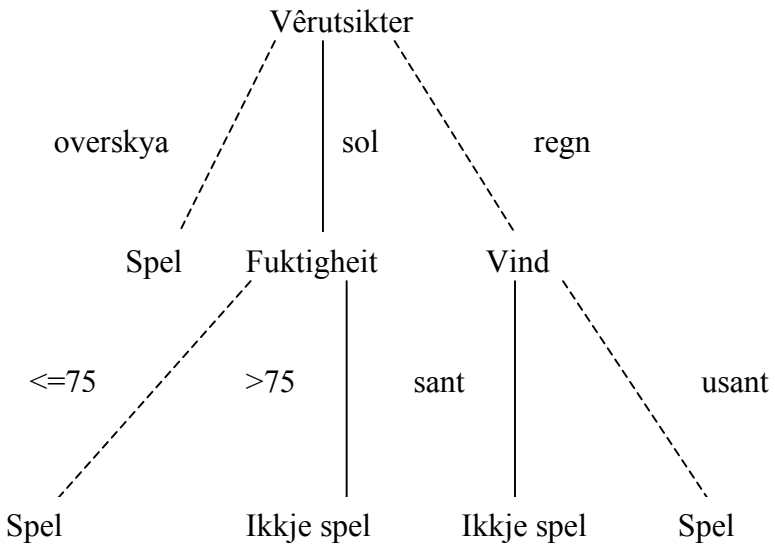
Tabell 2 : Attributt og verdiar for vêrforhold (*Building Classification Models:ID3 and C4.5*)

Treningsdata er bygd opp av eksempel bestående av desse attributtverdiane, i tillegg til ønska klasse for kvart eksempel. Kvar rad står for eit eksempel:

Vêrutsikter	**Temperatur**	**Fuktigheit**	**Vind**	**Klasse**
Sol	85	85	Usant	Ikkje spel
Sol	80	90	Sant	Ikkje spel
Overskya	83	78	Usant	Spel
Regn	70	96	Usant	Spel
Regn	68	80	Usant	Spel
Regn	65	70	Sant	Ikkje spel
Overskya	64	65	Sant	Spel
Sol	72	95	Usant	Ikkje spel
Sol	69	70	Usant	Spel
Regn	75	80	Usant	Spel
Sol	75	70	Sant	Spel
Overskya	72	90	Sant	Spel
Overskya	81	75	Usant	Spel
Regn	71	80	Sant	Ikkje spel

Tabell 3: Treningsdata (*Building Classification Models:ID3 and C4.5*)

Beslutningstreet som vert danna, summerer opp fakta om eksempla i treningsdata:



Figur 2: Eit beslutningstre for golfeksempelet (*Building Classification Models: ID3 and C4.5*)

Kvar node i beslutningstreet spesifiserer ein test for eit av attributta til eksempelet (t.d. 'vêrutsikter'), og kvar grein som kjem frå denne noden, korresponderer til ein mogleg verdi (t.d. 'regn') for dette attributtet. Eit blad refererer til ein kategori ('Spel/Ikkje spel'). Når treet vert gjennomgått ved klassifikasjon, startar ein ved rota til treet og går igjennom heile treet heilt til ein kjem til eit blad (Quinlan, 1992). Først vert attributtet som rotnoden representerer testa ('vêrutsikter'), og ein går deretter ned igjennom den greina som korresponderer til verdien til dette attributtet. Dette vert repetert heilt til ein kjem til eit blad, som representerer eksempelet sin kategori. (*Overview of Decision Trees*)

Klassifikasjonsmodellar i form av regelsett er ekvivalente med beslutningstremodellar. Modellane har ulik utsjånad, men fungerer i prinsippet på same måte ved klassifikasjon.

4.4. Dei spesifikke algoritmane.

Tre ulike maskinlæringsalgoritmar vert nytta ved etterprøving av tilordningsreglane til Trosterud. To av dei, *RIPPER* og *C4.5RULES*, er grådige metodar, og er valt fordi dei genererer klassifikasjonsmodellar i form av reglar. C4.5RULES genererer eit regelsett utifrå eit beslutningstre, medan RIPPER genererer reglar direkte frå treningssettet. Reglane som vert genererte vert samanlikna med regelsettet i Trosterud (2001). Føremålet er å finne ut noko om korleis RIPPER og C4.5RULES nyttar den semantiske, morfologiske og fonologiske informasjonen uttrykt i eksempla, ved klassifikasjon av substantiv i genusklasser. Ein er interessert i graden av samsvar med Trosterud sitt regelsett, og i tillegg eventuelle nye regelmessigheiter. Essensielt er også spørsmålet om visse typar attributt (semantiske, morfologiske, fonologiske) bidreg meir enn andre ved klassifikasjon. I tillegg er ein interessert i kor stor del av substantiva i datasettet systema greier å klassifisere riktig.

Den tredje algoritmen som er nytta, *TiMBL* (*Tilburg Memory-Based Learner*), er ein minnebasert metode, og genererer difor ingen abstrakt klassifikasjonsmodell. TiMBL er av den grunn lite informativ i forhold til dei to grådige algoritmane. Det ein er interessert i er sjølve feilraten ved klassifikasjon, for samanlikning med prestasjonane til dei regelbaserte metodane.

4.4.1. TiMBL.

TiMBL lagrar alle eksempla i treningssettet i minnet, og klassifiserer nye ukjente eksempel i eit testsett, ved at kvart eksempel vert tildelt klassa til det mest like eksempelet i treningssettet.

Inndeling i trenings-og testeksempel vert helst gjort ved bruk av ein av dei to metodane *10-fold cross-validation* eller *leave-one-out*: *10-fold cross-validation* vil seie at det vert utført ti eksperiment, der kvart av eksperimenta nyttar 90% av datasettet som treningssett, og 10% som testsett, slik at kvart eksempel vert nytta som testeksempel berre ein gong. Ved bruk av *leave-one-out* vert kvart eksempel i datasettet nytta som testeksempel ein gong, og klassifikatoren vert trena på resten av eksempla. Fordi det ikkje krev ei føreåt inndeling av datasettet før trening, har eg valt å nytte *leave-one-out*.

For TiMBL kan det gjerast ei rekkje val med omsyn til algoritmar og innstillingar for avstandsmålingar. Eit val står mellom algoritmane *IB1* og *IGTREE*. IB1 gir oftast eit meir presist resultat, men på bekostnad av hurtigheit. IGTREE på den andre sida, er meir effektiv, men gir ofte eit dårlegare resultat. Sidan IB1 gir eit jamnt betre resultat, er denne default, og vil bli nytta i denne oppgåva. IB1 går ut på at ein måler avstanden (likskapen) mellom to eksempel ved å slå saman avstanden mellom attributtverdiane. Avstanden mellom to eksempel er summen av verdiane.

Når IB1 vert nytta, kan ein velje mellom to innstillingar som påverkar definisjonen av likskap. Med *vekta overlapping* får kvart attributt tildelt ein verdi som bestemmer relevansen attributtet har med omsyn til det aktuelle klassifikasjonsproblemet. Defaultmetode for vekting av attributtrelevans er *Gain Ratio*, som er ein normalisert versjon av *Information Gain*. *Information Gain* ser på kvart attributt for seg, og måler kor mykje informasjon det bidreg med ved klassifikasjon. *Information Gain* overestimerer ofte relevansen av attributt med mange verdiar, og *Gain Ratio* er ei forbetring som gjer at talet på verdiar ikkje har betydning ved vekting av eit attributt. Om ein ikkje vil nytte *Gain Ratio* som vektingsmetode, er det mogleg for brukaren av TiMBL å setje sine eigne vekter til kvart attributt. Med vekta overlapping vert to attributtverdiar sett på som enten like eller ikkje like. Det finst ei anna innstilling, *MVDM (modified value difference metric)*, som tillet grader av likskap mellom verdiar. Likskapen mellom kvart par av verdiar av same attributt vert då rekna ut, og dette vert gjort for alle attributt. Alle verdipar får såleis tildelt eit mål på avstanden mellom seg. I vårt tilfelle kan til dømes to fonologiske verdiar verte målte som meir like enn to andre, basert på at dei opptrer oftare i same omgjevnader.

Ved klassifikasjon av eit ukjent eksempel er det mogleg, i staden for å berre ta omsyn til det eksempelet i treningssettet som liknar mest, å ta omsyn til fleire eksempel, eller fleire næraste naboar. (Daelemans, Zavrel, van der Sloot, van den Bosch, 2001)

4.4.2. C4.5 og C4.5RULES

C4.5 genererer eit beslutningstre som synt i 4.3. For å gjere treet mindre komplekst, vert også eit forenkla beslutningstre generert ved at delar av det komplekse treet som ikkje bidreg til nøyaktig klassifikasjon, vert tekne bort. Dette skjer ved at ein del subtre vert erstatta med blad eller med dei

hyppigast nytta greinene i subtrea. Når eit tre vert kutta, vil det oftast feilklassifisere ein del tilfelle i treningssettet, og blada på det kutta treet vil difor nødvendigvis ikkje innehalde tilfelle berre innanfor ei klasse. Kvart blad vert av den grunn ikkje assosiert med ei bestemt klasse, men med ei sannsynligheit for at eit eksempel i dette bladet vil tilhøyre ei bestemt klasse. Kutting av eit beslutningstre gjer treet både enklare og meir nøyaktig og kan til og med gi ei lågare feilrate ved klassifikasjon. (Quinlan, 1992: 35)

Målet er ikkje berre ein nøyaktig klassifikator, men også ein som er forståeleg og såleis kan gi innsikt i korleis eit klassifikasjonsproblem vert løyst. Ved kompliserte klassifikasjonsproblem vert sjølv det forenkla beslutningstreet så stort og infløkt at det er umogleg for menneske å forstå det fullt ut. For å gjere klassifikasjonsmodellen meir forståeleg kan han uttrykkjast i form av reglar som vert danna ut i frå beslutningstreet, av C4.5RULES. Ei forenkla form for produksjonsreglar vert nytta, L -> R. Regelen si venstreside L inneheld dei føresetnadene som må vere tilfredsstilt for at eit eksempel skal klassifiserast som det som regelen si høgreside R uttrykkjer (dersom L så R). Klassa som er uttrykt i R, er den same som er uttrykt i eit av blada i beslutningstreet, og føresetnadene i L er alle dei føresetnadene ein finn ved å følgje stien frå rota til treet og til det bladet som står for den aktuelle klassa. Om alle stiane som fører til eit blad i beslutningstreet, skulle verte omskrive til ein regel, så vert regelsettet like komplisert som treet. Av den grunn vert irrelevante føresetnader utelatne i reglane, det vil seie føresetnader som ikkje bidreg til å skilje den klassa det er snakk om frå andre klasser.

Reglar vert genererte frå kvar sti som fører til eit blad i beslutningstreet, men ein del av desse reglane vert utelatne frå klassifikasjonsmodellen av reglar. Ein regel vert utelaten om han har ei feilrate som er for høg, eller om han dupliserer reglar som er genererte frå andre stiar. Regelsettet vil difor ha færre reglar enn det er blad i beslutningstreet. Ein konsekvens av dette er at det vil finnast tilfelle i datasettet som beslutningstreet er generert frå, som ikkje vert dekt av nokon av reglane. Ein *defaultregel* vil difor bli inkludert blant reglane, det vil seie ein regel som tilordnar ei klasse til alle dei tilfella som ikkje vert dekt av nokon av dei andre reglane. Systemet vel som defaultklasse den klassa som inneheld flest treningseksempel som ikkje vert dekt av nokon regel. Ein annan konsekvens av generalisering av reglar, er at det vil finnast eksempel som vert dekt av meir enn ein regel. Dette vert løyst ved at systemet set opp ei prioritering av reglar, slik at den første regelen som dekkjer eit eksempel, vert nytta til å klassifisere eksempelet.

Ved klassifikasjon av eit eksempel ut i frå ein modell av reglar, vert reglane gjennomgått, og når systemet finn ei venstreside av ein regel som samsvarar med eksempelet, vert høgresida gitt som klasse. Om ingen venstresider samsvarar med eksempelet, får eksempelet defaultklassa. (Quinlan, 1992)

I samband med vår problemstilling vil C4.5RULES bli nytta, sidan reglane dette programmet genererer vil vere enklare å samanlikne med Trosterud sine reglar enn eit beslutningstre.

4.4.3. RIPPER.

RIPPER står for *Repeated Incremental Pruning to Produce Error Reduction*, og er ei forbetring av *IREP* (*Incremental Reduced Error Pruning*). I staden for å nytte eit beslutningstre som grunnlag for eit regelsett, dannar IREP eit regelsett ved å generere ein regel om gongen. Etter at ein regel er danna, vert alle eksempel som er dekt av regelen sletta, og denne prosessen vert gjenteken til det ikkje finst positive tilfelle, eller til regelen har ein uakseptabelt høg feilrate. Forbetringane i RIPPER går ut på at regelsettet som er generert av IREP, R*1*.....R*k*, vert optimalisert på følgjande

måte: For kvar regel R*i*, vert det konstruert to alternative reglar. Den eine, den såkalla erstatninga for R*i*, vert danna ved at ein regel R*a* vert generert, og deretter simplifisert ved til dømes å slette enkelte føresetnader, for å minimere feilraten til heile regelsettet. Den andre alternative regelen, den reviderte utgåva av R*i*, vert danna ved at fleire føresetnader vert lagt til R*i*. Til slutt vert det bestemt om den originale regelen, erstatninga eller den reviderte utgåva skal inkluderast i det endelege regelsettet. Etter at eit regelsett er konstruert og optimalisert, vert reglar lagt til ved hjelp av IREP, for å dekkje gjenståande positive tilfelle. (Cohen, 1995)

5. Skildring av databasen.

5.1. Innleiing.

Dette kapittelet skildrar korleis ein database av substantiv med attributtverdiar har vorte bygd opp i forhold til Trosterud sitt forslag til tilordningsreglar. Sjølve substantiva som er inkluderte er omtala, og i tillegg korleis desse har fått tildelt attributtverdiar i forhold til reglane til Trosterud. Databasen er bygd opp slik at den vert direkte innputt til læringsalgoritmane.

5.2. Substantiva.

Substantiva i databasen er henta frå Trosterud sitt materiale. Han hevdar å ha nytta 31500 substantiv frå Nynorskordboka som grunnlag for sitt framlegg til regelsett. Ideelt sett skulle alle desse 31500 vore inkluderte, men på grunn av mangel på tilgang til alt av Trosterud sitt materiale, inneheld databasen berre 13384 substantiv. 2968 av dei har feminint genus, 7761 har maskulint genus, og 2655 er nøytrum.

I tillegg til dei 13384 substantiva som er inkluderte i databasen, inneheldt materialet frå Trosterud ein del substantiv som har vorte kasta ut ved oppbygging av databasen. For det første fanst det ein del duplikat som eg ikkje var interessert i å behalde. I tillegg inneheldt materialet nokre typar ord som eg av ulike årsaker ikkje ville inkludere. Dette gjeld mellom anna ord med genusvariasjon for same tyding. Desse hevdar Trosterud å ikkje inkludere i sitt korpus, og dei vert difor heller ikkje inkluderte ved etterprøvinga av hans reglar. Dei to følgjande orda er døme på ord med genusvariasjon, som har vorte utelatne frå databasen (frå *Nynorskordboka, 3. utgåva, 2001*):

> talg f1 el. m1 (norr tolg f, uvisst opph)
> feitt frå drøvtyggjarar, særleg i innmat og innvolar .
>
> II snork [II snurk] m1 el. n1 snorking;
> einskild snorkande lyd høyre s- frå soverommet .

Elles er, som Trosterud sjølv påpeikar, identiske ord med ulikt genus inkluderte, til dømes (frå *Nynorskordboka, 3. utgåva, 2001*):

> I bank m1 *banke (I) .
>
> II bank m1 (gj fr frå it. banca, banco, eigl 'pengevekslardisk'; opph
> germ, sm o s *benk)
> 1 institusjon som tek mot innskot, gjev lån, driv handel med
> verdipapir og yter ymse tenester når det gjeld
> veksling og overføring av pengar og valuta setje, låne pengar i b-en
> / sikker som b-en
> 2 bygning der ein *bank (II,1) held til gå i b-en
> 3 pengesum (av innsats og innbetalte tap) som gevinstane blir
> utbetalte av; spelebank sprengje b-en vinne
> så mykje at kassa blir tom
> 4 (reserve)lager, opplagsstad blodb-

Berre usamansette ord er i følgje Trosterud grunnlag for hans regelsett, og samansetnadar er difor heller ikkje inkludert i vår database. Ein del samansetnadar som fanst i materialet frå Trosterud,

vart utelatne ved oppbygging av databasen. Grunnen til at samansetnadar er uinteressante med omsyn til genustilordning, er at dei utgjer ei produktiv klasse av substantiv. Sisteleddet i ein samansetnad eksisterer oftast også som eit uavhengig ord med same genus som den aktuelle samansetnaden. Det samansette ordet vil difor ikkje bidra med noko nytt med omsyn til genustilordning, og det er ikkje hensiktsmessig å inkludere samansetnadar. Nedanfor følgjer eit døme på ein samansetnad, og eit ord som er identisk med sisteleddet i samansetnaden (*Nynorskordboka, 3. utgåva, 2001*):

> farty~ el. fartøy|byggjar [~byggar] m1
>
> byggjar [byggar] m1 person som byggjer brub- / husb- / innb- .

Det usamansette ordet *byggjar* finst i datasettet. Å i tillegg inkludere *fartøybyggjar* hadde ikkje tilført relevant informasjon til databasen, sidan ein *fartøybyggjar* også er ein *byggjar* eller *ein person som byggjer. Fartøybyggjar* er difor utelate.

I tillegg til samansetnadar har eg valt å utelate ei gruppe av ord som oppfører seg på same måte, i den forstand at dei ikkje tilføyer noko til databasen med omsyn til semantiske eigenskapar. Dette gjeld ein del avleiingar der suffikset er eit leksikonoppslag med ei særskild tyding (*frå Nynorskordboka, 3. utgåva, 2001*):

> I -no´m m1 (frå gr, sjå *-nomi)
> 1 -kunnig, t d i agronom, sosionom og økonom
> 2 i namn på apparat som gjev regel for noko, t d i metronom

I-Nom er ei produktiv ending og har to spesifikke tydingar. Her, og for liknande tilfelle, som *-fil, -sofi* med fleire, er berre suffikset inkludert i databasen, og ord som inneheld suffikset er ikkje tekne med. *-Nom* er ført opp to gongar fordi den eine tydinga får verdien 'person' (jf. tabell 4.5), og den andre ikkje. Når det gjeld suffiks som også finst som sjølvstendige ord, som *–mani/ mani*, er suffikset utelate, og berre det sjølvstendige ordet behaldt. Substantiv danna med suffiks som *–ing* og *-heit*, er inkluderte i databasen. Slike suffiks har inga særskild tyding, men er nytta til å danne substantiv. Det er naudsynt å inkludere desse i forhold til etterprøving av ein del morfologiske reglar som tilordnar genus etter desse suffiksa (jf. tabell 4.10).

Også ein del klammeformer som finst i Trosterud sitt materiale, er utelatne frå datasettet. Ei klammeform vil oftast ha same genus som hovudforma av det aktuelle substantivet, og semantikken og morfologien vil ikkje variere mellom formene. Berre når det gjeld fonologi kan ei klammeform syne variasjon i forhold til hovudforma. Fordi klammeformer ikkje vil få tildelt attributtverdiar som varierer mykje frå hovudformene, er klammeformer ikkje inkluderte i databasen. Når det gjeld klammeforma *[byggar] m1* og hovudforma *byggjar m1* til dømes, er ein *j* etter *g* det einaste som skil klammeforma frå hovudforma. Det er dessutan hovudforma av substantivet som er den mest nytta forma, og difor den som vert fokusert på.

Med omsyn til nokre av substantiva som skulle inkluderast måtte det takast eit par avgjersler angåande tyding og form. Ei av desse avgjerslene gjaldt substantiv med fleire tydingar, der den eine ikkje er tilknytta den opprinnelege tydinga eller opphavet til det aktuelle ordet, men kjem frå metaforisk bruk av ordet og har etter kvart vorte til ei eiga tyding i leksikonoppslaget. Til ei slik metaforisk tyding vil det vere tilknytta andre semantiske verdiar enn den opprinnelege tydinga, og dette kan lage ugreie i resultata av eit maskinlæringseksperiment. Tilstrekkelig mange slike tilfelle kan føre til andre resultat enn om metaforiske tydingar er sett bort i frå, fordi reglane for den

opprinnelge tydinga vil påverke genus for den metaforiske tydinga. Metaforiske tydingar er difor i utgangspunktet ikkje tekne omsyn til når substantiva i datasettet har vorte tildelt attributtverdiar. Eit døme på eit substantiv med ei metaforisk tyding er *klyse* (Frå Nynorskordboka, 3. utgåva, 2001):

> klyse f2 (smh med kleime, jf eng *cluster 'klase' og lat. gluten 'lim')
> 1 klatt av seig og tjukk væske spyttek-
> 2 (slimet) vase (II,1) få ei k- av tang i garnet
> 3 slapp, ekkel person

Opphavet til *klyse* er knytta til tyding 1 og 2 i leksikonoppslaget, og tyding 3 må difor vere metaforisk. *Klyse* av tyding 3 vil få den semantiske verdien 'person' (jf. tabell 4.5). Om ein ser for seg at dei fleste substantiva med attributtverdien 'person' er maskuline, og at ein regel i tilknytning til dette vert generert av ein regelbasert maskinlæringsmetode, vil mange tilfelle av feminine substantiv med ei metaforisk tyding 'person', føre til ei rekkje unntak til denne regelen. Fordi tyding 3 er sett bort i frå, får ikkje *klyse* verdien 'person' i databasen. Det kunne derimot vere interessant å i tillegg utføre eksperiment der slike tydingar er inkluderte, for å sjå forskjellen i resultata. Dette vert ikkje gjort i denne omgang. Det kan diskuterast om slike tydingar bør utelatast i det heile, sidan dei er inkluderte i ordboka, og dermed er ein del av språket.

Det måtte også takast stilling til kva som skulle gjerast med substantiv med to ulike former. Desse har vorte rekna som to ulike tilfelle i databasen. For tilfelle som *'mynd f1 el IImynde n1*' er dette sjølvsagt, sidan dei to ulike formene får ulike genus. Men det finst også tilfelle der eit substantiv har to ulike former som får same genus, til dømes *'I bed el. I bedd m1*' Her er det óg naudsynt å skilje dei to formene, fordi dei vil generere ulike verdiar for fonologiske attributt, nærare bestemt dei siste tre bokstavane i ordet (jf. tabell 4.12-4.14)

5.3. Attributt og verdiar.

Innhaldet i databasen står i *csv* (*comma separated values*)-format med ei linje for kvart substantiv, der kvart attributt har ein fast plass som eit av dei aktuelle verdiane til dette attributtet kan ta opp. I tillegg til ein plass for kvart attributt inneheld kvar linje det aktuelle substantivet sitt genus:

Substantiv, a2, a3, a4, a5, a6, a7, a8, a9, a10, a11, a12, a13, a14, genus.

Om eit substantiv ikkje kan skildrast av nokon av verdiane til eit visst attributt, vert plassen som tilhøyrer attributtet oppteken av eit spørsmålsteikn. Eit utdrag frå datasettet syner korleis datasettet er oppbygd:

```
barndom,?,?,?,?,?,?,?,?,?,C,d,o,m,M.
bauxitt,?,?,stein,?,?,?,?,?,?,C,i,t,t,M.
feiring,?,?,?,?,?,?,?,verb,ing,C,i,n,g,F.
femininum,gram_kat,?,?,?,?,?,uten_segm_form,?,?,C,n,u,m,N.
fromheit,?,?,?,?,?,?,?,adj,heit,C,e,i,t,F.
havre,plante,?,?,?,?,?,?,?,?,E,v,r,e,M.
Ibry,?,?,?,?,?,?,uten_segm_form,verbalsubst,?,V,b,r,y,N.
Idokk,?,?,?,?,hol,terreng,omlydssubst,?,?,C,o,k,k,F.
Idominikanar,?,?,?,person,?,?,?,?,?,C,n,a,r,M.
kvalme,?,?,?,?,?,?,?,adj,?,E,l,m,e,M.
fot,parvis,kroppsdel,?,?,avlang,?,?,?,?,C,f,o,t,M.
hals,?,kroppsdel,?,?,avlang,?,?,?,?,C,a,l,s,M.
```

Attributta med ulike verdiar er organiserte i forhold til tilordningsreglane til Trosterud. Før dette kunne gjerast måtte det takast stilling til om det skulle nyttast binære eller symbolske/diskrete verdiar (jf. 4.1.). Eit problem med binære verdiar er ineffektivitet som følgje av eit stort tal på verdiar med relativt liten grad av informasjon. Grunnen til dette er at mange av attributta utelet kvarandre. Ved bruk av binære verdiar ville ein, i staden for eit attributt 'form' med fire verdiar 'avlang', 'hol', 'flate', og 'funksj_holrom' (jf. tabell 4.6), trenge fire attributt, kvart med verdiane 'T/NIL'. For substantivet *pinne* til dømes, vil attributtet 'avlang' få verdien 'T'. Dei tre attributta 'hol', 'flate' og 'funksj_holrom' (funksjonelt holrom), får alle verdien 'NIL' fordi desse fire attributta utelet kvarandre. Det vert difor nytta diskrete verdiar, noko som gjer at ein ved hjelp av berre tretten attributt for kvart substantiv kan få med all nødvendig informasjon. Resultat frå Hendrickx og van den Bosch (2003), der ulike eksperiment vart utførte både med binære og diskrete verdiar, syner dessutan at det generelt sett ikkje finst fordelar i samband med bruk av binære verdiar.

Ved organisering av semantisk, morfologisk og fonologisk informasjon i attributt og verdiar, er målet at verdiane til eit attributt ikkje overlappar, altså at kvart substantiv ikkje kan ha meir enn ein verdi av kvart attributt. Dette for å få med all relevant informasjon om kvart enkelt substantiv. For å gi eit tydelegare inntrykk av korleis denne inndelinga vart gjort, vil kvart av attributta med verdiar verte kommentert, og eventuelle problem i forhold til desse verte diskutert. For lettare å sjå forholdet mellom dei ulike verdiane og Trosterud sine reglar er dette sett opp skjematisk, med ein tabell for kvart attributt. Tabellane nedanfor syner alle moglege verdiar av det aktuelle attributtet, kva for tilordningsregel/-reglar som har vore utgangspunkt for kvar verdi, og døme på substantiv som får denne verdien. Sjølve substantiva vil opptre som attributt 1 i datasettet, men dette er berre til hjelp for å lettare kunne finne att eit substantiv. Substantiva vert ikkje tekne omsyn til som verdiar ved maskinlæringseksperiment. Dei seks første attributta (attributt 2-7) er semantiske, attributt 8-10 er morfologiske og 11-14 er fonologiske. Namna på attributta er ikkje sjølvforklarande, men må sjåast på i forhold til Trond Trosterud sine tilordningsreglar (jf. appendiks A).

Det er ikkje noko ein-til-ein-forhold mellom attributtverdiar og genustilordningsreglar. Fleire av reglane inneheld ein kombinasjon av semantiske og fonologiske, semantiske og morfologiske eller morfologiske og fonologiske komponentar, eller ein kombinasjon av ulike semantiske komponentar. Eit døme på det første er STR 6, som uttrykkjer at *"ord på -e for kroppsdelar er m "* (jf. appendiks A) Denne regelen gir eit behov for ein attributtverdi som referer til substantiv som endar på trykklett *-e*, og ein verdi som refererer til kroppsdelar. Verdien 'e' kjem under attributt 11, og er, som tabellen syner, også nødvendig i forhold til fleire semantiske reglar, ein overgripande fonologisk regel, og ein morfologisk regel. 'Kroppsdel' er ein verdi av attributt 3, og er relevant i samanheng med i alt fire tilordningsreglar, STR 5, 6, 8 og 9, som alle tek for seg substantiv i det semantiske feltet ytre organ på menneskekroppen (Trosterud: 37). Ein attributtverdi viser til ein eller fleire tilordningsreglar.

Attributt 2: Diverse semantikk.

Verdi	Tilsvarande tilordningsregel	Døme på substantiv
Hyperonym	STR 4	folk, idrett
Bokstav	STR 21	a, b, c, d, e
Parvis	STR 22	bukse, saks, fot
Heimleg_tre	STR 23	bjørk, eik, furu
Plante	STR 24	eføy, einer, anis
Meieri	STR 25	fløyte, myse, kefir
Gram_kat	STR 26	adjektiv, kasus, infinitt
Lyd	STR 20	bjeff, ekko, hokuspokus

Tabell 4.2: Diverse semantikk

Attributt 2 inneheld verdiar som ikkje nødvendigvis har nokon samanheng med kvarandre, anna enn at dei alle er semantiske. Grunnen til at dei er sette opp som verdiar av same attributt, er at dei ikkje overlappar. Hyperonym vil seie overordna substantiv, som *dyr, folk, instrument* osb. (jf. Trosterud: 36). Plantar utelet blomar, som i følge Trosterud følgjer vanlege reglar, og ikkje STR 24 (*ord på –e for plantar er m*). 'Gram_kat' står for grammatiske kategoriar, og substantiv som får verdien 'lyd', er substantiv som står for lydar, språkhandlingar og sitatord (jf. Appendiks A: STR 20).

Attributt 3: Menneskekroppen.

Verdi	Tilsvarande tilordningsreglar	Døme på substantiv
Kroppsdel	STR 5, 6, 8 og 9	bein, hovud
Sentral	STR 7	munn, nase

Tabell 4.3: Menneskekroppen

'Kroppsdel' vil seie ytre organ på menneskekroppen (Trosterud: 37). I samband med STR 6 nyttar Trosterud *finne* og *flanke* som døme på kroppsdelar, sjølv om dei er ord for dyrekroppsdelar. *Finne*, *flanke* og andre dyrekroppsdelar har i datasettet ikkje fått attributtverdien 'kroppsdel'. Verdien 'sentral' står for sentrale kroppsdelar. Med det meinast ord knytte til dei viktigaste sanseapparata for kognitiv verksemd (Trosterud: 37).

Attributt 4: Stoff.

Verdi	Tilsvarande tilordningsregel	Døme på substantiv
Stein	STR 17	bentonitt, gneis, kalk
Jord	STR 18	leire, mold
Stoff	STR 19	gull, deig, harpiks

Tabell 4.4: Stoff

Attributt 4 har tre ulike verdiar: ’Stein’ viser til ord for stein og mineralar, ’jord’ til ulike ord for jord, og ’stoff’ til ord for stoff og masse. Ord for stoff og masse vil seie substantiv som kan stå utan artikkel og som ikkje er tellelege.

Attributt 5: Biologisk kjønn.

Verdi	**Tilsvarande tilordningsregel**	**Døme på substantiv**
Hankj	STR 1	bamse, greve
Hokj	STR 1	binne, grevinne
Person	STR 2	admiral, kunde, kurdar
Avkj	STR 3	beist, foster, fruentimmer

Tabell 4.5: Biologisk kjønn

’Hankj’ og ’hokj’ står for biologisk hankjønn og biologisk hokjønn. Substantiv som refererer til personar utan kjønnsspesifisering (Trosterud: 35), får verdien ’person’. ’Avkj’ viser til ”ord som refererer til personar og har nedsetjande tyding , og ord for avkjønna vesen” (Trosterud: 36).

Attributt 6: Form.

Verdi	**Tilsvarande tilordningsregel**	**Døme på substantiv**
Avlang	STR 8, 11 og 16	canyon, haug, finger, fot, påle
Hol	STR 9 og 12	grop, gjel, navle
Flate	STR 27	blad, diskett, duk
Funksj_holrom	STR 28	åk, andlet, hjul, rør, øyre

Tabell 4.6: Form

Verdien ’avlang’ refererer til ord for mannlege kjønnsorgan og andre avlange organ (jf. STR 8), og til avlange objekt (jf. STR 16). Ord for avlange organ og mannlege kjønnsorgan får i tillegg verdien ’kroppsdel’ av attributt 3, og terrengtoppar får verdien ’terreng’ av attributt 7, medan ord for avlange objekt ikkje nødvendigvis har andre semantiske verdiar enn ’avlang’. Dei er karakteriserte nettopp ved denne verdien. Noko liknande som for ’avlang’ gjeld også for verdien ’hol’, ved at både kvinnelege kjønnsorgan og kroppshol (jf. STR 9) i tillegg til terrengfordjupingar (jf. STR 12), får denne verdien.

Elles finst verdien ’flate’ som refererer til flak og flater, og ’funksj_ holrom’. Med omsyn til det sistnemnde er det uklart kva ord Trosterud viser til, med unntak av døma i STR 28, *andlet, hjul, rør, øyre, åk*. Av den grunn har berre dei nemnde døma fått verdien ’funksj_holrom’.

Attributt 7: Stad.

Verdi	Tilsvarande tilordningsregel	Døme på substantiv
Terreng	STR 10, 11 og 12	dal, gard, grend, geysir, haug
Ikkje_perm	STR 13	fabrikk, hotell, brakke
Serv_stad	STR 14	bar, café, bistro
Samfunnsinst	STR 15	amt, arkiv, skole, klinikk

Tabell 4.7: Stad

Verdien 'ikkje_perm' vert tildelt "ord for bygningar som ikkje er permanente husvære for menneske" (Trosterud: 39). 'Serv_stad' står for 'servervingsstad', og refererer til "ord for bygningar der det vert servert mat og drikke" (Trosterud: 39). 'Samfunnsinst' er ei forkorting for 'samfunnsinstitusjon'. Med dette meiner ein kulturelt skapte stader og samfunnsinstitusjonar (jf. STR 15). Verdien 'terreng' refererer både til ord for geografiske stader generelt, ord for terrengfordjupingar, og ord for terrengtoppar og avlange terrengformasjonar (jf. STR 10, STR 11 og STR 12). Forskjellen mellom dei tre ulike typane av substantiv som vert tildelt denne verdien, kjem til syne ved at dei ulike typane terrengformasjonar får ulike verdiar av attributt 6 'form'. Terrengtoppar og avlange terrengformasjonar får verdien 'avlang' av attributt 6, medan terrengfordjupingar får verdien 'hol'. Ord for andre geografiske stader (jf. STR 10) får ingen verdi i attributt 6.

Attributt 8: Bøying.

Verdi	Tilsvarande tilordningsregel	Døme på substantiv
Omlydssubst	MTR 1	and, bok, mor
Uten_segm_form	MTR 2	ljos, yrke, hus

Tabell 4.8: Bøying

Omlydssubstantiv vert tildelt verdien ‚omlydssubst'. Substantiv som ikkje har segmental formativ i ubunden form fleirtal, får verdien ‚uten_segm_form'.

Attributt 9: Avleiing.

Verdi	Tilsvarande tilordningsregel	Døme på substantiv
Verb	MTR 5	aldring, erfaring
Subst	MTR 6	sogning
Adj	MTR 4, 6 og 7	blide, dumheit, ekling
Verbalsubst	MTR 3	hiv, dirr, gnål

Tabell 4.9: Avleiing

Verdien 'verb' vert gitt til substantiv på *–ing* som er avleidd av verb (jf. MTR 5), og verdien 'subst' vert gitt til substantiv på *–ing* som er avleidd av substantiv (jf. MTR 6). Både ord som består av adjektivstamme og *–e* (jf. MTR 4), og ord på *–ing* som er avleidde av adjektiv (jf. MTR 6), får verdien 'adj'. Desse vert skilde frå kvarandre ved at dei får henholdsvis verdiane 'E' (attributt 11) og 'ing' (attributt 10). Verbalsubstantiv av verbstammen får verdien 'verbalsubst'. Alle substantiv

som kjem av eit verb og har same form som verbstammen, har vorte rekna som verbalsubstantiv. Døme på verbalsubstantiv (frå *Nynorskordboka, 3. utgåva, 2001*):

II hiv n1 (av *hive)
1 det å hive; lyft (med vinsj, kran o l) ta fram sekker til luka og gjer klart til h- / last(mengd) som blir lyfta i ein gong (i vinsj, kran o l) den store krana svinga h- etter h- inn på bryggja
2 sleng, kast; fart det er h- i den karen
3 slingring, rulling, overhaling skøyta tok eit h- .

I hiv m1 (av *hive) i uttr: vere på ein h- el. på h-en vere litt drukken

dirr m1 (smh med *dirre) dirring han hadde ein liten d- i målet da han sa det

døyst n1 (jf *døyste) slag, støyt

drags n1 (sjå *dragse)

gnål n1 gnåling, mas slutt med dette g-et (om pengar)! .

Attributt 10: Suffiks.

Verdi	**Tilsvarande tilordningsregel**	**Døme på substantiv**
Ing	MTR 5 og 6	aldring, erfaring, sogning
Eri, ment, skop	MTR 8	bryggjeri, medikament, horoskop
Ine, inne, enne, esse, øse, ette	MTR 9	blondine, jødinne, komedienne, delikatesse, mitraljøse
Heit	MTR 7	godheit

Tabell 4.10: Suffiks

Verdiane av attributt 10 står for alle dei suffiksa som er relevante med omsyn til MTR 5-9. Det finst andre suffiks enn desse, som *–ling* og *–ning* (Beito, 1986 (1970), som er omtala med omsyn til genustilordning i diverse litteratur. Det kunne vore interessant å inkludere ein del av desse som verdiar. Men sidan dette er ei etterprøving av regelsettet til Trosterud, er berre suffiks som er inkluderte i hans reglar tekne omsyn til.

Attributt 11: Fonologi.

Verdi	**Tilsvarande tilordningsregel**	**Døme på substantiv**
C	STR 5	dyr, knapp, godheit
V	Regel 3	bru, kro
E	Regel 2, STR 6, STR 24, STR 25 og MTR 4	klokke, bukse, blondine

Tabell 4.11: Fonologi

Attributt 11 har tre fonologiske verdiar: Verdiane 'V' og 'e' finst på grunnlag av to overgripande fonologiske reglar, henholdsvis regel 3 og regel 2. 'V' vert gitt til einstava ord som endar på vokal, og 'e' vert gitt tostava ord som endar på trykklett *-e*. Denne trykklette *e*-en går igjen som ein faktor i fleire semantiske og ein morfologisk regel. På grunn av STR 5 (*ord på –C for kroppsdelar er n*) er det er óg naudsynt med ein verdi 'C' for ord som endar på konsonant.

Blondine er døme på eit substantiv der fonologi og morfologi overlappar. *Blondine* krev den fonologiske verdien 'E' (jf. tabell 4.11) og den morfologiske verdien 'ine' (jf. tabell 4.10). Alle ord som endar på *-ine* endar også på trykklett *–e.* Det finst altså ei ikkje tilfeldig overlapping av MTR 9 *(ord på –ine, -inne, -enne, -ette, -øse, -ette, er f)* og Regel 2 *(tostava ord på trykklett –e (svake substantiv) er f)*, som begge tilordnar femininum. Slike tilfelle av overlapping gjer det vanskeleg å vite kva kriterium som eigentleg tilordnar genus.

Attributt 12. Tredje siste bokstaven i substantivet:

Verdi	**Tilsvarande tilordningsregel**
Tredje siste bokstav i det aktuelle substantivet	FTR 1, FTR 2 og FTR 3

Tabell 4.12: Tredje siste bokstaven i substantivet

Attributt 13. Nest siste bokstaven i substantivet:

Verdi	**Tilsvarande tilordningsregel**
Andre siste bokstav i det aktuelle substantivet	FTR 1, FTR 2 og FTR 3

Tabell 4.13: Nest siste bokstaven i substantivet

Attributt 14. Siste bokstaven i substantivet:

Verdi	**Tilsvarande tilordningsregel**
Siste bokstav i det aktuelle substantivet	FTR 1, FTR 2 og FTR 3

Tabell 4.14: Siste bokstaven i substantivet

Attributt 12 -14 har grunnlag i dei fonologiske endingane i FTR 1, FTR 2 og FTR 3 (Trosterud, 2001). Desse reglane tilordnar genus etter visse fonologiske sluttsekvensar i substantiv. Eit mogleg alternativ er eit attributt 'fonologiske endingar' med desse fonologiske endingane som verdiar. Ein bakdel er at dette utelet andre moglege regelbundenheiter enn dei i FTR 1, FTR 2 og FTR 3. Dette kan gi misvisande resultat ved maskinlæring, og det vil vere meir fordelaktig om ein kan få med fonologiske endingar for alle substantiva i databasen. Det som er gjort er å, ved hjelp av sed-kommandoar i Unix, automatisk generere dei tre siste bokstavane i kvart substantiv og nytte desse som verdiar av tre ulike attributt; ein for den tredje siste, ein for den nest siste, og ein for den siste bokstaven i substantivet. Substantivet *jakke* vil då få verdiane 'k', 'k' og 'e' av attributt 12, 13 og 14. Substantiv på færre enn tre bokstavar vil ikkje verte tildelt nokon verdi av attributt 12, og eventuelt ingen verdi av attributt 13.

Problemet med denne tilnærmingsmåten er at ein del av dei fonologiske endingane i FTR 1, FTR 2 og FTR 3, er baserte på fonologiske eigenskapar som lengde, artikulasjonsmåte og artikulasjonsstad, noko som ikkje kjem fram av verdiane i attributt 12-14. Ei delvis løysing på nokre av problema kan vere å behandle fonem som bokstavar, altså gjere om kvar av endingane i dei fonologiske reglane til moglege bokstavsekvensar som kan samanliknast med verdiane i attributt 12, 13 og 14.

Tabellen nedanfor syner skjematisk kva verdi eit substantiv må ha av attributt 12, 13 og 14 for å samsvare med endingane i FTR 1, FTR 2 og FTR 3. Der rubrikken for attributt 12 er tom, er verdien av dette attributtet irrelevant med omsyn til den endinga det er snakk om.

Regel	**Genus i regel**	**Ending**	**Mogleg verdi av attributt 12**	**Mogleg verdi av attributt 13**	**Mogleg verdi av attributt 14**
FTR 1	F	-idd	i	d	d
FTR 1	F	-emd	e	m	d
FTR 1	F	-erd	e	r	d
FTR 1	F	-Cn		Konsonant*	n
FTR 1	F	-rg		r	g
FTR 1	F	-ft		f	t
FTR 1	F	-o:d		o	d
FTR 1	F	-vd		v	d
FTR 1	F	-pt		p	t
FTR 1	F	-kt		k	t
FTR 1	F	V[+høg]:l		i, e, o, u, y	l
FTR 1	F	-gd (for einstava ord)		g	d
FTR 1	F	-V: (for einstava ord)	Konsonant*, evnt. tom	Konsonant*, evnt. tom	Vokal*
FTR 2	N	-V:d		Vokal*	d
FTR 2	N	-om		o	m
FTR 2	N	-e:m		e	m
FTR 2	N	-um		u	m
FTR 2	N	-ym		y	m
FTR 2	N	-a:r		a	r
FTR 2	N	-ie:r	i	e	r
FTR 2	N	-iv		i	v
FTR 2	N	-a:t		a	t
FTR 3	M	-Rbe	l, m	b	e
FTR 3	M	-V:be	Vokal*	b	e

Tabell 5 : Samsvar mellom fonologiske reglar i Trosterud (2001), og automatisk genererte verdiar i databasen.

* Konsonant vil seie ein av alle moglege konsonantar: *b, c, d, f, g, h, j, k, l, m, n, p, q, r, s, t, v, w, x, z*

* Vokal vil seie ein av alle moglege vokalar. Vokalar med ulike aksentteikn vil bli sett på som eigne verdiar: *a, à, â, e, é, è, ê, i, ì, o, ò, ó, ô, u, ù, y, æ, ø, å*

Problemet med lengde vert ikkje løyst. Ei mogleg løysing er ein distinksjon mellom korte og lange vokalar som verdiar, men dette kan ikkje gjerast automatisk. Tal på stavingar er også eit problem, men berre med omsyn til einstava ord som endar på *-gd* (FTR 1). Einstava ord på *V:* (jf. FTR 1) er allereie merka i attributt 10 på grunn av regel 2 (*tostava ord på trykklett –e er f*). Å manuelt leggje til lengde og tal på stavingar for alle substantiva i databasen er for tidkrevjande. Endingar som *-erd,- pt* og *-kt* (jf. FTR 1), som tilsvarar nøyaktig dei automatisk genererte verdiane av attributt 12-14, kan samanliknast direkte med attributtverdiane. Endingar som *-V[+høg]:l* og *–Cn* (jf. FTR 1) derimot, kan ikkje samanliknast direkte med dei automatiak genererte verdiane, sidan *C* og *V[+høg]* ikkje vert genererert automatisk. *C* vert erstatta med ein einskild konsonant, til dømes *r*, og *V[+høg]* med ein av vokalane *i, e, o, u, y*. I slike tilfelle må ein leite etter grupper av aktuelle bokstavar i reglane som maskinlæringsalgoritmane genererer, for samanlikning med Trosterud sine reglar.

5.4. Genustilordning basert på fonologi i nederlandsk.

Genus i nederlandsk har vore sett på som meir eller mindre arbitrært. Ein del tilordningsreglar finst, men desse dekkjer på langt nær alle substantiv i språket. Daelemans og Durieux (2000) har utført ei rekkje eksperiment med minnebasert læring, med det mål å finne ut i kva grad fonologisk informasjon om substantiva bidreg til genustilordning i nederlandsk.

Nederlandsk har historisk sett eit genussystem med tre genus: maskulinum, femininum og nøytrum. Det skjer no eit skifte til to genus, med eit skilje mellom nøytrum og ikkje-nøytrum. Dette gjeld for artiklar, medan genus ved pronomen framleis skil mellom maskulinum og femininum. I Nederland skil ein mellom maskulinum og femininum berre ved antecedentar som denoterer menneske, medan i Flandern finst dette skiljet også for ikkje-menneskelege antecedentar.

Daelemans og Durieux utførte seks eksperiment, alle med IB1-IG (IG=*Information Gain*) og eit *leave-one-out* testsystem. Tre av eksperimenta (A1-A3) omfatta 6090 substantiv inndelt i klassene maskulinum, femininum og nøytrum. Dei tre andre (B1-B3) omfatta 7651 substantiv med klassene HET (nøytrum) og DE (ikkje-nøytrum). A1 og B1 inkluderte *onset, nucleus* og *coda* (byrjing, kjerne og slutt), for den siste stavinga, som attributt. A2 og B2 inkluderte i tillegg den same informasjonen for den første stavinga i substantivet. For eksperiment A3 og B3 vart i tillegg trykk og tal på stavingar inkludert. Tabellen nedanfor syner koding for substantivet *tafel*:

Eksperiment	**Klasse**	**Onset siste staving**	**Nucleus siste staving**	**Coda siste staving**	**Onset første staving**	**Nucleus første staving**	**Coda første staving**	**Trykk**	**Tal på stavingar**
A1	F	f	@	l					
A2	F	f	@	l	t	a	-		
A3	F	f	@	l	t	a	-	10	2
B1	DE	f	@	l					
B2	DE	f	@	l	t	a	-		
B3	DE	f	@	l	t	a	-	10	2

Tabell 6: Koding for *tafel* (Daelemans og Durieux: 20)

Resultata frå eksperimenta syner at auka fonologisk informasjon fører til fleire riktig klassifiserte tilfelle. Eksperiment A3 og B3 syner dei beste resultata med henholdsvis 84,3% og 87,65% riktig klassifiserte substantiv. Riktig klassifiserte substantiv ved A1-2 og B1-2 er rundt 1 % dårlegare. Daelemans og Durieux konkluderer med at desse resultata er betydeleg betre enn ein skulle tru etter påstander om at genus i nederlandsk er tilfeldig.

I lys av eksperimenta for nederlandsk kan det tenkjast at meir fonologisk informasjon ville føre til betre resultat også for liknande eksperiment i norsk. Ved ei etterprøving av reglane til Trosterud burde det med omsyn til FTR 1, FTR 2 og FTR 3 inkluderast tal på stavingar, lengde, ein vokal/konsonantdistinksjon, artikulasjonsstad og artikulasjonsmåte som attributt. På grunn av tidsmangel må eg nøye meg med ei best mogleg tilnærming, det vil seie dei tre siste bokstavane i kvart substantiv. Meir enn tre bokstavar er det ikkje behov for, ettersom databasen berre dekkjer informasjon som er nødvendig i forhold til Trosterud sine reglar.

6. Modellering og diskusjon av resultat.

6.1. Innputt til systema.

Som innputt ved klassifikasjon krevst det i tillegg til datasettet med eksempel ei oversikt over namn på klasser, attributt og verdiar. TiMBL genererer ei slik oversikt automatisk. Denne oversikta er innputt til C4.5. Klassene er lista opp først, og deretter alle attributta med tilhøyrande verdiar (a=attributt, ?=irrelevant attributt):

```
M,N,F.

a1ord:  ignore
a2diversesemantikk:
?,plante,bokstav,lyd,gram_kat,heimleg_tre,meieri,hyperonym,parvis.
a3menneskekroppen: ?,kroppsdel,sentral.
a4masse: ?,stoff,stein,jord.
a5biologiskkjønn: person,?,hankj,hokj,avkj.
a6form: ?,avlang,funksj_holrom,hol,flate.
a7stad: ?,samfunnsinst,terreng,ikkje_perm,serv_stad.
a8bøying: ?,uten_segm_form,omlydssubst.
a9avleiing: ?,subst,verb,verbalsubst,adj.
a10suffiks:
?,ing,ment,skop,inne,eri,ine,ette,esse,heit,adj,enne,oese.
a11fonologi: C,V,?,E.
a12tredjesistebokstav: a,-
,i,k,d,e,f,o,y,t,r,n,l,s,?,b,j,m,g,v,h,u,aa,p,oe,ae,c,z,w,o+,o-,e-
,x.
a13nestsistebokstav:
r,b,n,a,e,l,t,m,y,o,g,s,f,oe,p,?,d,i,j,k,u,aa,ae,h,v,c,e-
,z,w,o+,o-,o^,e^,x.
a14sistebokstav: k,u,g,r,s,i,t,f,e,l,m,a,aa,ae,d,n,v,c,o,e-
,p,b,y,oe,j,x,w,h,z,q.
```

C4.5 taklar ikkje *æ, ø, å* og vokalar med aksentteikn. Desse måtte endrast i datasettet, og dermed også i namneoversikta. Teikna som er nytta, alle i attributt 12-14, kan tolkast slik:

```
ae = æ
oe = ø
aa = å
e- = é
e^ = ê
o+ = ò
o- = ó
o^ = ô
```

RIPPER igjen taklar ikkje dei same teikna som C4.5 og TiMBL. RIPPER krev difor ei eiga namneoversikt:

```
M,N,F.
a1:  ignore.
a2:  bokstav, lyd, plante, gram_kat, heimleg_tre, meieri,
hyperonym, parvis.
a3:  kroppsdel, sentral.
a4:  stoff, stein, jord.
a5:  person, hankj, hokj, avkj.
a6:  funksj_holrom, avlang, hol, flate.
a7:  samfunnsinst, terreng, ikkje_perm, serv_stad.
a8:  uten_segm_form, omlydssubst.
a9:  subst, verb, verbalsubst, adj.
a10:  ing, ment, skop, inne, eri, ine, ette, esse, heit, adj,
enne, oese.
a11:  C, V, E.
a12:  a, _, i, k, d, e, f, o, y, t, r, n, l, s, b, j, m, g, v, h,
u, aa, p, oe,
ae, c, z, w, o_, e_, x.
a13:  r, b, n, a, e, l, t, m, y, o, g, s, f, oe, p, d, i, j, k, u,
aa, ae, h, v,
c, e_, z, w, o_, o1, e1, x.
a14:  k, u, g, r, s, i, t, f, e, l, m, a, aa, ae, d, n, v, c, o,
e_, p, b, y,
oe, j, x, w, h, z, q.
```

Teikna som er ulike i forhold til i innputt for TiMBL og C4.5 (alle i a12-14), kan tolkast slik:

```
e_ = é
e1 = ê
o_ = ò
o_ = ó
o1 = ô
```

Det er mogleg å la C4.5 og RIPPER ignorere attributt ved å merke gjeldande attributt med ”ignore” i namneoversikta. TiMBL ignorerer attributt ved opsjonen ”I”. Attributt 1 i treningsettet refererer til sjølve substantivet, og er inkludert i treningsdata av av praktiske årsaker. Substantiva skal ikkje nyttast som attributtverdiar ved klassifikasjon, og attributt 1 vert difor alltid ignorert. Attributt 2-14 kan inkluderast eller ignorerast etter ønske. Vi er interesserte i å lage ulike klassifikatorar ved å ignorere ulike typar informasjon om substantiva. Feilratane til dei ulike klassifikatorane kan syne i kva grad ulike typar informasjon, som semantikk, morfologi og fonologi, bidreg ved genustilordning. Kapittel 7.1 samanliknar resultat for ei rekkje slike eksperiment.

6.2. Klassifikator 1: C4.5RULES.

Det beste resultatet for C4.5RULES, med 7.5% feilrate (1008 av 13384 tilfelle er feilklassifiserte), vart oppnådd ved å inkludere alle attributta, med andre ord attributt 2-14. Først vart eit beslutningstre med 6018 greiner generert av C4.5, og eit forenkla tre med 157 greiner danna ut i frå dette. Ut i frå det forenkla beslutningstreet genererer C4.5RULES eit regelsett på 12 reglar i tillegg til default:

```
Rule 2:
     a8bøying = uten_segm_form
     ->  class N  [98.2%]

Rule 129:
     a5biologiskkjønn in {person, hankj}
     ->  class M  [97.5%]

Rule 15:
     a2diversesemantikk in {plante, bokstav, meieri}
     a14sistebokstav in {k, r, s, i, t, f, l, m, a, ae, d, n, v,
c, o, e-, p, b,
y, oe, j, x, w, h, z, q}
     ->  class M  [88.8%]

Rule 159:
     a9avleiing = subst
     ->  class M  [86.1%]

Rule 55:
     a12tredjesistebokstav in {a, i, k, d, e, f, o, y, t, r, n, l,
s, b, j, m,
g, v, h, u, aa, p}
     a13nestsistebokstav = b
     a14sistebokstav in {k, r, s, i, t, f, l, m, a, ae, d, n, v,
c, o, e-, p, b,
y, oe, j, x, w, h, z, q}
     ->  class M  [81.5%]

Rule 37:
     a12tredjesistebokstav in {a, i, k, d, e, f, y, t, n, l, s, b,
j, m, g, v,
u, p}
     a13nestsistebokstav in {r, n, a, e, t, m, y, o, g, s, f, oe,
p, d, k, u,
ae}
     a14sistebokstav in {k, r, s, i, t, f, n, o, p}
     ->  class M  [79.0%]

Rule 48:
     a12tredjesistebokstav in {a, e, o, r, l}
     a13nestsistebokstav = l
     a14sistebokstav in {k, r, s, i, t, f, l, m, n, v, o, e-, p,
y}
     ->  class M  [70.5%]

Rule 83:
     a5biologiskkjønn = hokj
     ->  class F  [96.0%]
```

```
Rule 120:
     a9avleiing = verb
     ->  class F  [95.7%]

Rule 90:
     a8bøying = omlydssubst
     a14sistebokstav in {k, r, s, i, t, f, l, m, a, ae, d, n, v,
c, o, e-, p, b,
y, oe, j, x, w, h, z, q}
     ->  class F  [93.7%]

Rule 46:
     a2diversesemantikk = heimleg_tre
     ->  class F  [89.1%]

Rule 114:
     a12tredjesistebokstav in {i, k, d, e, f, o, y, t, r, n, l, s,
b, j, m, g,
v, u, aa, p, oe, ae, o+}
     a13nestsistebokstav in {r, b, n, a, l, t, m, y, o, g, s, f,
oe, p, d, i, j,
k, u, v}
     a14sistebokstav in {u, e, aa}
     ->  class F  [67.1%]

Default class: M
```

Nokre av reglane (til dømes regel 90) inneheld testar med grupper av attributtverdiar i staden for ein enkel verdi. Dette kjem av at opsjonen *–s* (Quinlan: 85) er nytta. Denne grupperer attributtverdiar for testar i eit beslutningstre. Utan denne opsjonen hadde C4.5 generert ei separat grein og eit separat subtre for kvar moglege verdi av eit attributt. Reglar genererte frå eit slikt beslutningstre hadde vore mange og svært spesifikke. Regel 90 hadde til dømes vore oppdelt i mange små reglar. Det same gjeld ein del andre reglar, særleg dei med føresetnader for attributt 12-14, som inneheld ei lang rekkje verdiar. Dette hadde komplisert regelsettet, og gruppering av verdiar er difor ønskjeleg.

Utputt frå C4.5RULES består i tillegg til reglane av ei statistisk evaluering av kvar regel, ei samla evaluering av reglane (samla feilrate), og ei *forvirringsmatrise* som syner distribusjon av feilklassifikasjon i datasettet.

6.2.1. Reglane.

Ein del av reglane som C4.5RULES genererer, samsvarar med tilordningsreglar i Trosterud (2001). Defaultregelen er den same som hos Trosterud:

```
Default class: M
```

Nummera på reglane er tilfeldige, dei kjem frå rekkjefølgja til blada i beslutningstreet og fungerer berre som identifikasjon av reglane (Quinlan, 1992). Den første regelen, regel 2, ser slik ut:

```
Rule 2:
     a8bøying = uten_segm_form
     ->  class N  [98.2%]
```

Venstresida til regelen inneheld ein test som seier at attributt 8 må ha verdien 'uten_segm_form'. Eit tilfelle i treningssettet, som tilfredsstiller denne testen, får klasse N, gitt av høgresida i regelen. Programmet predikerer at 98,2% av tilfella som tilfredsstiller venstresida i denne regelen, vert klassifiserte som N. Regel 2 tilsvarar MTR 2 hos Trosterud (*substantiv utan segmental formativ i ub.pl. er n*). Systemet genererer i tillegg til regel 2, tre andre reglar som tilsvarar heilt eller delvis morfologiske tilordningsreglar hos Trosterud:

```
Rule 159:
     a9avleiing = subst
     ->  class M  [86.1%]

Rule 120:
     a9avleiing = verb
     ->  class F  [95.7%]

Rule 90:
     a8bøying = omlydssubst
     a14sistebokstav in {k, r, s, i, t, f, l, m, a, ae, d, n, v,
c, o, e-, p, b,
y, oe, j, x, w, h, z, q}
     ->  class F  [93.7%]
```

Regel 159 uttrykkjer noko av det same som MTR 6 (*ord på –ing som er avleidde av substantiv og adjektiv, er m).* MTR 6 inkluderer ord avleidd av både substantiv og adjektiv, medan regel 159 inkluderer berre ord avleidde av substantiv. Regel 159 uttrykkjer ikkje eksplisitt at det er ord på –*ing* det er snakk om, men dette seier seg sjølv, sidan berre desse orda har fått tildelt verdien 'subst'. Med omsyn til Trosterud sine reglar er ikkje denne verdien relevant for andre enn ord med suffikset –*ing*. Det same gjeld regel 120, som tilsvarar MTR 5 (*ord på -ing som er avleidd av verb, er f*): Berre ord på –*ing* har fått verdien 'verb', og regel 120 kan difor berre gjelde desse. Regel 90 samsvarar til ei viss grad med MTR 1 (*omlydssubstantiv er f*), ved at han gir 'omlydssubst' som eit kriterium. I tillegg inkluderer regelen krav for den siste bokstaven i substantivet.

Det vert også generert ein del reglar som samsvarar med semantiske tilordningsreglar hos Trosterud:

```
Rule 129:
     a5biologiskkjønn in {person, hankj}
     ->  class M  [97.5%]

Rule 83:
     a5biologiskkjønn = hokj
     ->  class F  [96.0%]

Rule 46:
     a2diversesemantikk = heimleg_tre
     ->  class F  [89.1%]
```

Regel 46 uttrykkjer det same som Trosterud sin STR 23 (*ord for heimlege tre er f*). Regel 129 og 83 formulerer til saman det same som STR 1 (*der det er ulike ord for referantane med ulikt biologisk kjønn, har orda tilsvarande genus*) og STR 2 (*ord som refererer til person utan kjønnsspesifisering, er m*). Innhaldet i STR 1 vert uttrykt både i regel 129 og 83, medan innhaldet i STR 2 vert uttrykt i regel 129.

Dei gjenståande reglane som C4.5RULES genererer, er ikkje like eintydige med omsyn til samsvar med Trosterud sine reglar. Dette gjeld mellom andre regel 15:

```
Rule 15:
     a2diversesemantikk in {plante, bokstav, meieri}
     a14sistebokstav in {k, r, s, i, t, f, l, m, a, ae, d, n, v,
c, o, e-, p, b,
y, oe, j, x, w, h, z, q}
     ->  class M  [88.8%]
```

Regel 15 tilordnar maskulint genus til bokstavar, planter og meieriprodukt, og kan såleis likne på STR 21 (*ord som refererer til bokstavar, er m)*, STR 24 *(ord på –e for plantar er m)* og STR 25 *(ord på –e for mjølkeprodukt er m).* Regel 15 set også visse krav for den siste bokstaven i substantivet. Men ikkje alle bokstavane i alfabetet er lista opp som verdiar i testen for attributt 14, sjølv om ein skulle tru det når alle finst i datasettet, med verdien 'bokstav'. Det finst element i regel 15 som totalt motseier STR 24 og 25. Desse uttrykkjer henholdsvis at planter og mjølkeprodukt på trykklett *–e* er maskuline, medan *e* ikkje er inkludert blant krava for siste bokstaven i substantivet i regel 15.

Fire av reglane har ingen samsvar med eller likskap med nokon av reglane i Trosterud sitt regelsett.

```
Rule 55:
     a12tredjesistebokstav in {a, i, k, d, e, f, o, y, t, r, n, l,
s, b, j, m,
g, v, h, u, aa, p}
     a13nestsistebokstav = b
     a14sistebokstav in {k, r, s, i, t, f, l, m, a, ae, d, n, v,
c, o, e-, p, b,
y, oe, j, x, w, h, z, q}
     ->  class M  [81.5%]

Rule 37:
     a12tredjesistebokstav in {a, i, k, d, e, f, y, t, n, l, s, b,
j, m, g, v,
u, p}
     a13nestsistebokstav in {r, n, a, e, t, m, y, o, g, s, f, oe,
p, d, k, u,
ae}
     a14sistebokstav in {k, r, s, i, t, f, n, o, p}
     ->  class M  [79.0%]
```

```
Rule 48:
     a12tredjesistebokstav in {a, e, o, r, l}
     a13nestsistebokstav = l
     a14sistebokstav in {k, r, s, i, t, f, l, m, n, v, o, e-, p,
y}
     ->  class M  [70.5%]

Rule 114:
     a12tredjesistebokstav in {i, k, d, e, f, o, y, t, r, n, l, s,
b, j, m, g,
v, u, aa, p, oe, ae, o+}
     a13nestsistebokstav in {r, b, n, a, l, t, m, y, o, g, s, f,
oe, p, d, i, j,
k, u, v}
     a14sistebokstav in {u, e, aa}
     ->  class F  [67.1%]
```

Desse fire reglane gir ei rekkje fonologiske kriterium for genustilordning, men kan vanskeleg samanliknast med FTR1, FTR 2 eller FTR 3 hos Trosterud, sidan dei fonologiske endingane er erstatta med einskilde bokstavar. Dei moglege samansetnadane av verdiar av attributt 12, 13 og 14, tilsvarar i kvar regel ei lang rekkje fonologiske endingar. Dei av reglane i regelsett 1, som tydeleg tilsvarar reglar hos Trosterud, tilsvarar semantiske, morfologiske eller overgripande fonologiske reglar.

6.2.2. Evaluering av reglane.

Systemet evaluerer regelsettet som vert generert i forhold til treningsdata, og evalueringa av kvar einskild regel vert sett opp i ein tabell:

Evaluation on training data (13384 items):

Rule	Size	Error	Used	Wrong		Advantage		
2	1	1.8%	2679	44	(1.6%)	2595	(2635|40)	N
129	1	2.5%	2262	41	(1.8%)	71	(112|41)	M
15	2	11.2%	136	5	(3.7%)	0	(0|0)	M
159	1	13.9%	47	8	(17.0%)	0	(0|0)	M
55	3	18.5%	45	0	(0.0%)	0	(0|0)	M
37	3	21.0%	3053	143	(4.7%)	-28	(3|31)	M
48	3	29.5%	167	11	(6.6%)	-3	(0|3)	M
83	1	4.0%	220	0	(0.0%)	39	(39|0)	F
120	1	4.3%	764	17	(2.2%)	730	(747|17)	F
90	2	6.3%	35	2	(5.7%)	31	(33|2)	F
46	1	10.9%	10	0	(0.0%)	5	(5|0)	F
114	3	32.9%	1699	353	(20.8%)	998	(1346|348)	F

Statistikken for den første regelen, regel 2, skal ut i frå tabellen tolkast slik: (Quinlan, 1992) Regelen si venstreside inneheld ein test (*size*=1), og den predikerte feilrata ligg på 1.8% (*error*). Regelen vart nytta 2679 gongar ved klassifikasjon av treningssettet (*used*). 44 av tilfella (1.6%),

som tilfredsstilte regelen si venstreside, vart feilklassifiserte (*wrong*). At *advantage* er 2595 (2635|40), vil seie: Om regel 2 hadde vore utelaten, hadde 2635 tilfelle som no er klassifiserte riktig av denne regelen, vorte feilklassifiserte, og 40 tilfelle som no er feilklassifiserte av regel 2, hadde vorte riktig klassifiserte av andre reglar inkludert defaultregelen. *Nettofordelen* av å nytte denne regelen er då 2635-40=2595. For regel 37 og 48 er nettofordelen henholdsvis -28 og -3, det vil seie at ein taper på å inkludere desse reglane i klassifikasjonsmodellen. Systemet har difor kutta dei ut, og evaluert dei andre reglane på nytt. Resultatet av den nye evalueringa vert gitt i ein tabell som er litt ulik den første tabellen:

Rule	Size	Error	Used	Wrong		Advantage	
2	1	1.8%	2679	44	(1.6%)	2595 (2635|40)	N
129	1	2.5%	2262	41	(1.8%)	71 (112|41)	M
15	2	11.2%	136	5	(3.7%)	0 (0|0)	M
159	1	13.9%	47	8	(17.0%)	0 (0|0)	M
55	3	18.5%	45	0	(0.0%)	0 (0|0)	M
83	1	4.0%	237	3	(1.3%)	45 (48|3)	F
120	1	4.3%	764	17	(2.2%)	730 (747|17)	F
90	2	6.3%	53	2	(3.8%)	49 (51|2)	F
46	1	10.9%	12	0	(0.0%)	7 (7|0)	F
114	3	32.9%	1699	353	(20.8%)	998 (1346|348)	F

For regel 15, 159 og 55 er fordelen 0 (0|0). Grunnen til at ein regel får 0 (0|0) i fordel, er at det ikkje finst tilfelle i datasettet som tilfredsstiller venstresida i både denne regelen og ein annan regel. Det vil seie at alle tilfelle som regel 15, 159 og 55 no dekkjer, ville verte klassifiserte av defaultregelen om desse reglane ikkje var inkluderte i regelsettet. Dei tre reglane som vert nytta på flest tilfelle i treningssettet, er regel 2, 120 og 114. Både regel 2 og 120 har ei låg feilrate og ein stor nettofordel. Dei fleste av tilfella som no er klassifiserte av desse, hadde vorte feilklassifiserte om dei vart utelatne. Heile 20.8% vert feilklassifiserte av regel 114, men fordelen med å inkludere denne regelen er så stor (998), at ein ikkje vil utelate han. Regel 2 klassifisererer nøytrumssubstantiv, medan regel 120 og 114 klassifiserer feminine substantiv. Dei tre reglane som har ein nettofordel på 0, klassifiserer maskuline substantiv. Reglar som gir ei anna klasse enn default, vil ofte ha ein større fordel, i og med at defaultregelen aldri vil dekkje dei same tilfella som ein slik regel.

Av dei ti reglane som står att etter at regel 38 og 47 er utelatne, tilsvarar seks tilordningsreglar hos Trosterud; tre morfologiske reglar, MTR 2, MTR 5 og (ein del av) MTR 6, og tre semantiske reglar, STR 1, STR 2 og STR 23. Dei ti reglane i regelsett 1, i tillegg til defaultregelen, klassifiserer riktig 92,5% av substantiva i treningssettet. På grunn av at regel 15, 159 og 55 har ein nettofordel på 0, slik at defaultregelen hadde dekt tilfelle som no er dekte av desse om dei var utelatne, hadde eit regelsett utan regel 15, 159 og 55, altså åtte reglar i tillegg til default, hatt den same feilraten på 7,5%. Trosterud sitt regelsett på heile 42 reglar i tillegg til defaultregel, klassifiserer om lag 94% av substantiva i nynorsk, altså berre 1,5% meir enn regelsett 1 på ti (evnt. åtte) reglar. Dette syner at mange av Trosterud sine tilordningsreglar er overflødige, og at eit regelsett med langt færre reglar kan klassifisere ein nesten like stor del av nynorske substantiv. Eit optimalt regelsett vil vere eit regelsett med færrast moglege reglar, som klassifiserer riktig flest moglege substantiv. Når feilraten går opp med berre 1,5% ved å nytte eit regelsett med om lag fjerdeparten så mange reglar som Trosterud set opp, kan ein seie at regelsett 1 er meir effektivt enn Trosterud sitt.

Ulempen ved regelsett 1 er at nokre av reglane er for kompliserte til at dei har nokon praktisk nytteverdi. Regel 15, 55 og 114 (regel 37 og 48 er sett bort i frå fordi dei er utelatne av systemet)

ramsar opp ei lang rekkje verdiar av attributt 12, 13 og 14 som føresetnader for genustilordning. Desse reglane er lite intuitive og kompliserer regelsettet.

6.2.3. Distribusjon av feilklassifikasjon i treningsdata.

C4.5RULES genererer ei forvirringsmatrise som syner korleis feilklassifikasjon er distribuert i treningsdata:

```
(a)   (b)   (c) <-classified as
----  ----  ----
7351    40   370     (a): class M
  15  2635     5     (b): class N
 574     4  2390     (c): class F
```

Vertikalt syner matrisa dei riktige klassene, og horisontalt klassene som er gitt av klassifikatoren. Av eksempla som skal vere i klasse M, er 7351 riktig klassifiserte av systemet, medan 40 har fått tildelt klasse N og 370 er tildelt klasse F. 2635 tilfelle i klasse N er riktig klassifiserte, og henholdsvis 15 og 5 tilfelle er feilklassifiserte som M og F. Av eksempla i klasse F har 2390 tilfelle fått tildelt riktig klasse, medan 574 er feilklassifiserte som M, og 4 som N. Ei slik matrise kan gi nyttig informasjon om distribusjonen av feilklassifikasjon i treningssettet. I dette tilfellet er størstedelen av dei feilklassifiserte tilfella maskuline substantiv som er klassifiserte som feminine, og feminine som er klassifiserte som maskuline. Nøytrumssubstantiv er stort sett riktig klassifiserte, og få feminine og maskuline substantiv er feilklassifiserte som nøytrumssubstantiv. Som vi snart vil sjå, har dette samanheng med at alle nøytrumssubstantiv får verdien 'uten_segm_form'. Sidan svært få substantiv som ikkje er nøytrum får denne verdien, genererer systemet regel 16, som klassifisiserer dei fleste av nøytrumssubstantiva i datasettet riktig.

Verdien 'uten_segm_form' er inkludert på grunn av Trosterud sin MTR 2, ein regel det er grunn til å stille spørsmål ved. Ingen nøytrumssubstantiv i normalisert nynorsk har segmental formativ i ubunden form fleirtal. Ein regel som MTR 2, som seier at substantiv utan segmental formativ i ubunden form fleirtal er nøytrum, er problematisk fordi ein like godt kan snu han på hovudet og seie at "nøytrumssubstantiv vert bøygd utan segmental formativ i ubunden form fleirtal". I staden for ein genustilordningsregel, blir dette ein regel for bøying av nøytrumsord i fleirtal. Trosterud påpeikar sjølv problematikken i samanheng med MTR 2, men har likevel valt å inkludere denne regelen. Sidan dette er ei etterprøving av Trosterud sitt regelsett, kan ein ikkje sjå bort i frå MTR 2, og attributtverdien 'uten_segm_form' er inkludert. Men det har i tillegg vorte utført eksperiment der denne verdien er ignorert. Når 'uten_segm_form' er inkludert, blir det berre generert ein einaste regel for klassifikasjon av nøytrumsord, nemleg regel 2. Når ein utelet 'uten_segm_form', må systemet nytte andre kriterium for at eit substantiv skal klassifiserast som N, og må generere andre reglar for tilordning av nøytrum. Ein vil også få svar på om dette fører til at betydeleg færre av substantiva vert riktig klassifiserte, og såleis om MTR 2 er ein viktig grunn til at Trosterud sitt regelsett klassifiserer ein så stor del av nynorske substantiv som det gjer.

6.3. Klassifikator 2: Bøyingsmorfologi er ignorert.

Om ein utelet verdien 'uten_segm_form', må også 'omlydssubst' utelatast, sidan desse to er verdiar av same attributt, og systemet ikkje tillet ignorering av berre ein verdi av eit attributt. Heile attributt 8 må difor utelatast. Sidan regel 90 i regelsett 1, som tilordnar omlydsubstantiv feminint genus, vert

nytta på berre 53 tilfelle i treningssettet, vil det ikkje gjere store utslag på resultatet om verdien 'omlydssubst' vert ignorert.

Når attributt 8 vert ignorert må C4.5RULES generere ei rekkje reglar for tilordning av nøytrumsord, i staden for berre ein. Heile regelsettet består av 19 reglar i tillegg til default:

```
Rule 173:
     a5biologiskkjønn = avkj
     ->  class N  [88.9%]

Rule 64:
     a2diversesemantikk in {lyd, gram_kat, hyperonym}
     a14sistebokstav in {k, r, s, i, t, f, l, m, a, ae, d, n, v,
c, o, e-, p, b,
y, oe, j, x, w, h, z, q}
     ->  class N  [81.9%]

Rule 9:
     a12tredjesistebokstav in {k, f, t, r, n, l, s, b, m, v, h, c}
     a13nestsistebokstav in {a, e, y, o, oe, i, u, aa, o-}
     a14sistebokstav in {m, d, v}
     ->  class N  [70.0%]

Rule 190:
     a9avleiing = verbalsubst
     ->  class N  [67.8%]

Rule 85:
     a13nestsistebokstav in {a, e, u}
     a14sistebokstav in {m, a, v}
     ->  class N  [67.5%]

Rule 31:
     a12tredjesistebokstav = e
     a13nestsistebokstav = n
     a14sistebokstav in {k, t}
     ->  class N  [48.5%]

Rule 158:
     a13nestsistebokstav in {a, e, o}
     a14sistebokstav in {u, g, e, aa}
     ->  class N  [44.3%]

Rule 20:
     a13nestsistebokstav = e
     a14sistebokstav in {k, r, s, i, t, l, m, d, n, v, o, p, y}
     ->  class N  [34.7%]
```

```
Rule 26:
     a5biologiskkjønn = hokj
     ->  class F  [96.0%]

Rule 183:
     a9avleiing = verb
     ->  class F  [95.7%]

Rule 52:
     a2diversesemantikk = heimleg_tre
     ->  class F  [89.1%]

Rule 159:
     a14sistebokstav = e
     ->  class F  [65.8%]

Rule 116:
     a11fonologi = V
     a13nestsistebokstav in {r, n, e, l, t, m, y, o, g, s, f, oe,
p, d, i, j, k,
u, aa, ae, h, v}
     ->  class F  [42.6%]

Rule 147:
     a5biologiskkjønn in {person, hankj}
     ->  class M  [97.5%]

Rule 17:
     a2diversesemantikk in {plante, bokstav, meieri}
     a14sistebokstav in {k, r, s, i, t, f, l, m, a, ae, d, n, v,
c, o, e-, p, b,
y, oe, j, x, w, h, z, q}
     ->  class M  [88.8%]

Rule 180:
     a9avleiing = subst
     ->  class M  [86.1%]

Rule 103:
     a12tredjesistebokstav in {a, i, k, f, o, y, n, l, s, b, j, v,
p}
     a13nestsistebokstav in {n, a, l, t, m, o, s, p, i, k, u}
     a14sistebokstav in {k, r, i, t, l, n, p}
     ->  class M  [80.9%]
```

```
Rule 43:
     a12tredjesistebokstav in {a, i, k, d, e, f, y, t, n, l, s, b,
j, m, g, v,
u, p}
     a13nestsistebokstav in {r, n, a, e, t, m, y, o, g, s, f, oe,
p, d, k, u,
ae}
     a14sistebokstav in {k, r, s, i, t, f, n, o, p}
     ->  class M  [79.0%]

Rule 54:
     a12tredjesistebokstav in {a, e, o, r, l}
     a13nestsistebokstav = l
     a14sistebokstav in {k, r, s, i, t, f, l, m, n, v, o, e-, p,
y}
     ->  class M  [70.5%]

Default class: M
```

Feilraten aukar mykje når attributt 8 er inkludert. Den totale feilraten ligg no på 24%, det vil seie at 3216 av 13384 substantiv vert feilklassifiserte. Forvirringsmatrisa som systemet genererer, syner at ein stor del av tilfella som er feilklassifiserte, er nøytrumssubstantiv:

```
  (a)   (b)   (c)     <-classified as
 ----  ----  ----
 6824   363   574     (a): class M
 1455   878   322     (b): class N
  474    28  2466     (c): class F
```

1455 substantiv som skulle ha vore klassifiserte som N, har fått klasse M, og 322 tilfelle som skulle ha fått klasse N, har vorte klassifiserte som F. Berre 878 av 2655 nøytrumsord er klassifiserte riktig, i forhold til 2635 av 2655 av klassifikator 1. Tala for feilklassifikasjon av feminine og maskuline substantiv til klasse N har også auka.

Ved å ta reglane og evalueringa nærare i augnesyn, ser ein at det er fråveret av attributt 8, og framfor alt verdien 'uten_segm_form', som har ført til den store auken i feilklassifikasjon. Når det gjeld reglar som tilordnar klasse F og M til substantiv, er desse mykje dei same som i regelsett 1. Reglane for tilordning av femininum er om lag dei same, med unntak av at klassifikator 2 ikkje kan generere ein regel for tilordning av femininum til omlydssubstantiv. Det vert derimot generert to andre reglar for tilordning av femininum:

```
Rule 159:
     a14sistebokstav = e
     ->  class F  [65.8%]
```

```
Rule 116:
     a11fonologi = V
     a13nestsistebokstav in {r, n, e, l, t, m, y, o, g, s, f, oe,
p, d, i, j, k,
u, aa, ae, h, v}
     ->  class F  [42.6%]
```

Regel 159 liknar på regel 2 (*tostava ord på trykklett –e (svake substantiv) er f*) hos Trosterud. Men regel 159 dekkjer ikkje nødvendigvis berre svake substantiv, sidan trykklett *–e* er spesifisert som ein eigen verdi 'E' av attributt 11. Om regelen hadde dekt berre svake substantiv hadde 'E' vore sett som krav. Regel 116 minner om Trosterud sin regel 3 *(einstava ord på vokal er f)*, men regel 116 spesifiserer i tillegg ei rekkje verdiar for den nest siste bokstaven i substantivet som kriterium.

Dei to regelsetta inneheld dei same reglane for tilordning av maskulinum, med unntak av at regelsett 2 inkluderer ein ekstra regel:

```
Rule 54:
     a12tredjesistebokstav in {a, e, o, r, l}
     a13nestsistebokstav = l
     a14sistebokstav in {k, r, s, i, t, f, l, m, n, v, o, e-, p,
y}
     ->  class M  [70.5%]
```

I tillegg finst det små avvik mellom dei to regelsetta med omsyn til kva verdiar av attributt 12-14 som er inkluderte som kriterium i ulike reglar.

Regel 2 i regelsett 1 er erstatta med åtte andre reglar som gir klasse N. Ein del av dei samsvarar med nokre av Trosterud sine tilordningsreglar for nøytrum:

```
Rule 173:
     a5biologiskkjønn = avkj
     ->  class N  [88.9%]

Rule 64:
     a2diversesemantikk in {lyd, gram_kat, hyperonym}
     a14sistebokstav in {k, r, s, i, t, f, l, m, a, ae, d, n, v,
c, o, e-, p, b,
y, oe, j, x, w, h, z, q}
     ->  class N  [81.9%]

Rule 190:
     a9avleiing = verbalsubst
     ->  class N  [67.8%]
```

Regel 173 tilsvarar STR 3 hos Trosterud (*ord som refererer til personar og har nedsetjande tyding, og ord for avkjønna vesen, er n*). Regel 64 inneheld dei same semantiske føresetnadene som STR 20 (*ord for lydar og språkhandlingar, og for sitatord, er n)*, STR 26 *(ord for grammatiske kategoriar er n)* og STR 4 (*hyperonym er n)*. Regel 64 tilordnar som desse nøytrum. Men i motsetnad til desse reglane inneheld regel 64 også krav til den siste bokstaven i substantivet. Regel 190 tilsvarar Trosterud sin MTR 3 (*verbalsubstantiv av verbstammen er n*).

Fem av reglane for tilordning av nøytrum i regelsett 2, er reglar som inneheld krav for dei tre siste bokstavane i substantivet. Desse er som nemnt ikkje særleg samanliknbare med Trosterud sine spesifikke fonologiske reglar:

```
Rule 9:
     a12tredjesistebokstav in {k, f, t, r, n, l, s, b, m, v, h, c}
     a13nestsistebokstav in {a, e, y, o, oe, i, u, aa, o-}
     a14sistebokstav in {m, d, v}
     ->  class N  [70.0%]

Rule 85:
     a13nestsistebokstav in {a, e, u}
     a14sistebokstav in {m, a, v}
     ->  class N  [67.5%]

Rule 31:
     a12tredjesistebokstav = e
     a13nestsistebokstav = n
     a14sistebokstav in {k, t}
     ->  class N  [48.5%]

Rule 158:
     a13nestsistebokstav in {a, e, o}
     a14sistebokstav in {u, g, e, aa}
     ->  class N  [44.3%]

Rule 20:
     a13nestsistebokstav = e
     a14sistebokstav in {k, r, s, i, t, l, m, d, n, v, o, p, y}
     ->  class N  [34.7%]
```

Systemet finn ved evaluering av reglane på treningsdata at regel 158 og regel 20 har negative nettofordelar, og utelet difor desse reglane. Ei ny evaluering vert gjort, og tabellen nedanfor syner resultatet:

Rule	Size	Error	Used	Wrong		Advantage	
173	1	11.1%	34	2	(5.9%)	30 (31|1)	N
64	2	18.1%	259	42	(16.2%)	74 (105|31)	N
9	3	30.0%	354	111	(31.4%)	85 (140|55)	N
190	1	32.2%	444	153	(34.5%)	146 (289|143)	N
85	2	32.5%	51	20	(39.2%)	14 (31|17)	N
31	3	51.5%	127	63	(49.6%)	3 (64|61)	N
26	1	4.0%	238	6	(2.5%)	37 (40|3)	F
183	1	4.3%	776	29	(3.7%)	718 (747|29)	F
52	1	10.9%	12	0	(0.0%)	8 (8|0)	F
159	1	34.2%	2191	790	(36.1%)	894 (1399|505)	F
116	2	57.4%	145	71	(49.0%)	38 (74|36)	F
147	1	2.5%	2064	7	(0.3%)	0 (0|0)	M
17	2	11.2%	136	9	(6.6%)	0 (0|0)	M
180	1	13.9%	47	8	(17.0%)	0 (0|0)	M
103	3	19.1%	2236	503	(22.5%)	0 (0|0)	M
43	3	21.0%	1986	540	(27.2%)	0 (0|0)	M
54	3	29.5%	104	23	(22.1%)	0 (0|0)	M

Alle reglane for tilordning av maskulinum har ein nettofordel på 0. Desse kan difor utelatast utan at feilraten går ned, og substantiv som desse reglane ville ha dekt, vil få tilordna maskulint genus av defaultregelen. Regelsettet kan såleis reduserast frå 17 til 11 reglar.

Ut i frå resultata frå klassifikator 1 og 2, kan ein slutte at bøyingsmorfologi bidreg i vesentleg grad til at Trosterud sitt forslag til regelsett tilordnar genus til ein så stor del av nynorske substantiv som 94 %. Sidan MTR 2 vanskeleg kan forsvarast som ein gyldig genustilordningsregel, må eit regelsett utan denne vurderast som eit betre alternativ. Nokre av nøytrumssubstantiva kan tilordnast nøytrum av alternative reglar, men desse reglane vil ikkje dekkje alle nøytrumssubstantiv, og resultatet er eit regelsett med langt lågare dekningsgrad enn 94 %.

6.4. Klassifikator 3: RIPPER med bøyingsmorfologi.

For eit betre grunnlag for verifisering av slutningane trekt i 6.3, vart også RIPPER testa både med og utan bøyingsmorfologi. Om RIPPER syner liknande resultat med omsyn til reglar og feilratar, forsterkar dette konklusjonen i 6.3, om viktigheita av bøyingsmorfologi for tilordning av genus til ein stor prosent av nynorske substantiv.

Når bøyingsmorfologi er inkludert (klassifikator 3), genererer RIPPER i tillegg til default 10 reglar, alle for tilordning av femininum og nøytrum:

```
N :- a8=uten_segm_form (2635/44).
F :- a11=E (1586/498).
F :- a9=verb (747/29).
F :- a11=V (86/55).
F :- a8=omlydssubst (57/3).
F :- a14=d, a12=i (10/6).
F :- a5=hokj (34/3).
```

```
F :- a14=d, a13=l, a12=oe (3/0).
F :- a14=d, a12=y (9/2).
F :- a7=terreng, a6=hol (16/4).
default M (7127/430).
```

Feilraten er 8.02%, altså litt større enn for C4.5RULES. Defaultklassa er sett til M, og systemet genererer difor i motsetnad til C4.5RULES, ingen reglar for tilordning av maskulint genus. Reglane ser litt annleis ut enn reglane genererte av C4.5RULES, med klassa først og deretter føresetnader for denne, men innhaldet er ekvivalent. Tre av reglane tilsvarar morfologiske tilordningsreglar i Trosterud sitt regelsett. I ordna rekkjefølgje, tilsvarar følgjande reglar MTR 2 (*substantiv uten segmental formativ i ub.pl. er n*), MTR 1 (*omlydssubstantiv er f*) og MTR 5 (*ord på –ing som er avleidd av verb, er f*):

```
N :- a8=uten_segm_form (2635/44).
F :- a8=omlydssubst (57/3).
F :- a9=verb (747/29).
```

Desse tre reglane er identiske med henholdvis regel 2, 90 og 120 i regelsett 1, med unntak av at regel 90 inneheld nokre tilleggskriterium. Tala i parentesen tyder følgjande: Talet før skråstreken er talet på tilfelle som er riktig klassifiserte av den aktuelle regelen, medan talet etter skråstreken er feilklassifikasjonar.

To av reglane i regelsett 3 samsvarar med semantiske reglar hos Trosterud. Den eine er ekvivalent med regel 83 i regelsett 1, og tilordnar femininum til referentar med biologisk hokjønn:

```
F :- a5=hokj (34/3).
```

Substantiv som refererer til biologisk hankjønn, er som vi ser av regel 129 i regelsett 1, maskulinum. Desse vert av RIPPER dekt av defaultregelen. Den neste regelen er identisk med STR 12 i Trosterud sitt regelsett (*ord for terrengfordjupingar er f*):

```
F :- a7=terreng, a6=hol (16/4).
```

Det vert generert ein regel som er identisk med Trosterud sin Regel 2 (*ord på trykklett –e (svake substantiv) er f*), og ein som er identisk med Regel 3 (*einstava ord på vokal er f*):

```
F :- a11=E (1586/498).
F :- a11=V (86/55).
```

Dei tre resterande reglane genererte av RIPPER, samsvarar ikkje med nokon av Trosterud sine tilordningsreglar:

```
F :- a14=d, a12=i (10/6).
F :- a14=d, a13=l, a12=oe (3/0).
F :- a14=d, a12=y (9/2).
```

RIPPER syner dei same tendensane som C4.5RULES. Ein del av reglane i regelsett 1 og 3 er identiske, samtidig som det finst variasjonar mellom setta. Men dei store trekka liknar, og for

tilordning av nøytrum genererer begge systema berre ein regel, basert på bøyingsmorfologi, og ekvivalent med MTR 2.

6.5. Klassifikator 4: RIPPER utan bøyingsmorfologi.

Reglane genererte av klassifikator 4 syner det same som regelsett 2, nemleg at tilordning av nøytrum vert erstatta av ei lang rekkje reglar i staden for den eine som er basert på attributt 8. Regelsett 4 inneheld heile 17 reglar for tilordning av nøytrum:

```
N :- a9=verbalsubst (379/168).
N :- a4=stoff, a14=n (142/25).
N :- a4=stoff, a11=C, a14=d (37/7).
N :- a4=stoff, a11=C, a14=v (18/2).
N :- a14=m, a13=u (74/11).
N :- a7=samfunnsinst (102/33).
N :- a13=e, a14=m (20/4).
N :- a4=stoff, a14=t, a13=a (20/3).
N :- a2=lyd (71/31).
N :- a14=t, a10=ment (57/0).
N :- a4=stoff, a11=C, a13=y, a14=l (11/0).
N :- a14=v, a13=i (49/14).
N :- a14=l, a7=ikkje_perm (9/2).
N :- a13=e, a7=terreng (15/6).
N :- a4=stoff, a11=C, a12=l (14/9).
N :- a13=e, a12=d, a14=r (19/15).
N :- a13=e, a12=r, a14=v (6/1).
F :- a11=E (1583/768).
F :- a9=verb (738/29).
F :- a11=V, a14=o (29/6).
F :- a5=hokj (47/5).
F :- a14=d, a12=y (9/2).
F :- a11=V, a14=u (11/8).
F :- a7=terreng, a6=hol (18/13).
F :- a14=d, a12=i, a13=l (4/1).
default M (6919/1820).
```

Feilraten for regelsett 4 er 22.29%. Utan å kome inn på detaljar for regelsett 4, vil eg konkludere med at ignorering av bøyingsmorfologi utgjer ein stor forskjell for feilraten.

6.6. Oppsummering av regelsett 1-4.

Regelsett 1-4 syner at C4.5RULES og RIPPER genererer ulike reglar ut i frå det same datasettet. Men nokre reglar går igjen for begge algoritmane. Når bøyingsmorfologi er teke omsyn til, genererer begge ein regel tilsvarande MTR 2. Dette er den einaste regelen som tilordnar nøytrum når bøyingsmorfologi er inkludert som attributt. Dette syner at bøyingsmorfologi er ein svært informativ faktor, om ein godtek at det kan bidra til genustilordning. Utan bøyingsmorfologi må fleire reglar for tilordning av nøytrum genererast. Desse varierer frå C4.5RULES til RIPPER. Biologisk kjønn vert teke omsyn til av begge systema, medan andre semantiske element varierer.

Ein del reglar som tilordnar genus på grunnlag av avleiingsmorfologi, er også identiske i regelsetta for RIPPER og C4.5RULES.

I alt vert det generert ein del reglar som tilsvarar tilordningsreglar hos Trosterud, medan ei rekkje av reglane hans ikkje vert tekne omsyn til. I staden vert det generert ein del andre reglar, og nokre av desse er lite intuitive for praktisk bruk. Systema genererer generelt mange færre reglar enn i Trosterud sitt regelsett, utan at dei resulterande regelsetta dekkjer betydeleg færre substantiv. Dette syner at ein del av Trosterud sine reglar er overflødige, og at hans regelsett kan begrensast ein del.

6.7. Klassifikator 5: TiMBL.

Også for TiMBL vart det gjennomført klassifikasjon både med og utan bøyingsmorfologi. Algoritmen som vart nytta var IB1 med vekting ved hjelp av *Gain Ratio*. Ein næraste nabo vart teken omsyn til, og *leave-one-out* vart nytta for inndeling i trenings- og testeksempel. Når bøyingsmorfologi er inkludert klassifiserer TiMBL 12466 av 13384 tilfelle riktig, det vil seie at 93,14% av substantiva har fått riktig genus, eller med andre ord at feilraten ved klassifikasjon er 6,86%. Dette er litt betre enn feilraten for klassifikasjon ved hjelp av C4.5RULES og RIPPER.

TiMBL rangerer dei ulike attributta sin relevans ved trening av datasettet, basert på *Gain Ratio* for kvart attributt:

```
Feature Permutation based on GainRatio/Values :
< 8, 9, 11, 5, 10, 4, 7, 2, 6, 14, 13, 3, 12, 1 >
```

Rangeringa syner dei same tendensane som resultata for dei grådige systema. Attributt som omfattar bøyings –og avleiingsmorfologi (8 og 9) er mest relevante ved klassifikasjon. Når det gjeld fonologi, er dei tre siste bokstavane i substantiva (attributt 12-14) lite relevante, medan meir generelle fonologiske endingar (attributt 11) er mellom dei faktorane som bidreg mest. Middels relevans har suffiks (attributt 10) og semantikk (attributt 5, 4, 7, 2 og 6), med unntak av attributt 3, som omfattar menneskekroppen. Dette er rangert som nest sist på lista, om ein ser bort i frå attributt 1, som er sjølve substantivet og difor er ignorert.

9911 av dei riktig klassifiserte eksempla er *eksakte matchar*. Ein *eksakt match* oppstår når to av eksempla er representerte av dei same attributtverdiane. Det finst også 184 *delte førsteplassar*. *Delte førsteplassar* oppstår når to eller fleire klasser er like frekvente i eit sett av næraste naboar. (Daelemans, Zavrel, van der Sloot, van den Bosch, 2001) Systemet får då problem med å velje den riktige klassa. 116 av dei 184 delte førsteplassane vart riktig klassifiserte av systemet. Det at klassifikasjon ved bruk av TiMBL gir mange eksakte matchar, tyder på at systemet ikkje må anstrenge seg noko særleg for å klassifisere nye eksempel. Ved delte førsteplassar derimot, er det vanskelegare for systemet å velje klasse. Mange eksakte matchar syner at tilordning av genus skjer meir på grunnlag av oppslag av identiske tilfelle enn på grunnlag av likskap. Ein kombinasjon av mange delte førsteplassar og mange eksakte matchar kan tyde på at datasettet er for fattig til at likskap kan verte rekna ut.

TiMBL genererer også ei forvirringsmatrise som syner distribusjonen av feilklassifikasjon i datasettet. Forvirringsmatrisa syner lite variasjon i forhold til den som er generert av C4.5RULES:

```
Confusion Matrix:
             M       N       F
        ---------------------
     M |    7344      33     384
     N |      20    2629       6
     F |     471       4    2493
```

Når bøyingsmorfologi er utelate, er feilraten for TiMBL 19.18%. Dette er også litt betre enn for dei regelbaserte systema.

7. Konklusjon.

7.1. Relevansen av ulike faktorar ved genustilordning.

For å finne ut meir om kva type informasjon som bidreg mest ved genustilordning, utførte eg ei rekkje eksperiment med både C4.5RULES, RIPPER og TiMBL, der ulike attributt vart ignorert. Her såg eg ikkje på spesifikke reglar for dei regelbaserte systema, men på feilratane ved dei ulike eksperimenta for alle tre systema. Såleis kunne eg samanlikne både prestasjonen til dei ulike systema, og prestasjonen generelt når ulik informasjon om substantiva er ignorert. Eg utførte til saman 21 ulike eksperiment, 9 for kvart system, der ulike kombinasjonar av attributt vart utelate. Tabell 7 syner feilratane til C4.5RULES, RIPPER og TiMBL for desse eksperimenta:

Eksperiment	Ignorerte attributt	C4.5RULES	RIPPER	TiMBL
1.	Ingen	7.5%	8.02%	6.86%
2.	Bøyingsmorfologi (a8)	24.0%	22.29%	19.18%
3.	Semantikk (a2-7)	8.0%	8.35%	8.26%
4.	Morfologi (a8-10)	26.4%	25.16%	20.7%
5.	Fonologi (a11-14)	40.3%	14.85%	14.62%
6.	Tre siste bokstavar (a12-14)	24.4%	8.12%	6.93%
7.	Semantikk (a2-7) og tre siste bokstavar (a12-14)	33.6%	8.41%	8.33%
8.	Bøyingsmorfologi (a8) og tre siste bokstavar (a12-14)	32.5%	23.98%	22.21%
9.	Semantikk (a2-7), bøyingsmorfologi (a8) og tre siste bokstavar (a12-14)	33.6%	26.37%	25.93%

Tabell 7*:* Feilratar for C4.5RULES, RIPPER og TiMBL.

TiMBL presterer best i alle eksperimenta. Dette er som forventa, fordi TiMBL, i motsetnad til dei to andre systema, er eit minnebasert maskinlæringssystem som ikkje nyttar nokon form for abstraksjon. C4.5RULES presterer mest variabelt av systema. Eksperiment 1-4 syner liten variasjon mellom RIPPER og C4.5RULES, medan feilratane for eksperiment 5-9 er betydeleg høgre for C4.5RULES enn for RIPPER. Felles for eksperiment 5-9 er at attributt 12-14 er ignorerte. Dette tyder på at C4.5RULES nyttar dei tre siste bokstavane i substantivet i mykje større grad ved klassifikasjon enn RIPPER og TiMBL.

Alle systema presterer best når ingen attributt er ignorerte. Men når dei semantiske attributta er utelatne (eksperiment 3), går feilraten opp med mindre enn 1.5% for alle tre systema. Dette kan tyde på at semantikk bidreg lite til genustilordning. Det same gjeld attributt 12-14 (dei tre siste bokstavane i substantivet). RIPPER og TiMBL presterer nesten like bra utan attributt 12-14 som når desse er inkluderte. C4.5RULES har derimot ei feilrate på heile 24.4% når attributt 12-14 er ignorerte. Men dei låge feilratane for RIPPER og TiMBL tyder på at attributt 12-14 ikkje er avgjerande for eit bra resultat, og dermed at dei tre siste bokstavane i susbtantivet bidreg lite ved genustilordning. Når generell fonologi (attributt 11) i tillegg til dei tre siste bokstavane er ignorert (eksperiment 5), går feilraten opp til ca 15% for RIPPER og TiMBL, og heile 40,3% for C4.5RULES. Dette syner at generell fonologi (om substantivet endar på konsonant, trykklett *–e* eller er eit einstava ord på vokal) bidreg meir til tilordning av genus enn dei tre siste bokstavane.

Når dei morfologiske attributta er ignorerte (eksperiment 4), er resultata frå dei tre systema eintydige: Feilklassifikasjon aukar betydeleg, til over 20% for alle tre systema. Men det er ikkje mykje forskjell mellom resultata frå eksperiment 4, og eksperiment 2, der berre attributt 8 er ignorert. Det er altså bøyingsmorfologi som bidreg i størst grad til riktig klassifikasjon, og ikkje så mykje avleiingsmorfologi og suffiks.

Sidan feilraten er så låg både ved ignorering av semantikk og dei tre siste bokstavane i substantivet (sett bort i frå resultata frå C4.5RULES), vart det utført eit eksperiment der begge desse typane informasjon vart utelatne. Resultata frå eksperiment 7 syner at morfologi og overgripande fonologiske endingar er nok for å oppnå klassifikasjon med låg feilrate. RIPPER og TiMBL presterer nesten like bra i eksperiment 7 som i eksperiment 1, der alle attributt er inkluderte.

Fordi tilordningsreglar på grunnlag av bøyingsmorfologi vanskeleg kan forsvarast, kan berre regelsetta utan a8 sjåast på som informative regelsett (eksperiment 2, 8 og 9). Ein taper lite med omsyn til feilrate ved at ein i tillegg til å ignorere bøyingsmorfologi, ser bort i frå dei tre siste bokstavane (eksperiment 8). Desse kompliserer dessutan reglane (jf. 6.2.2). Feilklassifikasjon aukar ikkje mykje meir om også semantikken er ignorert (eksperiment 9). Regelsetta genererte av RIPPER i eksperiment 8 og 9 kan sjåast som to alternative forslag til regelsett i forhold til Trosterud sitt:

I eksperiment 8 er både bøyingsmorfologi og spesifikke fonologiske endingar utelate. Regelsettet som RIPPER genererer inneheld 14 reglar. 7 av dei tilordnar femininum og 7 tilordnar nøytrum:

```
Final hypothesis is:
N :- a9=verbalsubst (379/168).
N :- a4=stoff, a11=C (358/298).
N :- a7=samfunnsinst (105/33).
N :- a2=lyd (76/31).
N :- a2=gram_kat (54/5).
N :- a10=ment (52/0).
N :- a5=avkj (32/2).
F :- a11=E (1583/760).
F :- a9=verb (738/29).
F :- a11=V, a5=hokj (8/1).
F :- a11=V, a7=terreng (13/11).
F :- a11=V, a6=flate (3/0).
F :- a5=hokj (42/5).
F :- a11=V, a3=kroppsdel (2/1).
default M (6729/1866).
```

Når i tillegg semantiske attributt er ignorert, genererer RIPPER berre fem reglar, to for tilordning av nøytrum og tre for tilordning av femininum (regelsett 9):

```
N :- a9=verbalsubst (379/168).
N :- a10=ment (58/0).
N :- a11=C, a10=skop (9/0).
F :- a11=E (1587/774).
F :- a9=verb (747/29).
default M (7074/2559).
```

Det kan diskuterast om regelsett 8 eller 9 er det mest optimale av desse regelsetta. Regelsett 8 klassifiserer 76.02% av substantiva, medan regelsett 9 klassifiserer 73.63%. Regelsett 9 klassifiserer altså 2.39 % fleire substantiv, men inneheld til gjengjeld over dobbelt så mange reglar som regelsett 8. Semantikk spelar altså ei viss rolle, men ei lita rolle ved genustilordning. Ein kan uansett trekkje den slutninga at avleiingsmorfologi, suffiks og generelle fonologiske endingar er viktige faktorar ved genustilordning i nynorsk, sidan dei åleine kan danne eit regelsett.

7.2. Diskusjon.

Resultata frå RIPPER og C4.5RULES tyder på at regelsettet som Trosterud legg fram kan begrensast mykje. Dette påpeikar han også sjølv. Sidan maskulinum er defaultgenus, kan blant anna reglane for tilordning av maskulinum kuttast ut. Andre reglar for tilordning av maskulinum vil difor vere overflødige. RIPPER genererer ikkje reglar for klassifikasjon av maskuline substantiv (utanom default), men presterer likevel betre enn C4.5RULES, som inkluderer reglar for tilordning av maskulinum. For RIPPER dekkjer defaultregelen alle maskuline substantiv, og resultatet er eit meir begrensa regelsett. Elles er mange av reglane til Trosterud svært spesifikke, og dekkjer berre få tilfelle. RIPPER og C4.5RULES genererer mange færre reglar enn Trosterud inkluderer, noko som gjer regelsettet meir oversiktleg og lettare å lære.

Dei regelbaserte læringsmetodane genererer ein del reglar som er kompliserte og lite intuitive, og difor lite praktisk nyttige. Det viser seg også at ulike algoritmar genererer ulike reglar ut i frå det same datasettet. Slike metodar kan difor ikkje produsere ferdige regelsett for genustilordning, men gi ein peikepinn på moglege reglar.

I kapittel 2.3.2 vert det stilt spørsmål om norsk er eit hovudsakleg semantisk system. Det høge talet på semantiske tilordningsreglar hos Trosterud kan tyde på at semantikk er eit viktig element for tilordning av genus i nynorsk. Resultata som er diskutert i 7.1 peikar derimot mot at semantikk er mindre viktig enn forventa. Visse semantiske reglar kan nok forsvarast, særleg dei som tilordnar genus etter biologisk kjønn. Slike reglar finst nemleg i alle regelsetta der dei semantiske attributta er inkluderte. Men å utelate semantiske attributt fører ikkje til ein særleg større feilrate. Resultata tyder på at morfologi er den viktigaste faktoren for genustilordning i nynorsk. I forhold til dette vil ein difor heller plassere norsk blant morfologiske tilordningssystem enn blant semantiske.

MTR 2 er nemnt som ein problematisk regel både av Trosterud sjølv og i 6.2.3 i denne oppgåva. Problematikken rundt bøyingsmorfologi som føresetnad for genustilordning i norsk, har også vore diskutert av andre. Graedler (1998) tek opp dette temaet for bokmål, der det finst ein klar samanheng mellom endinga *–er* i ubunden fleirtal og felleskjønn, og mellom inga fleirtalsending og nøytrum. Graedler påpeikar at sidan det finst berre to klasser av endingar i ubunden form fleirtal, og to genusklasser, er det vanskeleg å vite kva som er tilordna først. Det vert referert til Rand Schmidt, som meiner at fleirtalsformene er knytta til eit substantiv på eit seinare stadium enn eintalsformene, som har genusmerke. Om dette er tilfelle er det fleirtalsformene som rettar seg etter genus, og ikkje omvendt. Trosterud sin MTR 2 kan i så fall ikkje forsvarast.

7.3. Konklusjon.

Målet med denne hovudoppgåva var å etterprøve ein hypotese om at genustilordning i norsk er regelstyrt, gjennom ei datamaskinell etterprøving av Trosterud sitt konkrete framlegg til eit regelsett. Eg bygde opp eit datasett med substantiv som vart tildelte attributtverdiar på grunnlag av

reglane til Trosterud. Maskinlæringseksperimenta som vart utførte på dette datasettet, syner regelbundenheit ved genustilordning, og stadfestar generelt sett Trosterud si hypotese. Vidare kom ein fram til at hans regelsett kan begrensast mykje utan at det går utover dekningsgraden.

Trosterud legg fram eit regelsett med mange svært spesifikke reglar. Ved etterprøving av desse vert det ikkje oppdaga eventuelle andre element som tilordnar genus i nynorsk, enn akkurat dei som Trosterud legg til grunn for sitt regelsett. For å oppdage eventuelle andre faktorar er det behov for ei større mengde semantisk, morfologisk og fonologisk informasjon i tilknytning til substantiva. Det kan vere problematisk å inkludere nok informasjon til ikkje å utelate potensielle faktorar for tilordning av genus. Morfologisk og fonologisk informasjon er såpass begrensa at det er mogleg å inkludere tilstrekkjeleg, utan å utelate potensielle innverkande element for genustilordning. Å inkludere nok semantisk informasjon er derimot problematisk, sidan slik informasjon kan vere svært spesifikk og av stort omfang. Vidare eksperiment bør utførast, med ei grundigare koding av substantiv på grunnlag av semantikk, fonologi og morfologi.

Ut i frå den informasjonen som faktisk er inkludert i Trosterud sitt regelsett, fann ein at morfologi bidreg mest til genusstilordning, og at semantisk informasjon var mindre viktig enn forventa. Dei tre siste bokstavane i substantivet bidrog minimalt, medan generelle fonologiske endingar har ein del å seie for genustilordning.

Når det gjeld resultata må eg ta stort atterhald på grunn av databasen sitt manglande omfang. Under halvparten av dei substantiva som Trosterud hevdar å ha nytta som korpus, er inkluderte, og eg kan ikkje garantere at alle typar substantiv er representerte i databasen. Det er difor mogleg at det finst grupper av ord som ikkje er inkluderte, som hadde gitt stort utslag på resultata.

Resultata frå eksperimenta som er gjort gir ikkje noko svar på korleis eit optimalt regelsett for genustilordning i nynorsk bør sjå ut. Dei syner derimot at det finst regelbundenheit, og at regelbaserte læringsmetoder er eigna til å systematisere denne regelbundenheita, gitt tilstrekkeleg informativ koding av substantiv.

Referansar.

Beito, O. T. (1986 (1970)): *Nynorsk Grammatikk. Lyd – og ordlære*, 2. utgåva, Det Norske Samlaget, Oslo.

Building Classification Models: ID3 and C4. Tilgjengeleg på: http://www.cis.temple.edu/~ingargio/cis587/readings/id3-c45.html

Cohen, W. W. (1995): Fast Effective Rule Induction. Frå: *Machine Learning: Proceedings of the Twelfth International Conference (ML95).*

Corbett, G. G. (1991): *Gender*, Cambridge University Press, New York, USA.

Daelemans W. & Durieux G. (2000): Inductive Lexica. I Van Eynde, F. and D. Gibbon (Eds), *Lexicon Development for Speech and Language Processing*, Kluwer Academic Publishers, pp. 115-139.

Daelemans, W., Zavrel J., van der Sloot K., van den Bosch A. (2002): TiMBL: Tilburg Memory-based Learner. Version 4.3. Reference Guide. *ILK Technical Report-ILK 02-10.* Tilgjengeleg på: http://ilk.kub.nl/downloads/pub/papers/ilk0104.ps.gz.

Faarlund, J. T., S. Lie & K. I. Vanneboe (1997): *Norsk Referansegrammatikk*, Universitetsforlaget, Oslo.

Graedler, A-L. (1998): *Morphological, semantic and functional aspects of English lexical borrowings in Norwegian*, Universitetsforlaget AS (Scandinavian University Press), Oslo.

Hendrickx, I. & A. van den Bosch (2003): *Binary versus Multi-valued Features in Machine Learning of Natural Language: k-NN Versus SVM.* http://cnts.uia.ac.be/clin2003/abstracts/hendrickx.html

Husby, O. (1990): Norske ord. *Ordlagingslære med arbeidsoppgaver*, Friundervisningens Forlag, Oslo.

Leirvaag, O. G. (1999): *Norsk på grunnlag av samisk,* 2. utg*,* Samisk utdanningsråd, Kautokeino.

Nynorskordboka. Definisjons- og rettskrivingsordbok, 3.utgåva, 2001: Redaksjon: Hovdenak, M., L. Killingbergtrø, A. Lauvhjell, S. Nordlie, M. Rommetveit, D. Worren, Det Norske Samlaget, Oslo

Trosterud, T. (2001): Genustilordning i norsk er regelstyrt. I: *Norsk Lingvistisk tidsskrift* Årgang 19, pp. 29-58.

Overview of Decision Trees. Tilgjengeleg på: http://www.cs.uregina.ca/~hamilton/courses/831/notes/ml/dtrees/4_dtrees1.html

Quinlan, J. R. (1992 (1987)): *C4.5: Programs For Machine Learning,* Morgan Kaufmann Publishers, San Mateo, California.

Van Den Bosch, A. (with W. Daelemans, J. Zavrel & S. Buchholz): *Machine Learning / MBL for natural language processing*, Forskerutdanningskurs i statistisk språkbehandling, Bergen, 30.10-04.11- 2002.