UNIVERSITET I OSLO
Institutt for informatikk

Automatisk oversettelse av norske substantivkomposita

En eksperimentell studie

«Masteroppgave»
(60 studiepoeng)

Lars Bungum

1. september 2008

Sammendrag

I denne oppgaven omtales forsøk på å automatisk oversette norske substantivkomposita til engelsk. Det blir foretatt en diskusjon av komposita fra et lingvistisk ståsted og tidligere arbeid innenfor samme område, i særdeleshet arbeider med å oversette komposita fra engelsk til japansk og motsatt vei (Baldwin & Tanaka, 2004).

Det blir foretatt en gjennomgang av anvendt metodologi, og en diskusjon av ulike rammeverk, og de valgene som er gjort. Det blir også en vist til eksperimenter på oversettelse fra norsk til engelsk, som for de beste metodene kan rapportere en nøyaktighet som tilsier at for over 50% av de analyserte kompositaene klarte metoden å finne en riktig oversettelse. Bare 4% av disse kompositaene hadde oppslag i Kunnskapsforlagets «Engelsk stor ordbok: engelsk norsk/norsk-engelsk» (Eek, 2001). På bakgrunn av den lingvistiske diskusjonen av komposita blir resultatene og metodikken drøftet mot slutten av oppgaven.

Forord

Diskusjonen om komposita er svært gammel, og kan diskuteres på mange forskjellige nivåer. Måten komposita skrives på norsk, og spesielt utglidning i normene for dette (såkalt orddeling) vekker mye engasjement, og det er således enkelt å forklare hva oppgaven forsøker å gjøre. Hvordan den gjør det, er imidlertid mer innviklet, og for meg har det vært en stor glede å sette meg inn i et et saksområde som jeg har og har hatt et dagligdags forhold til på en systematisk og vitenskapelig måte.

Jeg vil først og fremst takke min veileder, professor Stephan Oepen, som med stø hånd og konstruktive tilbakemeldinger har veiledet meg gjennom både eksperimentfasen og skriveprosessen, som tidvis forekom meg bratt. Idet jeg er i ferd med å sette punktum vil jeg takke for denne hjelpen.

Dernest vil jeg takke Daniel Fredrick Mikkelsen som har stilt opp som informant, og utformet gullstandard og gjort sølv standardvurderinger. Jeg vil også takke medstudenter og venner fra og utenfor lesesalen, og ellers andre som har diskutert saken fra ymse sider med meg. Jeg vil også takke mine kolleger gjennom to somre på lønningskontoret hos IKEA, som med sin strålende vennlighet ga meg energi og pågangsmot til å bli ferdig.

September, 2008, Lars Bungum.

Innhold

1 Innledning **1**
- 1.1 Motivasjon . . . 2
- 1.2 Forskningsspørsmål . . . 4
- 1.3 Eksperimenter . . . 4
- 1.4 Oppbygning . . . 5
- 1.5 Bidrag . . . 5
- 1.6 Typesetting . . . 6
- 1.7 Språklige hensyn . . . 6

2 Tidligere relaterte arbeider **9**
- 2.1 Tidlige arbeider ved Rackow et al. . . . 9
- 2.2 Grefenstettes web-eksperiment . . . 12
  - 2.2.1 Tekstkorpora . . . 12
  - 2.2.2 WWW som korpus . . . 12
  - 2.2.3 Eksperimentet . . . 13
  - 2.2.4 Betydning av resultater . . . 14
- 2.3 Hanne Moas ReCompounder . . . 15
  - 2.3.1 Konstruksjonstyper . . . 16
  - 2.3.2 Kandidater . . . 16
  - 2.3.3 Oversettelse . . . 17
  - 2.3.4 Evaluering . . . 17
  - 2.3.5 Sammenlikning med Grefenstette . . . 18
- 2.4 Baldwin og Tanaka . . . 19
  - 2.4.1 Dyp og grunn oversettelse . . . 19
  - 2.4.2 Rangering etter CTQ . . . 21
  - 2.4.3 Maskinlæring og tospråklige leksika . . . 25

3 Komposita og oversettelse **35**
- 3.1 Om leksikalisering . . . 36
  - 3.1.1 Leksikalisering og orddannelse . . . 37
- 3.2 Om komposisjonalitet . . . 38

3.2.1 Semantisk komposisjonalitet 40
3.3 Allment om komposita 41
3.3.1 Inndeling av komposita 43
3.3.2 Universalitet 44
3.4 Komposita på norsk 44
3.4.1 Bøyning av komposita 46
3.4.2 Sammensetningsformer 47
3.4.3 Semantisk inndeling 48
3.4.4 Klassifikasjon etter ordklasse 49
3.4.5 Begrepsbruk i denne oppgaven 50
3.5 Oversettelse av komposita 51
3.5.1 Utbredelse av nominalkonstruksjoner 51
3.5.2 Automatisk analyse 52
3.5.3 Annet om sammensetningsanalysatoren 60
3.5.4 Perspektiver fra Baldwin og Tanaka 60

4 Ressurser og verktøy **63**
4.1 Engelsk Stor Ordbok 63
4.1.1 Søking i KF 64
4.2 Korpora 64
4.2.1 Norske korpora 64
4.2.2 Engelske korpora 65
4.3 VarCon 66
4.4 WordNet 66
4.5 RASP 67
4.6 TADM 68
4.7 Søkegrensesnitt for Yahoo 68
4.8 Oslo-Bergen-taggeren 69
4.8.1 Sammensetningsanalysatoren 70

5 Metodologi **71**
5.1 Support Vector Machines 71
5.2 Maksimalentropi 76
5.2.1 Informasjonsentropi 76
5.2.2 Et eksempel, å oversette *in* 77
5.2.3 MaksEnt-modeller 78
5.2.4 Parameterestimering 81
5.2.5 Anvendelser i NLP 82
5.2.6 Glatting 82
5.3 Eksperimentmiljø 83
5.3.1 Oversikt 83

5.3.2 Behandling av engelske korpora 85
5.3.3 Indeksering av parsede engelske korpora 86
5.3.4 Morfologisk analyse av norske korpora 88
5.3.5 Siling av komposita 89
5.3.6 Utarbeidelse av fasit og maler 90
5.3.7 Rangeringsprosessen 93
5.3.8 Bygging av MaksEnt-modell 93
5.3.9 Evaluering 95
5.4 Hypotesetesting 98

6 Eksperimenter **101**
6.1 Rangeringsmetoder 102
6.1.1 Heuristiske metoder 102
6.1.2 Maskinlæringsmetoder 103
6.2 Trekk 103
6.2.1 Korpusbaserte trekk 103
6.2.2 Tospråklige trekk 104
6.2.3 Mal-trekk 106
6.2.4 Eksempler 106
6.3 Resultater 107
6.3.1 Rangeringsmetoder 108
6.3.2 Analysedybde 109
6.3.3 Korpusstørrelse 111
6.3.4 Sølvstandard 113
6.4 Glatting 114

7 Diskusjon og konklusjoner **117**

A Oppsummering av resultater **125**
Referanser 127

Kapittel 1

Innledning

Denne oppgaven ønsker å vise hvordan et maskinoversettelsessystem kan oversette kjente og ukjente substantivkompostita som *campingbord* og *salamipakke*, helt naturlige norske ord brukt i større og mindre utstrekning, men som langt fra nødvendigvis står i en ordbok. Prosessen med å oversette dem består av to hoveddeler, først å generere mulige oversettelser, og dernest å velge ut den riktige blant dem. Dette gjøres ved hjelp av å hente inn data om hvor ofte oversettelseskandidatene og deres bestanddeler oppter i ulike korpora. Tolkningen av disse dataene blir et stort tema for oppgaven, det gjøres dels ved heuristikker, og dels ved hjelp av maskinlæringsteknikker.

Først deles ordene i to deler hvor begge kan identifiseres som substantiv, oversettes hver enkelt del til engelsk, og kombinasjonene av disse oversettelsene settes sammen til mulige oversettelser, og gjøres til gjenstand for spørringer mot data fra tekstkorpora. Slik fungerer systemet i sin enkleste form: *camping* oversettes til *camping*, *bord* til *table* og samensatt blir dette *camping table*. Men ofte vil ikke ordene kunne oversettes til engelsk direkte på denne måten. Noen ganger ved at den engelske ekvivalenten er et leksikalisert enkeltord urelatert til det norske utgangspunktet som vil være umulig for et slikt system å oversette, men ofte også ved at det gjøres systematiske skift, slik som at *salami·pakke* kan oversettes med *package of salami*. Disse systematiske skiftene fanges

opp ved et sett av maler som forteller at eksempelvis at oversettelsen av det første norske ordet skal stå i annen posisjon i oversettelsen og vice versa, med *of* imellom. Slik kan man nå oversettelsen i det siste eksempelet. Innføringen av slike maler innfører en ny faktor i beregningen av kryssproduktet. Antallet kandidater til oversettelse i målspråket kan bli høyt, ofte mange hundre. Det er enklere å lage et system for å generere mulige oversettelser hvor den riktige oversettelsen er en del av dem, enn å velge en riktig, eventuelt den riktige, av disse kandidatene.

Metodene som brukes i oppgaven baserer seg på eksperimenter gjort av Baldwin og Tanaka (2004) på liknende oversettelser mellom engelsk og japansk (begge retninger). Metoden for å generere kandidatoversettelser er de samme, men utvelgelsen skiller seg på et vesentlig punkt, valg av maskinlæringsteknikk. Baldwin og Tanaka (2004) bruker en SVM (støttevektormaskin) til å rangere kandidatene, mens en rangering basert på et masksimalentropirammeverk benyttes i eksperimentene omtalt i kapittel 6. I begge tilfeller trenes maskinlæringssystemene på data fra analyserte tektkorpora som forteller hvor ofte kandidaten og dens bestanddeler opptrer i hver mal samt i enkelte tilfeller oppslag i tospråklige ordbøker. Informasjonen fra korporaene og ordboken er representert som trekk ved hver oversettelseskandidat. Og kombinasjonen av hvilke trekk en ny oversettelseskandidat har, kombinert med en vekting som kommer som et resultat fra en maskinlærer, avgjør hvor høyt rangert kandidaten blir.

1.1 Motivasjon

I maskinoversettelsessytemer som benytter seg av oppslag i en ordbok vil man ofte ha problemer med å oversette komposita, fordi at det sammensatte ordet ikke finnes som oppslag. Som en teoretisk behandling av komposita i kapittel 3 antyder, så er slike ordsammensetninger i varierende grad produktive, i betydningen at nye komposita opprettes løpende, slik at et slikt system kan støte på et kompositum som teoretisk

har vært skrevet for første gang. Metoden som brukes i denne oppgaven har muligheten til å finne en riktig og/eller meningsbærende oversettelse også for et slikt ord. Et eksempel på en slik sammensetning kan være *lapskaus·utstilling*, uten at det påstås om at dette ordet aldri er ytret før nå (det er usannsynlig), så finnes det ikke i Google[1] eller Yahoo[2] sine søkemotorer. En teknikk som først deler sammensetningen opp i to og oversetter del for del har muligheten til å fange opp meningsinnholdet i sammensetningen likevel.

Koehn og Knight (2003) finner at samlet ytelse for statistiske maskinoversettelsessytemer går opp dersom nominalfraser fra kildespråket oversettes med nominalfraser også i målspråket. I en slik kontekst vil et system som dette kunne bidra, ved å kunne foreslå en oversettelse av sammensetninger, som kanskje ellers ville blitt hoppet over eller tatt over i målspråket i sin opprinnelige form. Johannessen og Hauglin (1996) slår fast at det er umulig å ha en liste over alle komposita i en ordbok, og når veldig mange systemer som behandler naturlige språk baserer seg på nettopp på ordbok, er systemer som behandler ord som ikke står der viktig. Et system som kan oversette ord som ikke står i en ordbok er i lys av dette en nødvendighet.

En artikkel i avisen Verdens Gang[3] datert 16. mai 2008 beskriver utvidelsen av et maskinoversettelsessystem mellom mange språk til også å inkludere norsk. Journalisten skriver blant annet:

> Spesielt det norske språkfenomenet særskriving, ofte kalt orddeling, sliter (...)-oversettelsen med. Ord som «sjimpanse-vold» blir dermed ikke oversatt og står i sin norske form midt i den oversatte teksten.

Dette tolkes som at journalisten påpeker at norske komposita skrives i

[1] `http://www.google.no/search?q=lapskausutstilling` besøkt 1. september 2008.

[2] `http://search.yahoo.com/search?p=lapskautstilling` besøkt 1. september 2008.

[3] `http://www.vg.no/teknologi/artikkel.php?artid=192875` besøkt 1. september 2008.

ett ord, og at dette representerer en utfordring for oversettelsessystemet. Det er sannsynligvis en av flere utfordringer, men det er fortellende at journalisten trekker fram nettopp problemer med komposita i sin første omtale av tjenesten.

1.2 Forskningsspørsmål

En oppsummering av de viktigste spørsmålene oppgaven ønsker å kommentere er:

- Hvilket utslag gjør antall ord i tilgjengelige korpora?
- Hvilket utslag gjør sammensetning av korpora?
- Hvilket utslag gjør analysedybde (parsing vs. tagging)?
- Hvilket utslag gjør valg av maler på ytelsen?
- Hvordan egner maskinlæringsteknikker seg til denne oppgaven?
- Hvordan fungerer teknikkene på nært beslektede språk norsk og engelsk i forhold til fjernt beslektede?

Eksmperimentresultatene vil gi innblikk i disse spørsmålene.

1.3 Eksperimenter

Eksperimentene utført i forbindelse med denne oppgaven er beskrevet i kapittel 6, og er gjort etter mønster av Baldwin og Tanaka (2004), målet er å gjøre et tilsvarende eksperiment som fra norsk til engelsk, som ble gjort for fra japansk til engelsk og motsatt vei. Ved å bruke den samme tilnærmingen til oversettelse fra norsk til engelsk vil kunne kaste lys egnetheten til denne metoden for oversettelse av også norske komposita til engelsk. Ved siden av dette ønsker forsøkene å si noe om betydningen av størrelse på korpora og analysedybde i korpuset som utgjør grunnlaget for rangeringen, samtidig som en alternativ maskinlæringsteknikk blir brukt.

Det ble trukket ut 750 komposita fra analyse løpende norsk tekst som ble forsøkt oversatt. I de fleste eksperimentene ble rangeringsmetodene testet for de komposita hvor de hadde muligheten til å finne frem til den ønskede oversettelsen. Av de 750 kompositaene, så var det mulig å finne en oversettelse av 444 av dem med en teknikk som går ut på å oversette kompositumene del for del. I tilfeller der den foretrukne oversettelsen bestod av bare ett ord, eller en nødvendig deloversettelse ikke fantes blant ordbokens oversettelser er det ikke mulig for en rangeringsmetode å velge riktig kandidat, fordi den ikke er blant kandidatene. Teknikken ble evaluert mot oversettelser innhentet fra en informant med morsmålskompetanse i begge språk.

1.4 Oppbygning

Oppgaven begynner med en diskusjon av tidligere relaterte arbeider, hvor (Baldwin & Tanaka, 2004) er særlig viktig. Det går over i en diskusjon av komposita fra et lingvistisk ståsted, og i særdeleshet hvordan de opptrer i norsk språk, og forsøk på å analysere og oversette dem. Dernest begynner beskrivelsen av eksperimentene med en gjennomgang av de ressursene som har vært brukt i kapittel 4, da mye forskjellige programvare og flere store tekstkorpora har vært satt sammen for å utføre eksperimentene. Dernest følger en beskrivelse av metodologien som brukes, med en kort presentasjon av maskinlæringsteknikker i kapittel 5, inkludert en beskrivelse av den praktiske implementasjonen av selve eksperimentet. Oppgaven avsluttes med en gjennomgang av eksperimentene og resultatene fra dem, etterfulgt av en kort oppsummering og diskusjon i kapittel 7.

1.5 Bidrag

Eksperimentene i denne oppgaven viser at en oppdelende oversettelsesstrategi for komposita fra norsk til engelsk er en metode som kan fungere. Av de 444 kompositaene som var mulige å oversette med teknikkene brukt

i eksperimentene var bare 18 (4%) av dem oppført med oppslag i Kunnskapsforlagets Engelsk Stor Ordbok (Eek, 2001). Likevel klarte den beste rangeringsteknikken i eksperimentene å finne riktig kandidat i over 50% av tilfellene. I en sekundær (og mer subjektiv) evaluering ble informanten bedt å om å vurdere om den høyest rangerte oversettelseskandidaten hadde fanget opp det sentrale meningsinnholdet i kompositumet den skulle oversette, slik at det var mulig å finne tilbake til hva som var oversatt. For den beste rangeringsmetoden var utgangspunktet *gjenfinnbart* i denne forstand i 70% av tilfellene.

Selv om resultatene ikke kan sammenlignes direkte med (Baldwin & Tanaka, 2004), så er resultatene kvantitativt i samme område som deres, og viser til en dekning på over 50% av de tilfeldig utvalgte kompositaene for løpende norsk tekst for den beste rangeringsteknikken. Dermed støtter denne oppgaven deres funn om at en maskinlæringstilnærming er formålstjenlig for et et problem som dette.

1.6 Typesetting

Oppgaven er skrevet i LaTeX, flytdiagrammet er laget i Kivio, og plottingen av resultatene er foretatt i OpenOffice Calc.

1.7 Språklige hensyn

Fordi oppgaven er skrevet på norsk, mens kildematerialet i nesten fullstendig grad er skrevet på engelsk og fagtermer i overveiende grad benyttes i sin engelske form er det en utfordring å bruke et godt og flytende språk samtidig som meningsinnholdet bibeholdes og formuleringene er presise. Det er dermed vanskelig å skrive en slik oppgave fri for «kaudervelske» formuleringer. Å komme frem til et dokument mest mulig fritt for engelske låneord har heller ikke vært et mål, da heller å presentere materialet på en forståelig måte i norsk språkdrakt.

Ord som *web*, *parser* og *tagger* brukes i teksten på en naturalisert måte i

konteksten av maskinell språkanalyse. Dersom det har vært nødvendig å sitere engelske originaler i omtale av en kilde gjøres dette i fotnoter, og norske oversettelser av eksempler og direkte sitater skrives direkte inn i teksten i parenteser. Hvis ikke annet er angitt er slike oversettelser forfatterens.

Kapittel 2

Tidligere relaterte arbeider

I dette kapittelet vil det bli gitt en presentasjon av utvalgte arbeider med automatisk oversettelse av komposita. Først diskuteres forsøkene til Rackow, Dagan og Schwall (1992), som også bruker kvantitative korpusdata til å velge rett oversettelse, og deretter (Grefenstette, 1999) og (Moa, 2005), som bruker WWW som et korpus til å foreta rangeringen. Sistnevnte forsøk dreier seg også om oversettelse fra norsk til engelsk og bidrar dermed med refleksjoner rundt denne prosessen. Avslutningsvis vil utviklingen i Timothy Baldwin og Takaaki Tanakas arbeider i feltet bli vist, hvorav metodikken i den siste angrepsvinkelen brukes i eksperimentene i denne oppgaven.

2.1 Tidlige arbeider ved Rackow et al.

Rackow et al. (1992) gjennomførte på begynnelsen av 1990-tallet forsøk på å automatisk oversette tyske substantivkomposita til engelsk. De begrenset seg til substantivkomposita hvis forledd også var et substantiv, som Rackow (1992) fant å være den mest utbredte typen både på tysk og engelsk. Rackow et. al skildrer to hovedutfordringer med å oversette de tyske sammensetningene til engelsk, først å segmentere det tyske ordet, og dernest å velge riktig engelsk leksem å oversette dem med, for så å velge riktig konstruksjonstype for engelsk. De

anser tyske komposita som entydig skrevet i ett ord, slik som på norsk, men at engelske komposita har flere konstruksjonstyper som substantiv-preposisjon-substantiv eksemplifisert med *in*, som i *chief of staff* eller adjektiv-substantiv som i *parliamentary debate*. Spørsmålet om disse engelske variantene har kompositumsstatus blir ikke behandlet inngående, men en definisjon som begrenser tyske komposita til en konstruksjonstype som resulterer i ett ord, mens engelske komposita tillates å bruke flere konstruksjonstyper som involverer preposisjoner og artikler er potensielt motstridende. Jamfør diskusjonen av komposita i avsnitt 3.3, så vil den tyske definisjonen være en syntaktisk definisjon som beskriver en konstruksjonstype, mens den engelske definisjonen rommer flere konstruksjonstyper, bærer bud om en mer leksikalsk tilnærming.

Rackow et. al har innenfor rammen av et semantisk transfer-system utviklet en algoritme for å behandle dertil usette substantivkomposita. Oversettelsesalgoritmen som er kalt COMPGE kalles dersom et ord på over fem bokstaver ikke står i leksikonet, eller finnes i ordboken Collins German-English Dictionary[1]. I et slikt tilfelle splitter algoritmen, implementert i PROLOG, opp ordet i retning fra venstre mot høyre fra tredje bokstav av, til den finner et ord som står i leksikonet. Dernest tar den hensyn til fugeelementene og leter etter et nytt ord i resten av ordet. Algoritmen blir kjørt om igjen for å se etter flertydigheter i oppsplittingen. En analyse som har to substantiviske ledd, som *husmanns·plass* blir foretrukket foran et med tre ledd som *hus·manns·plass*, et med tre foran et med fire, etc. Det fremgår ikke hvordan algoritmen velger mellom to analyser med samme antall ledd.

I neste omgang gjøres en semantisk analyse av sammensetningen, hvor sammensetningens syntaktiske og semantiske egenskaper følger sammensetningens *kjerne* (det engelske begrepet *head* blir brukt, forståelsen av *kjerne*-grepet blir presentert i avsnitt 3.4), som i de fleste tilfeller er det siste leddet. Dette sammen med morfologisk informasjon blir brakt inn til de semantiske transferreglene. Da står systemet overfor den andre hovedut-

[1]URL: `http://dictionary.reverso.net/german-english/`.

fordringen, å generere riktig engelsk kompositum av riktig type og med riktige ledd. Empiriske undersøkelser foretatt av Rackow (1992) viser at 54,4% av tyske substantivkomposita med substantiv som både forledd og etterledd ble oversatt til tilsvarende konstruksjonstypen, og derfor ble denne konstruksjonstypen valgt som forvalg. Men med utelukkende en slik strategi ville fremdeles om lag halvparten av ordene bli oversatt galt, og korpusbaserte teknikker ble derfor lagt til for å avhjelpe dette.

Dersom det er flere kandidater til oversettelse fordi det er flere mulige oversettelser av forledd og/eller etterledd, blir disse rangert etter hvor ofte de opptrer i et tagget korpus. Men mange komposita forekommer for sjeldent til at en slik rangering er mulig eller meningsfylt. I slike tilfeller blir kandidatene rangert etter hvor ofte oversettelsen av forleddet opptrer i en viss konstruksjonstype. Rackow et. al viser til *ecological* etterfulgt av et substantiv forekommer oftere enn *ecology* etterfulgt av substantiv. Derfor vil en konstruksjon med *ecological* som forledd foretrekkes dersom det ikke er mulig å skille oversettelseskandidatene som helhet etter frekvens. Eksempelet *Umwelts·probleme* (norsk: *miljøproblemer)* blir brukt til å illustrere poenget ved at metoden sørger for at ønskede *ecological problems* blir foretrukket fremfor *ecology problems*.

Forfatterne merker seg også at det i en rekke tilfeller er flere konstruksjontyper som gir gyldige oversettelser, og bruker dette som et argument for å bruke korpusdata til å rangere oversettelseskandidater. Komponentene i systemet var ikke fullt integrert, men utelukkende testet separat, slik at en evaluering ikke var tilgjengelig. Fordi at norske komposita i likhet med tyske skrives i ett ord som består av to ord føyet sammen, er det i begge tilfelle nødvendig med en oppsplitting. Måten ordene kan føyes sammen ved hjelp av et fugeelement likner mye på måten det gjøres på norsk (en nærmere diskusjon av norske komposita vil bli gitt i kapittel 3). Langer (1998) identifiserer 68 slike fugeelementer, inkludert en sjelden form hvor det første ordet står i flertall, som i *Prinzipien·reiter* (norsk: *prinsipp·rytter*), som viser til en større variasjonsbredde enn i norsk. Likevel kan sammenhengen mellom fugeelement og riktig engelsk konstruksjonstype som blir diskutert i (Rackow et al., 1992) være

relevant også for norsk. Dette blir ikke utforsket i eksperimentene utført i denne oppgaven.

2.2 Grefenstettes web-eksperiment

Idéen bak eksempelbaserte «Natural Language Processing»-systemer (NLP) er at attesterte lingvistiske hendelser kan brukes til å velge mellom teoretisk mulige lingvistiske hendelser eller språklige fenomener. «Attesten» betyr i denne sammenhengen at det er bekreftet at noen faktisk har sagt dette med den hensikt å kommunisere på et gitt språk, at det er autentisk. Forskjellige konstruksjonstyper av engelske komposita slik det er beskrevet i avsnitt 2.1 kan være eksempler på slike hendelser, og rangering etter korpusdataene eksempler på hvordan korpora kan brukes til å velge blant dem.

2.2.1 Tekstkorpora

Wikipedia[2] definerer et lingvistisk tekstkorpus som et strukturert samling tekster, brukt for å gjøre statistisk analyse, validere grammatiske regler, og å sjekke hvor ofte et fenomen, som en ord, frase eller setning opptrer. Hvor strukturert teksten er varierer, det kan være utelukkende råktekst, eller den kan være håndannotert, som for Penn Treebank[3]. «World Wide Web», (heretter kalt webben) har vært vurdert som for uorganisert og kaotisk for å kunne brukes til å tjene som en kilde av eksempler for nettopp i en slik sammenheng.

2.2.2 WWW som korpus

Grefenstette (1999) anerkjenner at det er mye støy på webben, men argumenterer for at dens størrelse veier opp for dette. Lingvisters eksempelsetninger vil for eksempel ligge tilgjengelig på webben, slik at

[2] `http://en.wikipedia.org/wiki/Text_corpus` besøkt 27. august 2008.

[3] `http://www.cis.upenn.edu/~treebank/`.

Frase	BNC	WWW 1998	WWW 2008
medical treatment	202	46 064	8 290 000
prostate cancer	28	40 772	10 900 000
deep breath	374	54550	6 560 000
acrylic paint	20	7308	1 040 000
perfect balance	28	9735	2 790 000
presidential election	74	23 745	12 700 000
electromagnetic radiation	24	17 297	1 740 000
powerful force	54	17 391	1 140 000
concrete pipe	8	3 360	304 000
upholstery fabric	5	3157	638 000
vital organ	30	7371	154 000

Tabell 2.1: *Webtreff for utvalgte fraser i British National Corpus og på World Wide Web. Kopi av tabell fra Grefenstette, med webtreff fra i dag i egen kolonne.*

det er en fare for en sirkulærargumentasjon som leder til selvoppfyllende profetier. En teori ville kunne bevises med eksempler en selv har lagt ut. Grefenstette argumenterer imidlertid for at denne og annen type støy blir forsvinnende liten på grunn av webbens enorme omfang. Eksempler fra lingvister vil aldri kunne overskygge ekte bruk av språket.

Han gir et noen eksempler på kompositas webtreff, og for å illustrere dets enorme ordtilfang gjengis en eksempel-tabell fra hans artikkel i tabell 2.1, oppdatert med resultater fra søkemotoren Google .

2.2.3 Eksperimentet

For å vise dette gjennomfører han et eksperiment hvor han velger ut alle spanske og tyske kompositа fra en offentlig tilgjengelig ordbok[4] ved å eliminere alle som ikke var «transparente oversettelser av sine bestanddeler», To eksperimenter ble gjort, fra tysk til engelsk og fra spansk til engelsk. En komplett liste av tyske substantivkomposita (det fortelles ikke hvordan disse ble valgt ut) og spanske nominalfraser som

[4]Ordboken som ble brukt finnes på `http://www.icp.grenet.fir/ELRA/cata/text_det.html\#basmullex`.

tilfredsstilte fire kriteria ble silet ut. De var:

- Kompositumet eller frasen kunne brytes opp i to andre ord som fantes i ordboken.
- Kompositumet eller frasen var oversatt til engelsk som to ord.
- Det var mulig å sette sammen den engelske oversettelsen av ordene fra det første punktet til den engelske oversettelsen i det andre punkt.
- Det var mer enn én mulig engelsk oversettelseskandidat.

724 tyske komposita og 1140 spanske nominalfraser tilfredsstilte disse kriteriene, og henholdsvis 3556 mulige oversettelser av de tyske kompositaene, og 6186 av de spanske nominalfrasene ble generert. Generering av oversettelseskandidater foregikk her ved å utarbeide kryssproduktet av de engelske oversettelsen av ordene utgangspunktet bestod av, Et eksempel er *Apfel·saft*, hvor kandidatene til oversettelse var *apple·juice* og *apple·sap*, fordi *saft* har to engelske oversettelser. Kandidatene ble sendt til søkemotoren Altavista[5] som et frasesøk. Treffene som ble returnert fortalte dermed hvor ofte de to ordene stod i rekkefølge, i motsetning til hvor ofte de forekom sammen på en side. For 87% av de tyske kompositaene og 86% av de spanske frasene fra listen over, var oversettelseskandidaten med høyest antall treff denne samme som selve termens oversettelse i ordboken.

2.2.4 Betydning av resultater

Grefenstette diskuterer ikke komposita eller komposisosjonalitet som fenomen, og det forstås som om poenget hans var å vise at webben kan brukes til å rangere mellom instanser av lingvistiske fenomener også i ubehandlet form. Han bruker imidlertid bare begrepet *kompositum* [6]

[5] `http://altavista.digital.com`.

[6] engelsk: *compound*.

om de tyske sammensetningene, men ikke de spanske, som tyder på at han betrakter komposita som en spesifikk konstruksjonstype. Resultatene for utvalget var overbevisende, men som Rackow (1992) viste, så er det bare om lag halvparten av tyske substantivkomposita av struktur substantiv-substantiv som får den samme strukturen på engelsk. Slik tas mye av rommet for feil vekk i måten listen blir valgt ut på. Videre er det fremdeles mulig at et tradisjonelt korpus ville gitt enda bedre resultater enn Grefenstette oppnådde i disse eksperimentene. En slik sammenheng hadde vært interessant, selv om poenget ikke nødvendigvis er å vise at WWW-rangering er bedre enn ved hjelp av tradisjonelle korpora, men at det kan brukes, og er mye enklere tilgjengelig.

I automatisk oversettelse av komposita, så er ikke den største utfordringen å oversette komposita som står i en tilgjengelig tospråklig ordbok, som vil være tilfellet for alle ord og fraser han har undersøkt, men heller å løse oversettelsen av dem som ikke står i ordboken, ofte fordi de nylig er dannet. Og dersom det også tas hensyn til flere konstruksjonstyper enn bare to jukstaposisjonerte substantiv, så øker antallet kandidater radikalt, og antallet nødvendige søk øker raskt så mye at det ikke lenger er gratis å utføre dem anno 2008.

Grefenstettes arbeider mer mer fokusert på brukeligheten til WWW som et korpus enn raffinering av rangeringsalgorimtnen basert på korpusdata, men føyer seg likevel inn i en rekke av prosjekter som bruker nettopp kvantitative data til å velge rett oversettelse, slik som også blir gjort i denne oppgaven.

2.3 Hanne Moas ReCompounder

Moa (2006) gjennomførte i sin masteroppgave arbeider som går videre på arbeidet til Grefenstette (1999) som viser at webben kan brukes til et troverdig korpus i lingvistisk disambiguering, og forsøkte å bruke webtreff til å rangere oversettelser av komposita som ikke står oppført i en tilgjengelig ordbok.

2.3.1 Konstruksjonstyper

Moa tar utgangspunkt i en leksikografisk definisjon av komposita, som sier at de er ord som består av to kjernemorfemer, som selv kan være komposita (Bergenholtz et al., 1997). Kjernemorfemer vil her si meningsbærende morfemer som kan stå alene, i motsetning til avledninger og bøyningsformer. Eksperimentene som utføres dreier seg likevel bare om substantivkomposita med substantiv som forledd, som også Grefenstette begrenset seg til. Mens Grefenstette valgte ut sine forsøksord ut fra at de hadde en oversettelse i en ordbok den forsøkte algoritmen kunne måles mot, åpner Moa for at norske substantivkomposita kan finne sine ekvivalenter i 5 engelske konstruksjonstyper. Disse var

- Jukstaposisjonering i ett, ord som i *rainforest*
- To substantiv ved siden av hverandre, som i *rain forest*
- To substantiv mellom preposisjonen *of*, som i *piece of mind*
- Et substantiv mellom preposisjonen *of* og et susbtantiv i flertall, som i *Master of puppets*
- To substantiv mellom preposisjonen *by*, som i *trip by foot*

Det er uklart hva som motiverer nettopp disse fem kategoriene, eksemplene er derfor bare ment som eksempler på hvordan konstruksjoner av disse typene kan se ut. Etter evalueringen av prosjektet ble to maler lagt til, et substantiv i flertall mellom *of* og et nytt substantiv, og et substantiv med genitivsmarkør etterfulgt av et nytt substantiv.

2.3.2 Kandidater

For å teste systemet ble en egenutviklet tagger brukt på 112 setninger fra et lite parallellkorpus[7], for å identifisere substantiv. Denne identifiserte 263

[7]Korpuset er utviklet av Ola Huseth ved NTNU i januar 2004. Utvalget sto oppført med både 112 og 114 setninger.

substantiv, som senere ble filtrert av en egenutviklet lemmatiseringsalgoritme inspirert av (Johannessen & Hauglin, 1996) og av manuell inspeksjon som reduserte tallet til 228. 151 av disse stod oppført i Kunnskapsforlagets store norsk-engelske ordbok (Eek, 2001), og ble fjernet for å fokusere nettopp på komposita som ikke hadde oppslag, i motsetning til det Grefenstette gjorde. Av de resterende 77 lot 57 seg oversette av algoritmen, og disse 57 oversettelsesforslagene ble evaluert. Årsaken til at 20 ikke lot seg oversette, var feil i filtreringen over, eller at enkeltord ikke stod i ordboken slik at det ble umulig å oversette, eller at oversettelseskandidaten ikke hadde noen treff på webben.

2.3.3 Oversettelse

Metoden som nå ble brukt til å oversette de resterende 57 kompositaene identifiserte først lemmaene de bestod av, eksemplifisert ved at *gårds·hus* består av lemmaene *gård* og *hus*. Disse ble i sin tur oversatt til engelsk, og dernest kombinert i overensstemmelse med konstruksjonstypene vist over. Dette eksempelet er hentet fra (Moa, 2006), og det resulterte i oversettelseskandidatene *farmhouse*, *farm house* og *house of farm*, osv. Disse kandidatene blir så søkt etter med frasesøk hos Google og Yahoo. Moa fant imidlertid at frasesøkene[8] returnerte mange treff hvor frasen ikke inngikk slik den stod skrevet, fordi skilletegn som «,.:» ble akseptert av søkemotorene. Algoritmen inspiserte derfor søkekonteksten[9] som ble returnert sammen med de første 10 treffene. Antallet webtreff som ble returnert per søk ble så justert med prosentandelen av de 10 første treffene som faktisk hadde søkestrengen i søkekonteksten.

2.3.4 Evaluering

Evalueringen av forsøkene foregikk ved at de første 5 treffene ble sammenliknet med 5 tilfeldig uttrukne treff fra kandidatlisten. En

[8] Søk i anførselstegn, som eksempelvis "*repair truck*".

[9] Søkemotorene returnerer 5 linjer med kontekst rundt hvert treff.

informant ble så forelagt disse to settene av 5 kandidater, og ble bedt om å vurdere dem. Informanten ble bedt om å vurdere om den høyest rangerte kandidaten var «ok» (36,9%), alternativt om minst en av de øverste 5 var ok (31,6%). Dersom ingen var dem var ok ble informanten bedt om å vurdere om et av forslagene kanskje kunne være ok.[10] (17,5%). Tallene i parentes viser hvor stor prosentandel av de 57 kompositaene som havnet i hver kategori. Det gjenstod da 14% som ikke kunne plasseres i noen av kategoriene nevnt over, og som ble vurdert som feilrate.

2.3.5 Sammenlikning med Grefenstette

Moa sammenlikner de tilsammen 86% av kompositaene hvor ikke samtlige kandidater produsert av systemet hennes ble vurdert som uakseptable av informanten med Grefenstettes suksessrate på 87% i oversettelsen av tyske komposita til engelsk. Forskjellene på eksperimentene er imidlertid så store at dette ikke er en rimelig sammenlikning. Grefenstettes forsøk gikk ut på å samle inn komposita fra en ordbok, slik at han visste at de hadde en ordsboksoversettelse, for å se om han kunne gjenfinne denne oversettelsen ved å kombinere oversettelsen av bestanddelene. Moa siler derimot bort alle slike ord, og konsentrerer seg om ord hvis oversettelse ikke finnes i ordboken. Jamfør diskusjonen av leksikalisering i avsnitt 3.1, så er ikke en oppføring i en ordbok noe sikkert tegn på leksikalisering, men det er likevel rimelig at Grefenstettes 744 komposita som står i ordboken er mer leksikaliserte som gruppe en de 57 utvalgt av Moa hvor ingen av dem har ordbokoppslag. Videre så har Grefenstette en automatisk evalueringsmetrikk, som sjekker om algoritmen hans kommer frem til det som står i ordboken, en «gullstandardevaluering» i betydningen at det er en kanonisk fasit, mens Moas evaluering er en svært liberal «sølvstandardevaluering» hvor informanten må utelukke at noen av de 5 øverst rangerte oversettelseskandidatene *kanskje* kan være ok for å registrere en feil. Fordi de norske kompositaene til oversettelse ble hentet fra et parallellkorpus, eksisterte allerede en oversettelse av samtlige av dem.

[10] Engelsk original: *None ok but at least one maybe*.

Hvis de hadde blitt brukt i evalueringen ville det liknet mer på det Grefenstette gjorde. Grefenstettes eksperiment hadde som et av sine mål å vise at webben var så stor at dets rene enormitet kompenserte for støy, og brukte dette som et argument for at grunn prosessering[11] kan fungere. Moa går imidlertid bort fra dette når hun innfører en justert webtreff-telling, og utfører en nærmere analyse av søkekonteksten. Det vises bare et eksempel på tall fra både rene og justerte webtreff, og i det tilfellet ville rangeringen blitt den samme, så det er derfor uklart hvilken effekt dette hadde. Dette utgjør en forskjell også i metodisk tankegang, som i tillegg til at selve oppgaven som skal løses er en annen gjør at eksperimentene synes kvantitativt inkommensurable.

2.4 Baldwin og Tanaka

Timothy Baldwin og Takaaki Tanaka har gjort flere arbeider på oversettelse av substantivkomposita, og deres siste arbeid på oversettelse mellom japansk og engelsk (begge retninger) danner grunnlaget for eksperimentene som er utført i denne oppgaven. Utviklingen i metodikk er likevel interessant, og et kort omriss av denne vil bli antydet i dette avsnittet. Samtlige arbeider er gjort på oversettelse mellom japansk og engelsk.

2.4.1 Dyp og grunn oversettelse

Tanaka og Baldwin (2003) gir en oversikt over måter å oversette substantivkomposita på, hvor de gjør et skille mellom dyp og grunn analyse som utgangspunkt for oversettelsen. Forskjellige måter å uttrykke samme semantiske forhold mellom språk motiverer en dyp analyse. Eksempelvis er det vanskelig å se for seg at meningsinnholdet i det norske ordet *slapp·fisk* kan overføres til engelsk ved hjelp av en grunn analyse. Ordet har utviklet seg fra å betegne en fisk som er *slapp* i kjøttet, i motsetning til *fast*, til å betegne en person som har en konsistens

[11] Engelsk: *shallow processing*.

som minner om en slapp fisks, brukt som en indikasjon på at denne kan utføre mindre den burde. Senere har det også blitt brukt til å betegne blant annet biler[12] som yter mindre enn forventet. (Tanaka og Baldwin bruker eksempelet *idobota·kaigi*, som direkte oversatt til engelsk blir *well-side meeting* (norsk: *møte ved brønnen*), men betyr *idle gossip* (norsk: *tomgangs·sladder*).

Forfatterne gjør så et annet skille enn mellom dyp og grunn analyse, og deler inn oversettelsesstrategier i minnebaserte og dynamiske. De minnebaserte metodene kalles dette fordi de har et begrenset antall komposita som kan oversettes «i minne». Den enkleste strategien, kalt MBMT_{DICT}[13], bruker en vanlig ordbok til å hente inn oversettelsespar, MBMT_{ALIGN} henter oversettelsespar fra et parallellkorpus ved å studere ordenes relative plassering (engelsk: *alignment*), og MBMT_{COMP} henter ut komposita fra et kildespråkkorpus, og oversetter deres bestanddeler ved å sette dem inn i oversettelsesmaler, i likhet med teknikker beskrevet ovenfor. De forholder seg bare til komposita som skrives i to ord, slik at oppstykkingsproblematikk som er relevant for tysk og norsk ikke blir det her. Den mest plausible kandidaten blir så valgt ut ved å «bruke evidens fra et målspråkskorpus»[14] uten at det forklares nærmere hvordan.

De nevner to dynamiske strategier, DMT_{COMP}[15], som er identisk med MBMT_{COMP} med unntak av at et hvilket som helst kompositum kan settes inn, og ikke bare dem hentet fra et korpus, og DMT_{INTERP} som tolker kompositumets semantikk, og gir dette en semantisk mellomrepresentasjon, som brukes til å generere en oversettelse.

Tanaka og Baldwin gjør så et eksperiment hvor de sammelikninger MBMT_{DICT} og DMT_{COMP}. Siden disse to metodene begge baserer seg på grunn prosessering, så tilbyr studien ingen sammenlikning av grunne og dype strategier. Argumentet for å utelate DMT_{INTERP} er at en dyp analyse

[12]http://www.dn.no/dnBil/article1061194.ece besøkt 24 juli 2008.

[13]Forkortelsen står for *memory-based machine translation*.

[14]Engelsk original: *Use empirical evidence from the target language corpus to select the most plausible translation candidate.*

[15]DMT står for *Dynamic machine translation*.

ikke har noen fordel overfor den grunne analysen fordi at komposita hvor kontekst er nødvendig for å finne frem til oversettelsen beholder sitt tolkningsrom dersom det oversettes ord for ord. Dermed vil en dyp analyse bare medføre en stor økning i produksjonskostnader, uten noen gevinst i ytelse. For å illustrere bruker de eksempelet *apple-juice seat* (norsk: *eple·jus·sete*) som kan bety både et sete noen har sølt eplejus på, eller der man må sitte for å drikke/få eplejus og andre tolkninger[16]. Eksempelet er godt, men likevel er ikke argumentet overbevisende. En dyp prosessering ville kunne oversette komposita som *slapp·fisk* nevnt over, det vil også kunne oversette komposita som har forskjellig betydning i forskjellige sammenhenger riktig, som er umulig med en grunn analyse, og det vil kunne uttrykke det semantiske innholdet med helt andre konstruksjoner enn i kildespråket, som å gå fra substantivkompositum til ett ord, eller fra nominal til verbalfrase. Et system med slike ferdigheter ville vært kostbart å utvikle, det er lett å forutsette et system med disse egenskapene men verre å utarbeide. Likevel kan det ikke avvises på teoretisk grunnlag fordi at enkelte fraser deler flertydighetsspektrum mellom språk.

2.4.2 Rangering etter CTQ

Tanaka og Baldwin (2003) utdyper og utvider den «grunne» strategien MBMT_{COMP}, spesielt blir rangeringsaspektet som manglet i forrige avsnitt grundig gjort rede for. Dette arbeidet dreide seg om oversettelse av japanske substantivkomposita (med substantiv som forledd) til engelsk. Det ble ikke gjort noen diskusjon av kompositumsbegrepet i forhold til japansk ut over at de kan anta formen som beskrevet under. Eksperimenter ble utført for å teste metoden.

Testdata

Oversettelsesprosessen blir mer formelt forklart, og deles nå inn i (a) å generere kandidater, *generering*, og (b) å velge blant dem, *utvelgelse*.

[16]De viser til en japansk oversettelse av ordet, men poenget er gyldig også med en norsk oversettelse.

Forutsetningene er de samme, 500 kandidater til oversettelse hentes ut fra et korpus[17]. Korpuset ble analysert med ALTJAWS[18], som deler teksten opp i ord og angir ordklasse for dem. Alle instanser hvor et substantiv etterfulgte et annet (kalt et NN-bigram) ble holdt til side. Disse ble så filtert, slik at alle som forekom mindre enn 10 ganger ble luket vekk. Så ble de 250 oftest brukte bigrammene valgt ut, sammen med 250 vilkårlige bigram.

Generering

For hvert japanske bigram ble hvert ord oversatt til engelsk, og kryssproduktet av oversettelsene satt inn i maler for å produsere listen over oversetteleskandidater. Kryssprodukter definert slik at hvis S og T er mengder (slik som oversettelsen av *ungdom* kan være mengden av ord utgjort av *youth* og *adolescent*), så er kryssproduktet av S og T mengden av alle par <s,t>slik at s er med i (er en del av S), formelt uttrykt $s \in S$, og t er med i T, formelt uttrykt $t \in T$.

Hvis mengden S er oversettelsene av *ungdom* (*youth* og *adolescent*) og mengden T oversettelsene av *skole* (*school* og *college*) blir kryssproduktet av mengdene utgjort av parene (*youth*,*school*), (*adolescent*,*school*),(*youth*,*school*) og (*adolescent*,*college*). Disse parene kan settes inn i maler for å utarbeide oversettelseskandidater til ordet *ungdomsskole*.

Malene kan se ut som eksempelvis [$N_1{}^E$ $N_2{}^E$] som forteller at det første japanske ordets oversettelse skal settes inn i posisjon 1 på den engelske kandidaten, hvis det er et substantiv, og det andre ordet tilsvarende i posisjon 2, eller [$N_2{}^E$ *in* $N_1{}^E$] som forteller at det oversettelsene skal stilles opp i motsatt rekkefølge, med preposisjonen i mellom. Eksempler på bruk av malene kan henholdsvis resultere i oversettelseskandidatene *machine translation* og *translation in machine*. Hvis *M* er antall maler, *m* og *n* fertiliteten, i betydningen antall mulige oversettelser, av hvert japanske ord, så er *O(m,n,M)* mengden oversettelseskandidater for hvert japanske

[17]Mainichi Newspaper Co., 1996.

[18]http://www.kecl.ntt.co.jp/icl/mtg/resources/altjaws.html.

kompositum.

Som i forrige artikkel så danner de et antall kandidater til en oversettelse av et japansk kompositum ut ifra et utvalg av maler, og fertiliteten til de enkelte ords bestanddeler. Med to maler og to ordbokoppslag for hver bestanddel, ender dermed opp med $2 * 2^2 = 8$ kandidater. Dette sluttfører genereringen av kandidater.

Utvelging

I utvelgelsesfasen innfører de så begrepet **corpus-based translation quality** (ctq) som er et interpolert beregning av sannsynligheten for at en generert kandidat opptrer i et korpus. I denne omgang presenterer de den som

$$CTQ(w_1^E, w_2^E, t) = \alpha p(w_1^E, w_2^E, t) + \beta p(w_1^E, t)p(w_2^E, t) + \gamma p(w_1^E, t)p(w_2^E, t)p(t) \tag{2.1}$$

hvor w_1^E står for oversettelsen av det første japanske ordet til engelsk, w_2^E står for oversettelsesen av det andre ordet, *t* står for maler (engelsk: *templates*) og $0 \leq \alpha, \beta, \gamma \leq 1$ og $\alpha + \beta + \gamma = 1$. Oversettelsen av de japanske ordene til engelsk kan være en frase.

Probabilitetene regnes ut fra *maximum likelihood estimate based on relative corpus occurrence*, (norsk: *maksimal trolighet basert på relativ korpusopptreden*), et anslag på hvor sannsynlig det er at de relativt sett opptrer i korpuset. Dette ved siden av at probabilitetene også omtales som betingede i artikkelen, (*conditioned on translation templates*) gjør det uklart hvordan probabilitetene regnes ut. Den matematiske formelen gir imidlertid inntrykk av at det er probabiliteten av snittet av hendelsene[19], av at malen og oversettelseskandidaten opptrer på likt, som regnes ut, og ikke sannsynligheten for at oversettelseskandidaten observeres gitt konstruksjonstype.

Grunnlag for beregning av CTQ brukte den skriftlige komponenten av

[19] engelsk: *joined probability*.

British National Corpus, som var dependentanalysert (*dependency parsed*) med taggeren RASP (E. Briscoe & Carroll, 2002) (versjon 1). En nærmere presentasjon blir gitt i avsnitt 4.5. Ut fra resultatet av denne parsingen kan substantiv-substantiv-relasjoner fanges opp med med *ncmod*-relasjonen. De bruker setningensfragmentet *the Jubjub bird's relation to the fruminous Bandersnatch*[20] som eksempel, som gir disse relasjonene:

```
ncmod(_,bird,Jubjub)
ncmod(POSS,relation,bird)
ncmod(to,relation,Bandersnatch)
```

Disse tre relasjonene representerer tre engelske nominelle konstruksjonstyper. Den første et substantivkompositum med to substantiv plassert rett ved siden var hverandre, som tilsvarer en mal av formen [$N_1{}^E$ $N_2{}^E$], den neste en possesivrelasjon som tilsvarer en mal av typen [$N_1{}^E$*'s* $N_2{}^E$] og den siste en mal av typen [$N_2{}^E$*to* $N_1{}^E$]. Ved å summere over slike analyser av hele korpuset kunne de regne ut probabilitetene over som er brukt til å regne ut *CTQ*.

Evaluering

For å evaluere, ble de 500 kandidatkompositaene vurdert mot sin oversettelse i ordboken ALTDIC (Ikehara, Shirai, Yokoo & Nakaiwa, 1995) med omtrentlig 400 000 oppslag hvorav halvparten er substantiv. De sammensetningene hvis engelske oversettelse, men ikke mer (flere ord), kunne gjenfinnes av maler som vist over, ble samlet i et datasett som ble kalt ALIGN_{GOLD}, hvor ordbokoversettelsen ble ansett som fasit. De fant 224 slike sammensetninger.

Algoritmen ble så brukt på disse 224, for forskjellige verdier av α og β. $\alpha = 1$ og $\beta = 1$ ble brukt som referanseverdier, den første velger den beste kandidaten etter hvor ofte den forekommer i sin fullstendige form, og i den andre etter hvor ofte de respektive engelske ordene opptrer

[20] Fragmentet er hentet fra Lewis Carrols eventyrverden, og vanskelig å oversette, så dette blir utelatt både fordi det er vanskelig og at det ikke bidrar til diskusjonen.

i en gitt kosntruksjonstype som henholdsvis forledd eller etterledd. Ytelsen ble målt etter to kriterier, *precision* (presisjon) som i denne sammenhengen betegner andelen av sammensetninger hvor algoritmen fant en riktig oversettelse, og *recall* (gjenkalling) som betegner andelen hvor algoritmen kunne generere en oversettelse. F_1*-score* betegner det harmoniske gjennomsnittet av disse, gjengitt i likning 2.2.

$$F_1 = 2 \cdot (\text{precision} \cdot \text{recall})/(\text{precision} + \text{recall}) \tag{2.2}$$

I tillegg til referanseverdiene, så ble det søkt etter verdier for α og β, (det ble ikke oppgitt hvilke inkrementeringsverdier som ble brukt), og høyere F-score-verdier ble funnet for alle datasett med verdier som skilte seg fra referanse, som betyr at alle tre ledd i likning 2.1 ble tillagt vekt.

2.4.3 Maskinlæring og tospråklige leksika

Baldwin og Tanaka (2004) presenterer et nytt utviklingstrinn i oversettelsen av NN-komposita. En revidert utgave av *CTQ* blir brukt til å oversette japanske og engelske komposita begge veier. Metoden henter sine komposita til oversettelse fra et korpus, slik at fremgangsmåten er lik den beskrevet som MBMT$_{COMP}$ i avsnitt 2.4.1, selv inndeling av dem er noe endret. Men den viktigste nyvinningen er at denne revisjonen blir målt mot en maskinlæringsteknikk i form av en *Support Vector Machine* (støttevektormaskin), heretter kalt en SVM.

Oversettelsesdata

For å finne kandidater til oversettelse, ble Reuters-korpuset, som er en del av korpuset beskrevet i avsnitt 4.2.2, og samme aviskorpus brukt for japansk som beskrevet i avsnitt 2.4.2. Reuters-korpuset ble først tagget med fnTBL (Ngai & Florian, 2001) og det japanske med ChaSen (Matsumoto, Kitauchi, Yamashita & Hirano, 1999), og deretter ble alle NN-bigram holdt til side. Denne listen ble filtrert slik at bare de som var etterfulgt av et ikke-substantiv ble beholdt. Videre ble entropien for

konteksten til venstre og høyre for bigrammet målt for å filtrere vekk forekomster hvor kontekst-entropien var under 1. Det er uklart hvor mye kontekst som regnes med her, og også hvordan den er regnet ut, så det er vanskelig å evaluere metodikken, men den har til hensikt å ekskludere NN-bigram som var en del av en større leksikal enhet.

Entropi som begrep i informasjonsteorien er formelt definert som i likning 2.3, hvor X er en diskret variabel som kan anta verdiene $x_1 ... x_n$.

$$H(X) = -\sum_{i=1}^{n} p(x_i) \cdot \log p(x_i) \tag{2.3}$$

Dersom distribusjonen er uniform, slik at alle hendelser er like sannsynlige vil entropien bli størst. Dersom sannsynlighetsfordelingen er skjev, slik at den har en slagside, så vil entropien gå ned. Ved å ha en slik nedre terskel, vil dermed komposita som har det samme ordet foran seg hver gang det forekommer kunne utelukkes. Derfor ble terskelen ikke brukt dersom bestemt eller ubestemt artikkel, *the* og *a(n)* forekom på venstre siden av kompositumet, eller skilletegn på dets høyre. Baldwin og Tanakas eksempel er gjengitt i eksempel 1, som ekskluderer *service department* fra å tas med.

(1) social service department
sosial tjeneste avdeling
avdeling for sosiale tjenester

Eksempelet er ikke spesielt godt, fordi at et raskt frasesøk på Googles søkemotor viser at *social service department* har 82 400 treff, og *service department* har 7 220 000[21]. Selv om *social service department* slik det er tolket her har en struktur hvor *social service* er enhet som modifiserer *department* og dermed ekskluderer at *service deparment* selv er enhet, så tyder nettsøket på at uttrykket opptrer så ofte i andre sammenhenger at det er konvensjonalisert nok til at det bør tas med. Kontekst tas ikke med i metodene for oversettelse. Det er mulig at dette er en tilfeldighet, og at

[21]Søk gjennomført 25. juli 2008.

service department bare forekommer i sammenheng med *social* i Reuters-korpuset, og at terskelens utelukker andre NN-bigram som ikke kan analyseres som komposita, men dette er det eneste eksempelet som ble gitt.

De kompositaene som ble hentet ut etter kriteriene nevnt over ble så delt inn i tre «frekvensbånd», som er en inndeling etter kompositaenes frekvens i grunnlagskorpuset. Grensen for å komme med i det høye frekvensbåndet var eksempelvis 336 for de japanske, og 344 for de engelske kompositaene. 250 kandidater ble trukket ut fra hvert av dem, og disse 750 kandidatene ble så oversatt med ALTDIC (Ikehara et al., 1995) og EDICT (Breen, 1995), samt et stort innslag av manuell oversettelse fra forfatterne[22]. I tillegg ble 0,5% av de japanske og 6,6% av de engelske kandidatene forkastet etter manuell inspeksjon.

Generering av oversettelseskandidater

Maler, slik som beskrevet i avsnitt 2.4.2 ble også brukt til å lage mulige oversettelser av hvert inngående kompositum. Det ble brukt 28 maler i oversettelse fra japansk til engelsk, og 4 motsatt vei. De ble laget ved å se på *alignment* mellom de utvalgte kandidatene til oversettelse og deres gullstandard-oversettelser. *Alignment* forstås som systematikken i hvordan ordene plasserer seg i forhold til hverandre. Det første japanske substantivet kan forekomme i enten den første eller andre posisjonen i oversettelsen. Hvis et kompositum stilles opp over sin oversettelse, kan det forstås som en *linjering*, en identifikasjon av de linjer som kan trekkes mellom hvert ord og dets oversettelse. Men det er jo ikke bare to maler som er generert, som er antall måter å trekke slike linjer på. Også skift i andre syntaktiske egenskaper som tall, kasus blir tatt med i malene, samt om preposisjoner og/eller artikler fremstår som en del av uttrykket mellom de to oversettelsene. Dermed kan malene ses på som en formel for å komme frem til den riktige oversettelsen på syntaktisk nivå, hvis man har fått oversettelser å sette inn i formelen som har med seg det rette

[22] 25% av de engelske, og 35% av de japanske kompositaene ble oversatt manuelt.

betydningsinnholdet på semantisk nivå.

Listen av oversettelseskandidater blir laget ved å sette inn kysspro-duktet av oversettelsene av enkeltordene fra hvert NN-bigram inn i disse «formlene».

Rangering etter CTQ

Med unntak av en revisjon av CTQ, så foregikk valget av den presumptivt beste oversettelseskandidaten på samme måte som i avsnitt 2.4.1. Evaluering av likning 2.1 i (Tanaka & Baldwin, 2003) viste at den tredje termen i likningen ikke påvirket ytelsen i nevneverdig grad. Dermed stod man igjen med likning 2.4. Videre ble α og β-verdier satt til henholdsvis .9 og .1 for alle eksperimentene, også som en følge av denne evalueringen.

$$CTQ(w_1^E, w_2^E, t) = \alpha p(w_1^E, w_2^E, t) + \beta p(w_1^E, t) p(w_2^E, t) p(t) \tag{2.4}$$

Rangeringen etter denne formelen ble brukt som referanse i dette eksperimentet. Evaluering viste også at problemet med denne måten å velge på ikke kunne skille mellom hvor vanlig eller perifer de ulike oversettelsene av hvert ledd i sammensetningen er. Dermed kan en situasjon oppstå hvor et ord er en perifer oversettelse av sin kilde, men samtidig et mye mer vanlig ord gjør at oversettelseskandidaten med dette ordet blir foretrukket. Eksempelet som ble brukt til å illustrere dette er *related item* (*relatert enhet*) som blir foretrukket fremfor *related article* (*relatert artikkel*) som er riktige oversettelsen i følge gullstandarden. Dersom *article* ble premiert for å være en vanligere oversettelse av sitt utgangspunkt ville riktig oversettelse bli valgt av algoritmen[23]. Derfor ble en alternativ utvelgelsesstrategi basert på en SVM brukt til å avhjelpe dette.

Rangering etter SVM-maskinlærer

Ved å anvende en SVM-maskinlærer kunne lingvistisk intuisjon repre-sentert ved CTQ kombineres med enspråklig informasjon (data hentet fra un-

[23]Det japanske ordet som dannet utgangspunkt for oversettelsene var *kaNreN·kiji*.

dersøkelser på ett språk) og informasjon fra et tospråklig leksikon. All denne informasjonen kan også brukes i en utvidelse av CTQ, men i så tilfelle måtte interpolering gjøres over for mange termer. Å bruke en maskinlæringsteknikk var en måte å unngå dette på, sommotiverte metodikken. TinySVM[24] var implementasjonen som ble valgt. En kort presentasjon av SVM-konseptet følger i avsnitt (5.1).

Baldwin og Tanaka bruker SVM-en binært, slik at den kan skille et utvalg oversetteleskandidater som enten riktige eller gale. Maskinen regner ut koeffisientene til en funksjon, kalt klassifikatorfunksjon eller bare klassifikator, som avgjør hvorvidt en ny observasjon ligger nærmest den gale eller riktige siden. Siden operasjonen er binær, kan de to sidene symboliseres med 1 og -1, og fortegnet til returverdien fra klassifikatoren avgjør klassifikasjonen. Måten klassifikatoren blir brukt i eksperimentet er å rangere etter returverdien, slik at jo høyere returverdien fra klassifikatoren er, jo bedre blir kandidaten ansett.

Alle oversettelseskandidater puttes først i en av to sekker, en for korrekte og en for gale, og en SVM brukes til å finne hyperplanet som skiller deres trekkvektorer. For hver kandidat ekstraheres en vektor med et antall trekk. Dette hyperplanet vil da skille en *riktig* oversettelse fra en *gal* oversettelse. Hvis hver av de 750 inndataene hadde igjennomsnitt 10 oversettelser[25], så vil planet skille mellom de 750 *riktige*, og de 6250 *gale* oversettelsene. Når klassifkatoren anvendes på en ny observasjon som SVM-en ikke er trent på, vil denne predikere hvorvidt den tilhører den *riktige* eller *gale* kategorien. Dette er en måte å klassifisere oversetteleskandidater på mellom gode og dårlige sådane. Problemet som ønskes løst er å *rangere* oversettelser av *noe*. Måten Baldwin og Tanaka har brukt SVM-en tar ikke høyde for at en oversettelseskandidats korrekthet betinges av det som skal oversettes. Samme oversettelseskandidat kan være både riktig og gal, avhengig av utgangspunktet. Det er uklart hvilken effekt en betinget rangeringsstrategi ville hatt på eksperimentet som ble gjennomført, fordi det

[24] `http://chasen.asit-nara.ac.jp/~taku/software/TinySVM`.

[25] Tallet er ukjent, men ligger sannsynligvis mye høyere, men det eksakte tallet er ikke nødvendig for å føre argumentet.

er ikke sikkert at samme oversettelseskandidat hadde opptrådt som både riktig og gal (et eksemplar av samme oversettelse i hver sekk), men i en reell implementasjon av større skala ville dette kunne få betydning.

En måte å bruke en SVM på som tar hensyn til at en oversettelse er en oversettelse av *noe*, slik at den skiller mellom gode og dårlige oversettelser av et gitt kompositum, og ikke strenger, er å formulere SVM-problemet slik at maskinen prøver å finne planet som skiller gode fra dårlige oversettelser *med bibetingelsen* at den skal ha færrest mulige tilfeller hvor to kandidater som har samme utgangspunkt blir rangert slik at den som er markert som *gal* foretrekkes framfor den som er markert som *riktig*. En måte å gjøre dette på er å gå igjennom treningsdataene, og å regne ut vektordifferansen mellom par av oversetteleskandidater med samme utgangspunkt hvor rangeringen er forskjellig. Disse vektordifferansene samles til et nytt sett treningsdata, som kan brukes som inndata til en SVM som slik det gjøres ved klassifikasjon. Se (Velldal, 2008) for en diskusjon av hvordan SVM-er kan brukes i rangeringsøyemed.

Trekk brukt i SVM-en

Trekkene som ble brukt gjengis i tabell 2.2. Trekkvektoren for hver oversettelseskandidat ble satt sammen av enspråklige trekk, som også var tilgjengelig for CTQ-rangeringen over, i tillegg til tospråklige trekk som kom fra oppslag i flere ordbøker. De enspråklige trekkene gir gir informasjon korpuset om bruk, mens de tospråklige gir et bilde av hvor ofte oversettelsen og delene av den er en direkte oversettelse av sitt utgangspunkt ut fra antagelse om at dersom oversettelseskandidaten er en oversettelse av kompositumet den skulle oversette, så vil den oftere være korrekt. For de enkeltordene, så vil en hyppig frekvens av en oversettelse gitt utgangspunktet kunne gi et bilde av dette er en vanligere måte å oversette ordet.

MWE$(w_1^{L_2}, w_2^{L_2}, t)$ er en normaliseringsparameter for *Multiword Expressions* (*norsk: flerordsuttrykk*), uttrykk bestående av flere ord, som regner ut hvor ofte hele uttrykket forekommer relativt til forekomsten av kjernen

Enspråklige	Tospråklige
frekvens($w_1^{L_2}, w_2^{L_2}, t$)	frekvens($w_1^{L_2}, (w_2^{L_2}, t \mid (w_1^{L_1}, w_2^{L_1}, t)$
frekvens($w_1^{L_2}, t$)	frekvens($w_1^{L_2}, t$)
frekvens($w_2^{L_2}, t$)	frekvens($w_1^{L_2}, t \mid w_1^{L_1}$)
frekvens($w_1^{L_2}$)	frekvens($w_2^{L_2}, t \mid w_2^{L_1}$)
frekvens($w_2^{L_2}$)	frekvens($w_2^{L_2} \mid w_2^{L_1}$)
frekvens(t)	frekvens($w_2^{L_2} \mid w_2^{L_1}$)
CTQ($w_1^{L_2}, w_2^{L_2}, t$)	
MWE($w_1^{L_2}, w_2^{L_2}, t$)	

Tabell 2.2: *Enspråklige og tospråklige trekk fra Baldwin og Tanaka (2004). Fordi disse forsøkene ble gjort i begge retninger mellom engelsk og japansk, brukes L_1 og L_2 til å betegne henholdsvis kildespråk og målspråk. ${w_1}^{L2}$ betegner dermed det første ordet (frasen) i oversettelsen til målspråket, L_2, mens ${w_2}^{L1}$ betegner det andre ordet (frasen) kildespråket, del 2 av det kompositum som skal oversettes.*

i uttrykket, definert som det siste ordet. Hvis en oversettelse w_1^2 eller w_2^2 består av mere enn ett ord, så blir de enspråklige frekvenstrekkene beregnes ved å multiplisere forekomsten av kjernen i den aktuelle konteksten (malen), med denne normaliseringsparameteren. Selve trekket MWE beregnes ved å multiplisere MWE-parameterne for w_1^2 og w_2^2 med hverandre. I de tilfellene oversettelsene består av kun ett ord settes parameterne til 1, som for at systemet skal foretrekke disse.

I tillegg kommer to trekk som beregnet ut fra malen som brukes til å generere den aktuelle oversettelseskandidaten, som identifiserer henholdsvis hvilken mal som ble brukt til å generere den, og hvilket ord i oversettelsen som er kjernen i uttrykket, ved siden av en normalisert versjon av de andre trekkene, relativt til deres maksimale verdi, som fordobler antall en- og tospråklige trekk.

Hvert av disse trekkene resulterer i en numerisk verdi, og i tillegg til hver numeriske verdis andel av den høyeste målte sum brukt i trekkvektoren. Når det var 8 enspråklige og 6 tospråklige trekk, blir det til sammen $8 \times 2 \times 2 = 32$ trekk. Hver oversettelseskandidat gir en trekkvektor $[0 : x_1, 1 : x_2, 2 : x_3 : x_n]$ hvor x står for den numeriske verdien tilknyttet hvert trekk ved oversettelseskandidaten. SVM-en avgjør så i

Frekvensbånd	Referanse		CTQ		SVM_{full}	
	G	S	G	S	G	S
Høyt	.425	.789	.445	.806	.462	.857
Middels	.315	.665	.368	.797	.480	.878
Lavt	.210	.393	.280	.569	.320	.720

Tabell 2.3: *Fra Baldwin og Tanaka. Utvalg av resultater fra oversettelse mellom japansk og engelsk.*

hvor stor grad hver av trekkene bidrar til å indikere enten god eller dårlig oversettelse, slik at modellen kan brukes til å predikere denne egenskapen ved en ny oversettelseskandidat av et nytt kompositum.

Evaluering

Metodene blir evaluert først ved å sjekke hvorvidt den høyest rangerte oversettelseskandidaten er gullstandardoversettelsen, en enkel binær vurdering. Ved siden av dette innfører Baldwin og Tanaka en «sølvstandard», som forsøker å etablere L1-gjenfinnbarhet[26] hos de høyst rangerte kandidatene. For å vurdere om en oversettelseskandidat er nettopp L1-gjenfinnbar, så skal den inneha de grunnleggende semantiske egenskapene av det som skulle oversettes, slik at det er mulig å finne tilbake til dette opprinnelige kompositum *med rimelig sikkerhet*[27].

Resultater

Det ble utført et referanseeksperiment som rangerte bare etter hvor ofte oversettelseskandidaten opptrådte i sin helhet i det ettspråklige korpuset (det samme som å sette $\alpha = 1$ og $\beta = 0$ i CTQ), samt rangering etter CTQ, så en SVM-lærer som brukte bare enspråklige trekk, bare tospråklige trekk, og avslutningsvis en kombinasjon av begge. Resultater for oversettelse fra japansk til engelsk gjengis i tabell 2.3.

Resultatene viser en klart stigende trend etter hvor rangeringsverktøy.

[26] Engelsk: *L1-recoverability*.

[27] Original engelsk tekst: *..with reasonable confidence.*

Både for gullstandard og sølvstandard gir SVM-en bedre ytelse. Videre er sølvstandardtallene svært gode over alle frekvensbånd, mens gullstandardytelsen er noe mangelfull for det laveste frekvensbåndet. Det er bare 100 av 250 komposita hvor gullstandardoversettelsen var en del av de genererte kandidatene, og av dem var 32% riktig. Det vil si at 32 av 250, eller 12,8% av kompositaene valgt ut for oversettelse fikk en korrekt oversettelse.

Forfatterne avviser at SVM-en har en urimelig fordel fremfor CTQ fordi den kan ta tospråklige trekk med i betrakningen, fordi så få av kompositaene (12%) står med oppslag i ordboken, og at bare 65% av disse oppslagene ble inkludert i gullstandarden. Men det er ikke bare i det tilfellet selve kompositumet står i ordboken de tospråklige trekkene vil komme til nytte, men de vil også gi et bilde av hvor ofte ett ord som opptrer som delledd i en oversettelseskandidat er en oversettelse av sitt utgangspunkt. At dette mangler i CTQ ble påpekt som en svakhet ved CTQ selv, og en motivasjon til å bruke tospråklige trekk. Det er praktiske hensyn som taler mot å inkludere dette i CTQ (interpolering over for mange termer), og ikke teoretiske. SVM-en kommer frem til en modell, slik at det ville vært mulig å presentere de ulike trekkenes relative betydning. Sammenlikningen mellom SVM-en der den bare bruker enspråklige trekk med CTQ er derfor mest interessant. SVM-en presterte bedre også med bare enspråklige trekk, men med en langt mindre margin enn dersom samtlige trekk kom til anvendelse.

Kapittel 3

Komposita og oversettelse

I dette kapittelet følger en presentasjon av komposita fra en lingvistisk synsvinkel, og en diskusjon av hvilke avgrensinger som er relevante for denne oppgaven. Diskusjonen om ordsammensetninger og deres egenskaper som kan føres helt tilbake til de gamle sanskrit-grammatikerne er mangslungen, og en presentasjon av deler av denne diskusjonen vi bli gitt for å skape en ramme rundt prosjektet. Forståelsen og tolkningen av begrepet *kompositum* hviler på hvordan *komposisjonalitet* blir forstått, hvis tolkning igjen bygger på hvordan *leksikalisering* og et språks leksikon forstås. Diskusjonen av dem vil derfor bygges opp etter denne listen i motsatt rekkefølge. Denne oppgaven er praktisk i sin natur, og ikke ment som et teoretisk bidrag til en filosofisk diskusjon av semantikk og sammenhenger som den har foregått gjennom århundrer og årtusener, men vil likevel bli diskutert for å gi en sammenheng som eksperimentenes forutsetninger og resultater blir tolket innenfor.

En semantisk diskusjon i ordets snevre og hverdagslige betydning, ords definisjon, er nødvendig på grunn av forskjeller mellom norsk og engelsk. Det engelske begrepet *compound* blir brukt parallelt med det norske *kompositum*, men samtidig vil en diskusjon om *komposisjonalitet* bli gjennomført. Fordi det muligens kan virke som en nødvendighet at et *kompositum*, noe som er komponert, må være *komposisjonelt*, enn et *compound* presiseres det at de to begrepene forstås som det samme i den

videre diskusjonen. Såvel det norske som det engelske begrepet kan føres tilbake til det latinske *componere*, som selv kan tjene som eksempel, da det er et kompositum med strukturen *com·ponere* med betydningen *sammen·å føre*[1]. Når et kompositum er en sammensetning, følger det at det er satt sammen av noe. Men selv om den morfologiske oppbygningen kan vise til en slik komposisjon, er det ikke like opplagt at det semantiske innholdet er bygd opp som en tilsvarende komposisjon, hvor bestanddelene kan pekes på.

3.1 Om leksikalisering

I grove trekk handler leksikalisering om at nye enheter tas opp i leksikonet, forteller Bakken (2006) i sin diskusjon av leksikalisering. Her menes ikke et leksikon som i ordets normale forstand, dvs. som i et familieleksikon, men som i et gitt naturlig språks leksikon, dets antatte lagerbeholdning av atskilte termer. Selv om hvorvidt et ord er leksikalisert ikke kan avgjøres av redaktører, slik som i et leksikon i vanlig forstand, så kan denne redaksjonelle prosessen være en tjenlig analogi. Nettopp synet på dette leksikon og hva det er, vil derfor være avgjørende for å avgjøre hvordan begrepet «leksikalisering» tolkes slik det er brukt.

Bakken beskriver et arbeid av Pawley (1986) som antyder en kontrast mellom et leksikon i tradisjonell strukturalistisk eller generativ forstand, og nettopp begrepet leksikon slik som leksikografer eller endog lekfolk forstår det. Strukturalisme forstås i denne sammenheng som et system som kan beskrive hvordan en komplekst system som naturlig språk fungerer, ut ifra et leksikon av ord, og hvilke prosesser som virker på dem for å si om en setning er grammatisk, eller hvilken morfologi som vil bli anvendt. Bakken forteller deretter om økt interesse for leksikon innenfor lingvistisk forskning, og at den «leksikografiske» forståelse av leksikonet er mer forenlig med bruksrettede rammeverk som kognitiv lingvistikk.

[1] Wiktionary-oppslag besøkt 8. juli 2008. URL: `http://en.wiktionary.org/wiki/compound`.

Her anses leksikonet for å være den fremste skueplass for menneskelig språkevne, hvis organisasjon kan beskrives som et nettverk, og skillet mellom produktive regler og lagrede enheter er i høy grad visket ut.

3.1.1 Leksikalisering og orddannelse

Produktive orddannelser danner nye ord, men i det de kommer til, så trer de ut av ordannelsesgrammatikken og inn i en annen sfære. Veien videre til en eventuell leksikalisering foregår i følge andre prinsipper enn dem som lå til grunn for deres dannelse. En av disse er å anse leksikalisering som ferden mot en bruk av uttrykket som en konvensjon. Tanken er at for at et ord skal bli en del av et sosialt eller mentalt leksikon, så må det etableres en konvensjonell sammenheng mellom konseptet og den nye ordformen. Et godt eksempel på dette kan være kompositumet *scooter·safari* som det kan belegges at brukes aktivt i norsk i dag[2] i betydningen *beundring av natur på snøscooter*, som dermed har etablert seg som en konvensjon brukt på denne typen opplevelsesreiser, og vil følgelig være leksikalisert ut ifra en slik definisjon. Hvis man som kontrastivt eksempel bytter ut *scooter* med en annen generisk betegnelse på et transportmidler som *sykkel* og setter det sammen til *sykkel·safari* er det dette ikke lenger en betegnelse på en konvensjon[3], det er langt mer uklart hva slags safari men også hva slags sykkel det er snakk om.

Leksikalisering som avvik

En annen måte å betrakte det på er å anse leksikalisering som avvik. Dersom leksikonet defineres som et område hvor lingvistiske enheter lagres i motsetning til å genereres, så følger det naturlig at leksikalisering følger av irregularitet. Avvikene må da skje etter orddannelsen, i en utvikling som potensielt leder til leksikalisering. Bauer (1983) kaller disse utviklingstrekkene som leksikaliseringstyper, og viser til syntaktisk, semantisk, morfo-

[2] 3,010 webtreff på Google 30. august 2008 fra norske sider.

[3] 521 treff på Google 30. august 2008 fra norske sider.

logisk og fonologisk leksikalisering. Fonologisk leksikalisering viser til tilfeller hvor fonotaktisk bevegelse i ordet gjør det uklart hvordan ordet opprinnelig var satt sammen. Et eksempel på dette er *Freksta* som er en brukt (om ikke utbredt), betegnelse på byen Fredrikstad, opprinnelig kompositumet *Fredriks·stad*. Bakken (1998) argumenterer i tråd med et slikt syn med at leksikaliseringsprosessen er en essensielt semantisk prosess hvor et opprinnelig sammensatt ord mister sin kompoisjsonelle betydning og blir et eget symbol, slik som *scooter·safari* beskrevet ovenfor.

Leksikalisering som et aspekt ved betydning

I betraktningen av leksikalisering som et aspekt ved betydning, har det vært utbredt å legge uforutsigbarhet til grunn som et kriterium. Dette blir som å snu konvensjonalitetsargumentet over på hodet å si at man ikke kan forutsi hva *scooter·safari* betyr ut ifra de to enkeltordene, siden dette nå er en konvensjon på en spesifikk utflukt. Svanlund (2002) problematiserer dette synet, og hele konseptet komposisjonalitet fra sitt ståsted i kognitiv lingvistikk, og sier at polysemi og metaforisk bruk vil undergrave ethvert forsøk på å forutsi hva et sammensatt ord skal bety ut ifra bestanddelene uansett. Han trekker frem motivasjon, forstått som årsaken til at språkbruker benytter et ord, som den viktigste faktoren i analysen av komplekse uttrykk. En slik analyse kan også forklare engangsforeteelser[4].

3.2 Om komposisjonalitet

Komposisjonalitetsprinsippet (KP) innenfor semantikk og språkfilosofi kan gjengis slik[5] (i forfatterens oversettelse fra engelsk):

> Betydningen av et komplekst uttrykk er en funksjon av betydningen av bestanddelene, og av reglene som har satt dem

[4] engelsk: *nonce words*.

[5] Wikipedia besøkt 11. juli 2008. Url: `http://en.wikipedia.org/wiki/Compositionality`.

sammen.

som innledningen forteller, er komposisjonalitet et tema utenfor diskusjonen om sammensatte ord og lingvistikk, også et tema fra filosofien.

Pelletier (2006) gir en presentasjon av den språkfilosofiske debatten om komposisjonalitet, med utgangspunkt i at teoretikere enten mener at komposisjonalitet faktisk finnes i språk, eller at dette ikke er tilfelle. Pelletier viser til tre forståelser av begrepet, hvor den første er den mest grunnleggende, som kan oppsummeres med setningen «et objekt er summen av sine deler». Innenfor denne diskursen kalles komposisjonalistene atomister, eller også reduksjonister, som mener verden består av objekter som alle kan beskrives av sammensetningene av noen grunnleggende atomiske elementer. Motsatt vil ikke-komposisjonalister kunne kalles holister som mener at kontekstuelle eller emergente egenskaper ved en sammensetning ikke kan fanges opp av en analyse av bestanddelene.

Videre diskuterer et syn på komposisjonalitet som bunner i det som kan kalles naturlig språks *magi*, oppsummerert i tre punkter hentet fra både vestlig og indisk filosofi:

1. Vi kan forstå et uendelig antall setninger, så lenge de bruker ord vi allerede forstår. Vi forstår setninger og kombinasjoner vi aldri har møtt.

2. Vi kan konstruere nye setninger som vi aldri har hørt eller brukt noen gang, og vi vet at de er passende i situasjonen.

3. Vi er skapninger med endelig tidshorisont, som eksponeres for en endelig mengde av informasjon om språket vårt. Likevel kan vi lære et system som kan komme til uttrykk på et uendelig antall måter.

Alle punktene impliserer at språk på den ene side er noe spesielt (uendelig, kreativt), men vi kan likevel bruke og forstå det med våre endelige forutsetninger som mennesker. Komposisjonalitet blir brukt som en forklaring på denne magien. Dette er en funksjonell definisjon av

komposisjonalitet som er mindre eksplisitt enn de andre, og som bare sier at komposisjonalitet er det *noe* som sørger for språks magi, uten å si nøyaktig hvordan.

3.2.1 Semantisk komposisjonalitet

Den siste forståelsen av komposisjonalitet er *semantisk komposisjonalitet*, som beskrevet i KP. Denne veien skiller seg ut fordi at det er oppbygningen til betydningen av noe som er det essensielle, og dermed er det dette utenforliggendes egenskaper som avgjør en eventuell komposisjonalitet, ikke entiteten selv. En semantisk komposisjonalitet forutsetter den første forståelsen av komposisjonalitet er gjeldende. Dersom kompositumet *fiske·stang* skal undersøkes for å se om betydningen kan hentes frem fra betydningen av henholdsvis *fiske* og *stang* forutsetter dette at kompositumets overflaterealisering, *fiskestang* allerede er satt sammen av de to «atomære» bitene.

Problemet med denne forståelsen av komposisjonalitet er at den er vag. Kastes et nytt blikk på KP, så er det her uklart både hvordan betydningen av bestanddelene avgjøres og hvordan reglene for sammensetning ser ut. Når ingenting i prinsippet begrenser hvilke egenskaper som kan tilknyttes ordenes betydning, eller hvilken fleksibilitet som kan inkluderes i sammensetningsreglene, så vil det være mulig å komme frem til komposisjonalitet for alt mulig, bare ved å utvide dem. Et eksempel på dette er rekken av komposita *rød·strupe*, *rød·sprit* og *rød·tang*. Her vil det kunne innvendes at *rød* har forskjellige betydninger i de tre tilfellene, selv om samtlige er satt sammen av de samme syntaktiske reglene, som et argument mot komposisjonalitet. Dette kan imidlertid imøtegås ved å si at betydningen *rød* av rød er den samme hele veien, men betydningen er disjunkt.

Sandu og Salo (2006) foretar en nærmere undersøkelse av semantisk komposisjonalitet på mer formelt grunnlag. De begynner med å sitere KP, og slår fast at det er svært svakt når det står alene, så vagt at enkelte filosofer og logikere anser det for å være metodologisk

innholdsløst, fordi det kan gis en slik tolkning. Dersom det imidlertid tolkes på en måte som forutsetter en sammenheng, formelt gjennom kontekstprinsippet, i betydningen «et utsagn betyr noe bare i den kontekst det forekommer» blir det mer interessant. Forfatterne viser til (Hodges, 1998) hvis utvidelsesteorem gjør det mulig å kombinere KP og kontekstprinsippet til å gi tolkninger av komplekse uttrykk som setninger, hvor betydningen av bestanddelene eller de syntaktiske operasjonene som virker på dem ikke begrenses av eksempelvis humør eller værforhold. Slike begrensninger forekommer imidlertid ofte i naturlige språk, og avslutningsvis siterer de Fodor og Lepore (2002) som slutter at de fleste teorier fra kognitive vitenskaper (engelsk: *cognitive science*) er uforenlige med KP, hvilket parallelt med diskusjonen om leksikalitet i forrige avsnitt, viser til innvendinger fra kognitiv lingvistikk.

3.3 Allment om komposita

Bauer (2006) begynner sin diskusjon av *komposita* med en enkel, og svakt selvmotsigende definisjon som går ut på at et kompositum er et ord utgjort av to andre ord. Denne definisjonen er imidlertid så romslig at den må presiseres både for alle språk, og også enkeltvis. Eksempelvis kan et kompositum bestå av flere enn to ord, men aldri færre, og «ordene» kan opptre i ulike varianter, som siteringsformer, ordstamme eller en egen sammensetningsformativ som i norsk, som i *sekke·løp* på norsk, og det er derfor en sammensetning av *leksemer* som menes. Bauer tegner et bilde av komposita innen lingvistikk som et potensielt forvirrende landskap, uten noen tilnærmet konsensus på det mest fundamentale, hvilke konstruksjoner som kvalifiserer til statusen *kompositum*

På tross av enkelte forskjeller, identifiserer Bauer to innfallsvinkler til komposita som i hovedsak brukes til å beskrive deres oppbygning. Enten anses et kompositum å være en spesifikk språklig konstruksjon med en formell definisjon. Alternativt vurderes det som en leksikal enhet som møter visse kriteria, jamfør diskusjon om leksikalisering i avsnitt 3.1.

For å illustrere forskjellen viser hun til eksemplene *black·bird*, *wind·mill* og *combination lock*, hvis norske motsvarigheter *svart·trost*, *vind·mølle* og *kombinasjons·lås* fungerer like godt, ved å slå fast at samtlige av disse vil få kompositumsstatus sett fra begge ståsteder. Ser man imidlertid på eksempel 2,

(2) (PM backs) mercy killings bill
mercy·killings·bill
barmhjertighets·draps·proposisjon
(PM) støtter barmhjertighetsdrapsproposisjon

så vil dette være et kompositum ut ifra et konstruksjonsperspektiv, men ikke et leksikalt. Det er helt klart flere leksemer satt sammen til en enhet, men kan vanskelig sies å være leksikalisert. De som forfekter et syn på at disse konstruksjonene er komposita legger vekt på at det ikke er noen formell forskjell mellom «nyhetsordene» og leksikaliserte komposita hvis leksikalisering avgjøres av frekvens. Selv om det er en vesensforskjell mellom de to måtene å se saken an på, vil det også være en del overlapp mellom definisjonene, som de første eksemplene forteller om. Dette gjør også at det ikke alltid er klart hvilken av dem som menes, når begrepet anvendes lingvistikken.

Bauer forteller at hun likevel har funnet frem til visse kriterier, som det er allmenn aksept for at ihvertfall til en viss grad samsvarer med kompositumsstatus[6]. Disse kriteriene er de samme som Bauer (1983) kalte leksikaliseringstyper, nemlig ortografiske, fonologiske, morfologiske, syntaktiske og semantiske trekk, som diskutert i avsnitt 3.1 med unntak av ortografi, som bare opptrer her. Det virker også rimelig at hvis en konstruksjon får kompositumsstatus hvis det er leksikalisert, så må kravene til leksikalisering av konstruksjonen være oppfylt. Det siste følger av det første, selv om det ikke er nødvendig at kompositum er leksikalisert.

Ortografi er et kriterium som er meget tydelig på norsk, da svært mange komposita opptrer med en spesiell sammensetningsformativ, som

[6]Engelsk *original: ...there are a number of criteria that are generally accepted as correlating with compound status, at least to a certain degree.*

vil bli diskutert mer inngående i avsnitt 3.4. Også verdt å merke seg fra diskusjonen om ortografi hvor hun bruker engelsk som eksempel er at dette er et meget uklart bilde. Hun trekker frem stavemåtene *rainforest*, *rain-forest* og *rain forest* og viser til at det er enkelt å finne belegg for bruken av dem alle i samme betydning. Bauer viser også til den kontroversielle praksisen i skandinaviske språk hvor sammensetninger i økende grad skrives med mellomrom, omtalt som «orddeling», som for eksempel *ananas ringer* i substantivisk forstand. En mulig tolkning av de tre ulike skrivemåtene for *rain·forest* viser selve leksikaliseringsprosessen fra orddannelse til leksem, men Bauer avviser dette som en pålitelig forklaring etter undersøkelser av ordbøker med forskjellig alder.

3.3.1 Inndeling av komposita

Klassifikasjon av komposita må være vanskelig når det er stor uenighet om hva som i det hele tatt er et kompositum, men Bauer (2006) legger vekt på de første kjente inndelingen fra sanskrit-grammatikerne i India som er fortsatt basis for mange moderne inndelinger, og selve merkelappene er dermed relevante. De fire grunnleggende typene komposita fra Sanskrit var:

- *Tatpurusa*, hvor man har en klar hierarkisk struktur, hvor et ledd er et hypernym til det som betegnes av kompositumet, og det andre leddet representerer modifikasjonselementet som utgjør selve undergrupperingen. Et eksempel på dette er *blikk·boks*, som er en undergruppe av *boks* nærmere spesifisert ved at den er laget av blikk[7]. Denne strukturen omtales ofte på engelsk som *modifier-head*, hvor *head* betegner hovedledd-egenskapen som her innehas av ordet *boks*.

- *Dvandva*, som betegner koordinerte ord, som kunne vært bundet sammen av «og». Eksempler på dette er stedsnavnene *Magny-Cours*

[7]Blikk er flertydig (som i å kaste et blikk), men dette får ikke konsekvenser for argumentet, og ses vekk ifra.

og *Garmisch-Partenkirchen*.

- *Bahuviri*, som betegner komposita av adjektivisk natur som beskriver en implisitt entitet som er eier av egenskaper som kommer frem gjennom kompositumets delledd. Et eksempel på dette er *krøll·topp*, som ikke er en type *topp*, men betegner en person hvis hode er preget av krøller.

- *Avyayubhava*, som betegner adjektiviske komposita brukt som adverb. Denne typen har hatt mindre betydning i moderne lingvistikk.

3.3.2 Universalitet

Bauer (1983) fastslår at det er umulig å si om komposita finnes i alle språk, på grunn av de sprikende definisjonene. Begge deler har vært argumentert for. Det kan godt være at komposita som konstruksjonstype er en universell egenskap, men ikke som leksikal enhet, heter det. Med tanke på diskusjonen om leksikalisering over, og hva dette impliserer av antagelser om hvordan selve «leksikonet» ser ut, vil det motsatte være minst like mulig. I denne diskusjonen kommer det frem at preposisjonsuttrykk på fransk kan regnes som komposita, som for eksempel *chemin de fer* («vei av jern» = «jernbane»), som er verdt å merke seg fordi mange norske komposita finner sine motstykker i engelske preposisjonsuttrykk i eksperimentene kapittel 6, uten at det er opplagt at disse oversettelsene også er komposita.

3.4 Komposita på norsk

Endresen, Simonsen og Sveen (2000) viser i sin innføringsbok i lingvistikk til en tradisjonell og enkel definisjon av en sammensetning,

> En sammensetning er et ord som består av flere ord.

Ord av flere forskjellige ordklasser kan være sammensetninger, og ord fra varierende ordklasser kan settes sammen til nye ord igjen. Begrepet

sammensetning i denne sammenhengen forstås i det videre som det samme som kompositum.

Forfatterne viser til disse eksemplene:

Ordklasse	Eksempel
substantiv	*måne·skinn*
adjektiv	*ild·rød*
verb	*kjede·røyke*
adverb	*bak·over*
preposisjoner	*inn·under*

Disse eksempelordene er delt i to, og de betegner det første leddet som *forledd* og det andre som *etterledd*. Men komposita kan nøstes, slik at forledd eller etterledd selv kan være et kompositum, slik som det flertydige eksempelet *overbygningskontor*, som kan analyseres som enten *over·bygningskontor* eller *overbygnings·kontor*. På norsk blir av og til ordet *bilbil* brukt for å betegne en bil som er konstruert for å frakte andre biler. Dermed kan man i prinsippet tenke seg en uendelig rekke av ordet *bil* som vil betegne noe som frakter en stadig større farkost ettersom den blir lengre.

Faarlund, Lie og Vannebo (1995) behandler sammensetninger mer utførlig i Norsk referansegrammatikk, og støtter denne definisjonen, som også betrakter komposita som en konstruksjonstype, uten å ta hensyn til leksikaliseringskriterier eller -prosessen. De nyanserer imidlertid bildet med å skille mellom sammensetninger og avledninger. I følge Faarlund et al. går skillet ut på at ord hvis forledd ikke kan stå alene avleder avledninger, hvor greske prefikser som *bio*, *maksi* og *pol* gir opphav til mange eksempler.

Faarlund et al. diskuterer også leksikalisering og et begrep de kaller *gjennomskuelighet*, som begge oppfattes som gradvise egenskaper. Gjennomskuelighet er det samme som komposisjonalitet, definert med forfatternes ord som *Når betydningen er gjennomskuelig, kan vi slutte oss fram til betydningen fra betydningen til de delene ordet er satt sammen av*. Med andre ord er betydningen av sammensetningen en funksjon av betydningen

av bestanddelene slik komposisjonalitetsprinsippet lyder. Leksikalisering definerer de med ordene *Når betydningen (...) ikke lenger uten videre kan leses ut av de delene det er sammensatt av, sier vi at sammensetningen er leksikalisert*, Jamfør diskusjonen av leksikalisering i avsnitt 3.1 så oppfatter Norsk referansegrammatikk leksikalisering som et aspekt ved betydningen. De vil dermed oppfatte *løve·tann* som mer leksikalisert enn *politi·betjent* fordi det er nokså greit å slutte seg til betydningen av det siste kompositumet fra dets bestanddeler men tilnærmet umulig i det første.

3.4.1 Bøyning av komposita

Endresen et al. (2000) slår direkte fast at norske komposita bøyes etter etterleddet, noe som betraktes likt i Norsk referansegrammatikk, hvor det formuleres ved at komposita får sin ordklasse etter etterleddet, og når det har en gitt ordklasse så må det bøyes i tråd med denne, hvilket gjør førstnevnte syn til en implikasjon av det neste. I Norsk referansegrammatikk innføres begrepet *kjerne*, som benyttes på kompositumets viktigste ledd, som da ofte er etterleddet. Begrunnelsen er dels syntaktisk, at kompositumet får sin ordklasse fra dette leddet, og dels semantisk fordi kompositumet henter størstedelen av sin betydning herfra. Alternativet til at etterleddet er kjerne er at det ikke er noen, som vil bli eksemplifisert i avsnitt 3.4.3. Norsk referansegrammatikk nyanserer imidlertid ved å vise til en serie unntak hvis ordklasse ikke er i samsvar med etterledd, som substantivene *kryp·inn* og *far·vel* hvis etterled ikke er av samme klasse.

At kjernen står til lengst mot høyre er en indikator på at norsk er et venstreforgrenende språk, forstått som at del(er) av en frase som er avhengig av en kjerne står til venstre for denne. Dette blir tydelig ved bøyning av nomenkomposita i tall og kasus, som vist for et eksempel fra hvert genus under:

fotball	**teskjekjerring**	**bokomslag**
fot·ball	*teskje·kjerring*	*bok·omslag*
fot·ballen	*teskje·kjerringa*	*bok·omslaget*
fot·baller	*teskje·kjerringer*	*bok·omslag*
fot·ballene	*teskje·kjerringene*	*bok·omslagene*

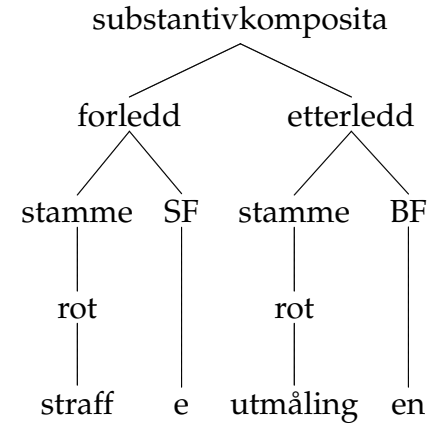
3.4.2 Sammensetningsformer

Forleddet i eksemplene vist til nå har opptrått konsekvent i entall, og i en form det kan stå alene, eller som etterledd. Det er imidlertid ikke alltid slik, som følgende eksempelliste fra Endresen et al. (2000) viser.

> *barne·hage*, *gutte·kor*, *preste·gård*, *kalve·dans*. *heste·handel*, *lønns·nemd*, *manns·kor*, *dåps·kjole*, *dags·marsj*, *tids·nød*, *barndoms·minne*

Samtlige forledd i denne listen står i en bøyd form som skiller seg fra ordene når de står i entall og nominativ. Formene kan minne genitiv som i *dåps·kjole* og flertall som i *blomster·kjole*, men Endresen et. al kaller dem *sammensetningsformer* og fugeelementene *sammensetningsformativer*. Dette går frem av eksempel (3), hvor SF viser sammensetningsformativen, og BF tydeliggjør bøyningsformativen som viser at ordet står i bestemt form. Roten av et ord betegner en minimal morfologisk enhet som ikke kan deles videre, men samtidig har en leksikal betydning.

(3)



3.4.3 Semantisk inndeling

Norsk referansegrammatikk identifiserer fire hovedtyper sammensetninger, nokså parallell med diskusjonen i avsnitt 3.3.1. Bauer (2006) finner altså her støtte for sin påstand om at moderne grammatikere i høy grad bruker Sanskrit-inndelingen som mønster. Faarlund et al. (1995) bruker Sanskrit-kategoriene direkte som benevnelser på to av dem. Bruken av betegnelsene samsvarer godt med den som er beskrevet ovenfor, men noen eksempler tas likevel fra Norsk referansegrammatikk, fordi det ikke nødvendigvis er helt den samme forståelsen av begrepene som er lagt til grunn. De fire kategoriene fra Norsk referansegrammatikk er:

- Determinative sammensetninger. Dette er det vanligste slaget, hvor forleddet avgjør den nærmere spesifikasjonen av etterleddet (dermed determinativt), hvor etterleddet opptrer som kjernen i frasen. Eksempler på dette er *havre·grøt* og *risengryns·grøt* som begge er undergrupper av *grøt*.

- Possesive sammensetninger. Dette er komposita hvor noe utenfor selve ordet er eier av egenskapen utgjort av dette possesive kompositumet. Eksempler på denne typen er typiske dyrenavn som *rød·strupe* og *bleik·nebb*, men også mange nedsettende betegnelser på medmennesker som *bleik·fis* og *lav·panne*. Poenget er da at brukeren av disse ordene søker å henvise til noen som har egenskapen som kommer frem i det, derav possesivt. Norsk referansegrammatikk kaller denne typen også bahuviri-sammensetninger.

- Imperativsammensetning. Dette er en mindre gruppe hvorav enkelte står i unntakslisten over komposita hvis ordklasse ikke samsvarer med etterleddets. De består av imperativformen av et verb som forledd, og en utfylling. Eksempler er *kryp·inn*, *forglemmeg·ei* og stedsnavnet *Kik·ut*.

- Kopulativ sammensetning. Dette disse kalles også dvandva-sammensetninger, og består av to likeverdige ledd hvis betydning kan betraktes som summen av leddene. Etterleddet anses ikke som kjernen

Forledd/etterledd	Substantiv	Verb	Adjektiv
Substantiv	x	x	x
Verb	x	x	x
Adjektiv	x	x	x
Preposisjon	x	x	x
Adverb	x		x

Tabell 3.1: *Mulige ordklasser som kan stå som forledd i norske komposita. Mulige forledd i rader.*

i slike komposita. Eksempler på dem er sammenføyninger av stedsnavn og adjektiv som *Aurskog-Høland* og *blå·gul*.

3.4.4 Klassifikasjon etter ordklasse

Med unntak av de unntakene som ble nevnt under diskusjonen av imperativkomposita over, så har de resterende komposita ordklasse etter sine etterledd. Norsk referansegrammatikk foretar en gjennomgang av disse ordklassene, med en diskusjon av hvilke ordklasser som er mulige forledd. Ordklassene er de nesten de samme som hos (Endresen et al., 2000), diskutert avsnitt 3.4 med unntak av klassen subjunksjoner som legges til. I Norsk referansegrammatikk gis sammensatte adverb, preposisjoner og subjunksjoner en mer perifer rolle, sannsynligvis fordi de skiller seg fra de andre ordklassene ved at de er sterkt innarbeidet i norsk språk, og diskusjonen av dem blir mer en diskusjon av deres opprinnelse. Således er disse kompositaene mer i tråd med en leksikal definisjon av komposita enn en konstruksjonstypedefinisjon. En tabellerisk oppsummering av diskusjonen for de tre viktigste ordklassene, substantiv, adjektiv og verb, viser at spennvidden er stor i hvilke ordklasser som kan sammenkoples til nye, norske komposita i tabell 3.4.4.

Som det går frem av denne tabellen, så er komposita som består av substantiv både som forledd og etterledd en av svært mange typer, selv om det er den oftest forekommende.

Mer om semantikk

Endresen et al. (2000) har også en liten diskusjon av kompositas semantikk, og slår fast at man ikke alltid kan komme frem til hele sammensetningens betydning ved å studere dens bestanddeler. De slår fast at *blå·bær* betegner en spesifikk planteart, uten at det er mulig å forklare hvorfor ikke alle blå bær er blåbær, ut ifra forledd og etterledd. Videre viser de til *jord·bær*, *brann·mann*, *bil·selger*, *dør·selger*, *barne·film* og *dyre·film*, og viser til at disse sammensetningene kunne bety flere ting ut ifra sine respektive ledd. Et jordbær kan være et bær av jord, en brannmann kunne vært en mann som stifter branner, en bilselger kunne vært en ambulerende sådan, og en dørselger kunne solgt dører, en barnefilm kunne vært en film av barn og ikke for barn, og tilsvarende kunne en *dyrefilm* vært en film for dyr, og ikke om dem.

Hvis man ser motsatt på det, så ville tilfellet hvor vi kunne finne frem til betydningen av et kompositum ved å studere bestanddelene være i tråd med komposisjonalitetsprinsippet omtalt i avsnitt 3.2. Som det kom frem av den diskusjonen er det tvil om det overhodet er mulig å finne tilbake til den hele betydningen av en kompleks frase, herunder et kompositum. Hvis man aksepterer den tvilen er det ikke mulig å finne tilbake til hele betydningen, og dermed er ikke dette noen interessant målestokk. Likevel går det an å si noe mer om kompositas semantiske egenskaper ved å dele dem inn i kategorier som gjort over, som videre kan underspesifiseres igjen. Et eksempel på dette er determinative nomenkomposita som kan deles inn etter hvilken preposisjon som kunne vært brukt til å forklare hva det betyr. Eksempelvis er en *tømmer·hytte* laget *av* tømmer, mens en *fjell·hytte* en hytte som er laget *for* bruk på fjellet. Det henvises til Norsk referansegrammatikk for en inngående undergruppering langs disse linjer.

3.4.5 Begrepsbruk i denne oppgaven

Som diskusjonen ovenfor har vist, så er diskusjonen av hvor grensene går for hva som er et kompositum, og når det eventuelt slutter å være

det en stor debatt som har pågått lenge. Eksperimentene foretatt, og omtalt i denne oppgaven har imidlertid et mer pragmatisk utgangspunkt, å oversette komposita, hvor det er hensiktsmessig med en funksjonell definisjon. Et kompositum vil bli omtalt som et kompositum dersom forfatterne betrakter dem som dette, under omtale av egne eksperimenter. Når det kommer til eksperimentene utført på oversettelse fra norsk til engelsk, så betraktes samtlige ord som er valgt ut til å bli oversatt som komposita, men det tas ikke stilling til hvorvidt konstruksjonene de blir oversatt til på engelsk, bør ha kompositumsstatus.

Innenfor rammene av et eksperiment som går ut på å oversette komposita ved å oversette bestanddeler, betraktes komposisjonalitet som den praktiske muligheten til å nå det ønskede målet for oversettelsen (den såkalte gullstandardoversettelsen) nettopp ved oversettelse av bestanddelene. Dersom denne er et består av bare ett leksem, vil dette være umulig.

3.5 Oversettelse av komposita

3.5.1 Utbredelse av nominalkonstruksjoner

Norsk og engelsk er nært beslektede språk, og det er dermed nærliggende å anta at forskjellene er små, men det er samtidig sannsynlig at det er en viss forskjell i utbredelsen av nominalkonstruksjoner generelt, og substantivkomposita spesielt. Og selv om utbredelsen skulle være omtrentlig, eller i prinsippet helt lik, er det ikke opplagt at alle uttrykk som uttrykkes nominelt på norsk også realiseres slik på engelsk.

Følgende eksempler viser slike situasjoner. Først et eksempel fra et oppslag i KF som viser et norsk kompositum som oversettes med en engelsk setning (4), og motsatt et eksempel fra Oslo Multilingual Corpus som viser et engelsk uttrykk, *stiffness* som oversettes med en norsk setning (5).

(4) blomsternektar

nectar secreted by flowers

(5) ... , but there was something about the stiffness of his gait ...
..., men det var noe med den stive måten han gikk på ...

Dette er to eksempler på nominalkonstruksjoner som oversettes med andre konstruksjoner i oversettelse mellom språkene, som viser at det ikke nødvendigvis er noe en-til-en forhold her. Selv om det i disse tilfellene godt kan finnes nominalkonstruksjoner på motsatt språk som vil klare å formidle det samme budskapet, så kan man ikke uten videre anta at en nominalfrase på engelsk vil være den mest naturlige måten å uttrykke noe, som på norsk uttrykkes med en slik. I denne oppgaven gjøres det imidlertid ingen forsøk på å beskrive slike transisjoner.

3.5.2 Automatisk analyse

Johannessen og Hauglin (1996) presenterer en algoritme for automatisk analyse av norske komposita, i forbindelse med utviklingen av en morfosyntaktisk tagger. De innleder med å si at dannelsen av komposita er ekstremt produktivt i norsk språk[8], for å motivere behovet for en sammensetningsanalysator. Som på tysk, så skrives norske komposita i ett ord, og deres interne struktur må undersøkes for å analysere dem. Oppgaven til en tagger er å gi rett ordklasse til hele sammensetningen, og selv om norske sammensetninger får ordklasse etter sine etterledd[9] er det nødvendig å vite hvordan ordene skal deles opp for å få tak i hva som nettopp er etterledd i konstruksjonen for en leksikon-basert tagger som Oslo-Bergen-taggeren. En sammensetningsanalysator ble defor utviklet som en modul til bruk i denne taggeren.

Sammensetningsmodulen fungerer ved å først finne frem til mulige analyser av et ord den blir forelagt i henhold til inndelingen under og et leksikon, og dernest å finne frem til den av analysene som er den

[8]Originaltekst: *compounding is extremely productive in Norwegian*.

[9]Med unntak som diskutert i dette kapittel.

grammatikalsk og semantisk riktige. Den arbeider direkte på ordet, uten kontekstuell informasjon.

Inndeling

Johannessen og Hauglin (1996) viser til fire typer komposita i norsk etter morfologi:

- **jukstaaposisjonerte** stammer, når de bare settes rett ved siden av hverandre, slik at *mor* + *far* blir til *mor·far*, eller *telefon* + *svarer* blir til *telefon·svarer*
- **suppletiv** stamme av typen *kles* eller *billed* som kan danne *kles·skap* eller *billed·språk*
- **epentisk** *s* når to ord fuges sammen med en *s* for å danne et kompositum som *mors·binding* eller *aluminiums·fabrikk*
- **epentisk** *e* når to ord fuges sammen av en *e* for å danne et kompositum som *barne·trygd* eller *heste·ekvipasje*

Dette dermed plasserer Johannessen og Hauglin seg i en tradisjon som definerer komposita ut ifra ortografiske og morfologiske kriterier jamfør Bauers kriterier for kompositumsstatus i avsnitt 3.3. Relatert til diskusjonen om norske komposita i avsnitt 3.4.2 så oppter de «epentiske» bokstavene -e- og -s- som sammensetningsformativer som fordi ordene føyes sammen til et kompositum.

Flertydighet i oppdeling

Akø (1989) forteller at utbredelsen av de forskjellige kompositatypene er at 75% tilhører den jukstaposisjonerte kategorien, 17% opptrer med epentisk -s-, og de resterende 8% med epentisk -e-. Fordi at det i norsk språk er en rekke ord som ender på -e- og -s-, er det nødvendig å utvikle regler for å skille mellom -e- og -s- som sammensetningsformativer og som deler av et leksem. Et eksempel på dette er *løvetemmer* som fra et morfologisk ståsted

like gjerne kan analyseres som *løv-e·temmer* i betydningen en temmer av løv, som *løve·temmer,* en temmer av løver.

Videre er det mulig at både forledd og etterledd selv er komposita, som øker antall mulige oppdelinger. For ordet *kulturforskeren*. er de titalls måter å sette det sammen på som deler av andre ord, hvorav et utvalg er vist i eksempel 6.

(6) a. *kultur·forskeren*

b. *kul·tur·forskeren*

c. *kultur·forske·ren*

d. *kultur·forsker·en*

e. *kul·tur·forsker·en*
etc.

Å velge rett analyse

Dermed blir en viktig utfordring å velge blant disse mulige analysene. Johannessen og Hauglin (1996) benytter seg av grammatisk kunnskap som ordklasse, bøyning, type ordstamme (en- eller flerstavelser) samt fonologiske og fonotaktiske hensyn gjennom ulike heuristikker. Et eksempel på en slik regel er 7.

(7) Hvis to analyser har det samme antall ledd og ingen epentisk bokstav er involvert, velg analysen som er et substantiv, hvis tilgjengelig.

Denne regelen gjør at riktig analyse blir valgt i eksempel 8

(8) a. *blomster·holder* (substantiv)

b. * *blomst·erholder* (verb i presens)

Denne heuristikken er klassifisert som en ordklasse-regel fordi den velger ut en ordklasse.

Modulen og ordklasser

I eksempel 6 kommer det også frem tydelig at denne analysatoren av komposita også har rom for andre ordklasser enn substantiv, som gir enda flere tolkninger av det samme ordet fordi flere leksemer fra leksikonet kan brukes i søk etter oppbygninger. Johannessen og Hauglin (1996) diskuterer imidlertid ikke direkte hvilke ordklasser som er mulige for- og etterledd i norske komposita, men gir eksempler som viser såvel substantiv som verb og adjektiv som etterledd, og dermed ordklassen til selve kompositumet. Som det går frem av tabell 3.4.4 så kan norske komposita være satt sammen av mange ordklasser, med varierende hyppighet og produktivitet. En spesifikasjon av hvilke kombinasjoner av ordklasser som modulen anser som produktive, og en rangering av deres produktivitet synes å være relevant informasjon som kunne brukes til å luke vekk antatt gale analyser.

Implementasjon og resultater

Algoritmen omtalt av Johannessen og Hauglin (1996) er implementert som en del av Oslo-Bergen-taggeren, men også gjort tilgjengelig som en separat applikasjon av Paul Meurer ved Universitetet i Bergen. Evaluering omtalt i (Johannessen & Hauglin, 1996) viste at analysatoren bare gir en helt gal analyse i 1.1% av tilfellene, definert som at den foreslår en oppdeling med galt etterledd. Dette riktignok etter at komposita hvis bestanddeler ikke stod i leksikonet ble luket ut, og dermed ikke talt med som feil.

Videre siterer de Munthe (1972) som forteller at 10,4% av ord i løpende tekst er komposita, og støtter dette med en egen stikkprøve fra avisen Aftenposten. Forekomsten av komposita er imidlertid veldig mye lavere i resultatene fra Oslo-Bergen-taggeren (OBT), nærmere beskrevet i avsnitt 4.8, som skal gjøre bruk av en sammensetningsanalysator som er basert på metodene brukt i denne oppgaven. I arbeidet med denne oppgaven ble det norske korpuset analysert med OBT og andelen ord merket med merkelappen «SAMSET», som er OBT sin måte å identifisere en sammensetning, var bare på 1,5%.

Morfologisk preprosessering

Videre viste en nærmere undersøkelse at ordene merket med SAMSET fordelte seg bare mellom to ordklasser, substantiv og adjektiv, hvor av 37058 (83%) var substantiv og 7569 (17%) var adjektiv fra det samlede norske tekstkorpuset omtalt i avsnitt 4.2.1. Men når forfatterne presenterer flere eksempler med verbkomposita virker det ikke som de mener at denne undergruppen av komposita ikke er produktiv. Likevel viser det seg at det er meget vanskelig for sammensetningsanalysatoren slik den behandler etterleddets morfologi å komme frem til at komposita er verb. Som vist av eksempel 6, *kulturforskeren*, så klarer modulen å identifisere at *forskeren* er et maskulint substantiv i bestemt form entall. Den kan også komme frem til at et kompositum som *liv·redde* består av leksemene *liv* (substantiv) og *redd* (adjektiv), eller *re* (perfektum partisipp av å *re* brukt som adjektiv). I disse to tilfellene blir morfologien analysert i sammensetningsanalysatoren, mens substantivkomposita som står i genitiv som *kultur·forskeren-s*, der blir kompositumet oversendt modulen både med og uten genitivsmerket. Resultatet av å sende et substantivkompositum i genetiv i fullform til analysatoren er at det returneres som et adjektiv med etterledd *ens* (i betydningen å være enig) eller som substantiv med etterledd *nes*.

Når det kommer til verbmorfologi, skjer ingen lemmatisering verken i taggeren i forkant. Ved å prøve tre eksempler på verbkomposita fra Norsk referansegrammatikk, *kniv·drepe*, *små·flire* og *lys·regulere* i sammensetningsanalysatoren så blir de analysert som substantiv når de står i presens (endelsen «er» kan også være en flertallsmarkør) eller adjektiv når de står i preteritum (endelse «te» kan også være en flertallsbøyd perfektum partisipp brukt adjektivisk). Når de samme ordene blir brukt rett i Oslo-Bergen-taggeren, så skjer heller ingen lemmatisering i forkant slik som ved substantiv i genitiv, og ordene blir klassifisert som adjektiv eller substantiv etter tid av verbet.

Behandlingen av setningen *Oslo lysregulerer kryssene fra neste år.* brukt i taggeren er gjengitt under:

```
"<Oslo>"
        "Oslo" subst prop @løs-np &sted <sted>
"<lysregulerer>"
        "lysreguler" subst appell mask ub fl samset @løs-np
"<kryssene>"
        "kryss" subst appell nøyt be fl @løs-np
"<fra>"
        "fra" prep @adv
"<neste>"
        "neste" det dem be <adj> @det>
"<år>"
        "år" subst appell nøyt ub ent fl @<p-utfyll
        "år" subst appell mask fem ub ent @<p-utfyll
"<.>"
        "$." clb <<< <punkt>
```

På tross av at verbkomposita regnes som produktive både av Norsk referansegrammatikk og Johannessen og Hauglin er det i praksis umulig å få ut en analyse av systemet. Regelen gjengitt i eksempel 7 sørger for at *lys·regulere-r* blir betraktet som substantiv og ikke verb.

Substantiv i genitiv blir heller ikke analysert som substantiv av selve sammensetningsanalysatoren. Men i OBT-implementasjonen er det motvirket ved at fjerne gentivs -s- før den sendes videre til dit. Fordi OBT baserer seg på en «Constraint Grammar»-formalisme (Karlsson, 1990) kan man derfor sette spørsmålstegn ved om sammensetningsanalysatoren er rett sted for å avgjøre hvilken ordklasse ordet skal få. I V2-språket norsk (i betydningen at verbet er annet setningsledd) synes derfor analysen i eksempelet ovenfor som lite plausibel. Dersom i stedet sammensetningsanalysatoren sendte videre både analysen som substantiv og verb, er det stor sjanse for at «Constraint Grammar»-reglene[10] ville lande på en verbal analyse av *lys·regulere-r*.

[10] *Constraint* kan i denne sammenhengen forstås som en restriksjon.

Analyse av forskjell i antall komposita

En nærmere undersøkelse av forskjellen i antall komposita mellom empiriske undersøkelser og Oslo-Bergen-taggerens resultater syntes interessant på grunn av den vesentlige forskjellen mellom de 10,4% sitert i innledningen, og de 1,5% identifisert av OBT. Som Johannessen og Hauglin (1996) ble en kort avisartikkel fra Aftenposten brukt [11], *Tatt i ed i Zimbabwe* gjengitt under:

> Presidenten avla sin ed under en seremoni i presidentboligen i Zimbabwes hovedstad Harare søndag ettermiddag.
>
> Valgkommisjonen offentliggjorde søndag resultatet fra det kritiserte presidentvalget, der Mugabe var eneste kandidat. De offisielle tallene viser at Mugabe fikk 2.150.269 stemmer mot 233.000 til opposisjonsleder Morgan Tsvangirai, som boikottet valget, men hvis navn fortsatt sto på stemmesedlene.
>
> Jeg erklærer derfor Mugabe, Robert Gabriel, til behørig valgt president i Zimbabwe, sa valgkommisjonens leder Lovemore Sekeramyi søndag.
>
> 42,37 prosent av de stemmeberettigede skal ha avlagt stemme ved valget, omtrent like mange som ved første valgomgang i mars, ifølge kommisjonen.
>
> Mugabe sa tidligere søndag i en TV-overført tale at han kom til å vinne en «overveldende seier».

I denne teksten som består av 110 ord, kommer følgende 12 komposita til syne: *president·boligen*, *etter·middag*, *Valg·kommisjonen*, *offentlig·gjorde*, *president·valget*, *opposisjons·leder*, *stemme·sedlene*, *valg·kommisjonens*, *stemme·berettigede*, *valg·omgang*, *TV-·overført* og *over·veldende*. Dette utgjør 10,9% av antall ord, og støtter således både Munthe (1972) og Johannessen og Hauglin (1996) sin stikkprøve.

[11] http://www.aftenposten.no/nyheter/uriks/article2511779.ece besøkt 30. juni 2008.

Men OBT finner bare 3 komposita, nemlig *valg·kommisjonen*, *valg·kommisjonens* og *TV-·overført*, eller 2,7% av ordene. Dette er mer i samsvar med funnene fra taggingen av det norske korpuset med OBT som nevnt ovenfor. Dersom ordene i listen ovenfor blir forsøkt prossesert med sammensetningsanalysatoren direkte, så finner dem imidlertid en korrekt analyse for alle 12, bortsett fra 4, nemlig *etter·middag*, *over·veldende*, *valg·kommisjonens* og *offentlig·gjorde*, og av disse er de to første oppført i Bokmålsordboka. De resterende 8 sammensetningene kom seg aldri til sammensetningsanalysatoren, av to grunner. Den ene grunnen er at de stod i ordlisten, som er tinærmet lik Bokmålsordboka (Wangensteen, 2005), som eksempelvis *stemme·sedlene*, fordi at ord i denne listen ikke betraktes som komposita i OBT, og blir sendt sammensetningsanalysatoren for videre analyse,

Alternativt ble en Kleene-stjerne ble brukt til å identifisere ordet. Merkelappen som settes på ordet *president·bolig*,

```
(SUBST APPELL MASK BE ENT <* BOLIG>)
```

viser dette. Kleene-stjernen betyr at alle sekvenser av bokstaver og tegn som direkte etterfølges av *bolig* dekkes av denne regelen, og får attributter som vist. Riktignok kan et ord som ender på *bolig* bli behandlet av en annen regel først slik at den aldri kommer til anvendelse, men dersom den blir brukt, er dette resultatet.

Undersøkelsen av denne avisartikkelen er for liten til å si noe entydig om hvor godt algoritmen virker, men gir en tydelig indikasjon på hvorfor OBT bare er i stand til å identifisere og analysere en brøkdel av de komposita som opptrer i løpende tekst med norsk språk. Selv om sammensetningsanalysatoren kan kjenne igjen de fleste på korrekt måte blir de ikke merket slik av OBT, så blir den i de fleste tilfellene ikke kalt opp. Dette i tillegg til at verbkomposita ikke kan identifiseres, forteller at OBT ikke kan brukes til å vurdere hvor stor andel av løpende tekst som er komposita.

3.5.3 Annet om sammensetningsanalysatoren

Dersom komposita bindes sammen av en bindestrek, så vil sammensetningsanalysatoren alltid si at forleddet er *ukjent*. Dermed er det umulig å bruke den til å analysere slike komposita for å gi et bilde av indre egenskaper, selv om de kan identifiseres av OBT. For eksempelet nevnt over, *TV-·overført*, så klarer OBT og sammensetningsanalysatoren å identifisere det med SAMSET-merket, men forleddet rapporteres som ukjent.

3.5.4 Perspektiver fra Baldwin og Tanaka

Baldwin og Tanaka (2004) innleder sin seneste artikkel "Translation by Machine of Complex Nominals: Getting it Right" med en diskusjon av hvorfor "Complex Nominals", i denne sammenhengen substantivkomposita representerer en utfordring for maskinell oversettelse. Deres arbeider representerer grunnlaget for arbeidene i denne oppgaven, og ble diskutert mer utførlig i 2.4.3, men deres diskusjon av den lingvistiske motivasjonen for arbeidet hører hjemme her. Selv om deres arbeider dreide seg om oversettelse mellom japansk og engelsk og tilbake, vil også denne diskusjonen være relevant i forhold til denne oppgaven.

De gir fem grunner til at substantivkomposita er et problem for maskinoversettelsessystemer (MT-systemer):

- **varierende konstruksjoner i oversettelser**: Oversettelsen av et kompositum kan realiseres som en forskjellig konstruksjon på engelsk. Et eksempel på dette er *anbefalings·brev* som på engelsk kan oversettes med *letter of recommendation*, som antar formen substantiv-preposisjon-substantiv (N of N), mens *bok·magasin* oversettes med *literary journal* som antar formen adjektiv-substantiv (ADJ-N) og *freds·lengsel* oversettes med *yearning for peace*[12], som antar formen substantiv-preposisjon-substantiv (N for N).

- **leksikalsk divergens**: dette vil si at samme leksem på norsk som etterledd kan bli oversatt med flere typer engelske etterledd. Et

[12]Oversettelse fra gullstandarden brukt i eksperimentene omtalt i kapittel 6.

eksempel på dette er *kirke·gang* som oversettes med *churchgoing*, mens *pass·gang* oversettes med *amble pace* og *kapp·gang* oversettes med *race walking*[13].

- **semantisk underspesifikasjon** komposita er flertydige, og kan bare oversettes på en sikker måte når konteksten er kjent.

- **ikke-komposisjonelle komposita** ikke alle komposita kan finne sitt motstykke på et nytt språk ut ifra å analysere dets bestanddeler. Enkelte fenomener heter noe helt annet på andre språk, som er satt sammen av andre leksemer. For eksempel oversettes *hjemmebrent* til engelsk med *illegally distilled spirits*, men ikke **home burnt*

- **produktivitet og frekvens** Baldwin og Tanaka (2004) forklarer ikke dette punktet, som dermed tolkes til å bety at denne typen konstruksjoner opptrer ofte, og at det dermed er viktig å ha en plan for å løse problemet.

Baldwin og Tanakas bruk av begrepet *komposisjonalitet* forstås her mer praktisk enn i den filosofiske debatten presentert i avsnitt 3.2. De definerer en oversettelseskomposisjonalitet (engelsk: *translation-compositionality*) som at det er mulig å finne frem til målspråkets oversettelse ut fra ordbokoversettelser av kildespråkets kompositums bestanddeler. Hvis oversettelsen av et kompositum kan gjenfinnes ved å oversette forledd og etterledd for seg, og kombinere dem på ulike måter, så er det oversettelseskomposisjonelt, men det er ikke (nødvendigvis) et innlegg for at det er mulig å gjenvinne den hele og fulle betydning av et sammensatt ord ved å se på delene, bare en korrekt oversettelse.

I tillegg til dette, så er det en kompliserende faktor for MT-systemer at det eksisterer flere gode oversettelser med samme semantiske innhold, men med varierende konstruksjonstyper. Et eksempel på en slik oversettelse fra gullstandarden er *fremskritts·politikk* som oversettes med både *progress policy* og *policy of progress*. Automatiske systemer som trenes opp etter

[13]Oversettelse fra KF.

å lære en fasit vil da kunne få en slagside mot en spesiell konstruksjonstype, selv om flere kan være like gode.

Forfatterne slår videre fast at 3-5% av ordene i deres undersøkte korpora bestod av ord som var en del av substantivkomposita, mens gjennomsnittlig frekvens for hvert kompositum var et beskjedent tall. Dermed vil et MT-system med stor sannsynlighet treffe på substantivkomposita, men de enkelte sammensetninger vil opptre sjelden. Dette underbygger at dannelsen av komposita er en produktiv side ved naturlige språk, som gjør det umulig å holde en liste potensielle mulige komposita oppdatert. Dermed er det behov for å ha en metode for å løpende prosessere hittil usette komposita.

Kapittel 4

Ressurser og verktøy

Dette kapittelet gir en presentasjon av de verktøy og ressurser som har vært brukt i forberedelsen av og utføringen av eksperimentene omtalt i kapittel 6. Slik blir det tydelig hvilke ressurser, programmer og versjoner av dersom er brukt i for eventuelt senere å kunne undersøke om dette kan ha påvirket resultatene.

4.1 Engelsk Stor Ordbok

Kunnskapsforlagets «Engelsk stor ordbok»(Eek, 2001) har vært tilgjengelig i elektronisk format i denne oppgaven. Dette er en ordbok som går i begge retninger, med tilsammen 217,000 oppslagsord, hvorav 60 534 er unike og norske, fordelt på 62 554 oppslag med 76 930 forskjellige betydninger av ordene, med 158 218 mulige engelske oversettelser. 45 992 av de norske oppslagene er substantiv, og 25 419 har oversettelser som består av flere enn ett ord.

Oppslagene representeres ved XML-entiteter som inneholder metainformasjon som ordklasse og kjønn, og en seksjon som inneholder samtlige betydninger av ordet, med hver sin engelske oversettelse. En seksjon med brukseksempler i begge språk følger også oppslagene.

4.1.1 Søking i KF

Gjennom LOGON-prosjektet (Oepen et al., 2004) er KF gjort tilgjengelig gjennom et eget søkegrensesnitt implementert i Common Lisp i prosjektets kildetre, som har vært tilgjengelig for denne oppgaven. Søkegrensesnittet er laget fra selve ordboken i XML-format.

4.2 Korpora

For å utføre eksperimentene, så var det nødvendig med norske korpora for å velge ut komposita til oversettelse, og engelske for å bruke som grunnlag til å velge rett oversettelse av dem.

4.2.1 Norske korpora

Det norske korpuset ble satt sammen fra to kilder, Oslo Multilingual Corpus og et spesialutviklet korpus for LOGON-prosjektet.

Oslo Multilingual Corpus

Oslo Multilingual Corpus (OMC)[1] er en flerspråklig samling parallelltekster. Johansson (2004) beskriver hvilke prinsipper som ligger til grunn for dets oppbygning. OMC er en samling tekster bestående av originaltekster og oversettelser fra mange språk. Av disse tekstsamlingene, er all unik norsk tekst, både fra originale og oversatte tekster er brukt. Tilsammen utgjør dette et korpus på 2,7 millioner ord.

LOGON korpus

Som en del av LOGON-prosjektet er det opprettet et norsk-engelsk parallellkorpus, som er beskrevet på `http://www.hf.uio.no/tekstlab/prosjekter/tourist/`. Korpuset samlet inn fra norske websider som

[1] URL: `http://www.hf.uio.no/forskningsprosjekter/sprik/english/corpus/index.html`.

dreier seg om turisme, og handler mye om fjellturer og fotturisme. I oppgaven brukes den norske delen av dette korpuset, til sammen på 253 000 ord.

4.2.2 Engelske korpora

Fordi et av spørsmålene som ble ønsket undersøkt i denne oppgaven var betydningen til størrelsen på grunnlagskorpuset, var det et poeng å bruke flere korpora av noen størrelse. I eksperimentene ble 3 engelske korpora brukt, som beskrives i det følgende.

British National Corpus

British National Corpus (BNC), som finnes beskrevet på `http://www.natcorp.ox.ac.uk/` er et tekstkorpus på 100M ord av generell karakter fra den britisk-engelske sfære, hvorav 80M fra skrevne kilder, .

Et av formålene med tekstkorpuset er å vise et bredt spekter av moderne britisk engelsk, i skrift og tale.

The North American News Text Corpus

«The North American News Text Corpus» (NANTC) er et amerikansk tekstkorpus bestående av en samling avisårganger fra 90-tallet, tilgjengelig fra `http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC95T21`.

Korpuset er på 350 milloner ord, og består av 2 årganger av New York Times, Reuters News og Wall Street Journal, samt 3 årganger av Los Angeles Times.

The AQUAINT Corpus of English text

«The AQUAINT Corpus of English text» (AQUAINT) er satt sammen av nyhetstekster fra Xinhua News Service, the New York Times og Associated Press, på tilsammen 375 millioner ord. LDC: `http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2002T31`.

4.3 VarCon

Fordi grunnlagskorpusene beskrevet over bestod av både britisk engelsk og amerikanske engelsk tekst, ble VarCon (Atkinson, 2004) i utgave 4.1 brukt til å gjøre normalisering. Fordi de amerikanske tekstkorporaene var størst, ble konverteringen gjort for BNC til amerikansk tekst.

VarCon endrer systematisk særegne britiske suffiks som *ise* til *ize*, men har også et forråd over ord og uttrykk som er forskjellige i de to variantene av engelsk, slik som at britiske *petrol* på amerikansk engelsk heter *gas(oline)*.

4.4 WordNet

Wordnet (Miller, 1995) er en stor, leksikalsk database over det engelske språket, hvor ordklassene substantiv, verb, adjektiv og adverb grupperes sammen i samlinger av synonymer. Ved siden av et web-grensesnitt, `http://wordnet.princeton.edu/perl/webwn` så er pakken fritt nedlastbar med en «open source»-lisens. WordNet tilbyr en rekke typer oppslag i disse databasene, som kan bl.a fortelle om synoymer, antonymer, hypernymer ut ifra et søkeord. En av disse søkene er å be om å få se en liste over tilknyttede ord til et gitt substantiv.

I den nedlastbare utgaven er tilgangen på spørringer noe større enn på WWW, og man kan spørre om avledende morfologi[2] ut ifra en stamme av et verb eller susbtantiv, og videre hvilke komposita et enkelt ord finnes som en bestanddel (presist: substreng) av. Denne funksjonen for avledende morfologi ble brukt til å generere oversettelser for de malene der hvor en transisjonen fra substantiv til adjektiv var nødvendig i eksperimentene i kapittel 6, ved at WordNet ble spurt etter avledninger, og resultatet av søket sortert på adjektiv.

WordNet i versjon 3.0 fra 2006 ble brukt.

[2]Engelsk:. *derivational morphology*.

4.5 RASP

RADISP/RASP (E. Briscoe & Carroll, 2002) (Robust Accurate Statistical Annotation of General Text) er et tekstanalyseverktøy som har som mål å fungere godt på ikke-domenespesifik tekst. Systemet ble utgitt i 2002 under en egen lisens som er gratis for ikkekommersiell bruk, og består av en kaskade av endelige tilstandsmaskiner og en LR-parser, som returnerer de n beste analysene av en setning. En siste komponent velger så blant dem basert på statistiske data.

RASP bruker en kombinasjon av endelige tilstandsmaskiner og statistiske metoder for å samtidig kunne tilby robust og presis analyse som statistiske tilnærminger har klart, og samtidig kunne dekke mange forskjellige teksttyper, som er en fordel med førstnevnte fremgangsmåte, uten at det er nødvendig med et stort tekstgrunnlag for opptrening.

Først gjøres et utdrag av analyseenheter (kalt *tokenization*) hvor ordene skilles fra tilhørende skilletegn, og setningsgrenser markeres. Dernest gjøres en ordklassemerking (kalt *POS tagging*), med en *Hidden Markov Model*-basert tagger, tillagt en utførlig utviklet komponent for å hanskes med ukjente ord. Disse klassifiserte ordene presenteres så for en morfologisk analyse (kalt *lemmatizer*) som ekstraherer deres respektive lemma og affikser. Disse blir så sendt videre til en probabilistisk LR parser med noe leksikalsk informasjon om sannsynligheten for at enkelte verb opptrer i fellesskap. I siste fase kommer en probabilistisk modell for å velge ut analyser basert på strukturell kontekst, støtte for underanalyse i blant analysene, og leksikalsk informasjon der det er tilgjengelig.

RASP kan presentere utdata i en rekke formater, men i hovedsak enten som grammatiske relasjoner (kalt *Grammatical relations*) (GR) eller som trær med større og mindre grad av annotering eller i et format som minner Penn Treebanks annoterte trær (PTB), men utvidet med *tags* fra et utvalg av CLAWS-tagsettet brukt i Susanne-korpuset (Sampson, 1995). Et eksempel fra (E. Briscoe & Carroll, 2002) på begge typer utdata for setningen *What debts did Qintex group leave?* følger:

```
(TOP (S (NP (DDQ What:1) (NN2 debt+s:2)) (S (VDD do+ed:3)
```

```
(S (NP (NP1 Quintex:4) (NN1 group:5)) (VV0 leave:6))))
(? ?:7))

(|obj| |leave:6_VV0| |debt+s:2_NN2|)
(|aux| |leave:6_VV0| |do+ed:3_VDD|)
(|ncsubj| |leave:6_VV0| |group:5_NN1| _)
(|ncmod| _ |group:5_NN1| |Quintex:4_NP1|)
(|det| |debt+s:2_NN2| |What:1_DDQ|)
```

T. Briscoe, Carroll og Watson (2006) presenterete en ny utgave, og det er nettopp RASP i annen versjon, som er brukt i denne oppgaven. Den nye utgaven retter opp svakheter i den første, hvor forfatterne trekker frem at gradvise forbedringer som gjør systemet bedre rustet til å dekke et videre spekter av tekster. Andre forbedringer inkluderer nytt design av GR-utdata, og flere muligheter for brukeren til å skreddersy bruken av programmet.

4.6 TADM

Toolkit for Advanced Discriminative Modeling (TADM) (R Malouf, 2006) er en C++-implementasjon av et rammeverk for å estimere parametre i MaksEnt-diskriminantmodeller, utgitt under en LPGL-lisens[3]. Med utgangspunkt i arbeider av Rob Malouf, føres TADM nå videre som et Sourceforge-prosjekt[4] i samarbeid med Jason Baldridge og Miles Osborne.

4.7 Søkegrensesnitt for Yahoo

Med støtte fra Yahoo, er det utviklet en modul til programmeringsspråket Python, tilgjengelig fra Sourceforge på `http://pysearch`.

[3]Lesser GNU Public License, URL: `http://www.gnu.org/copyleft/lesser.html`.

[4]URL: `http://tadm.sourceforge.net`.

`sourceforge.net/`. Modulen implementerer et sett av klasser og funksjoner for å arbeide mot Yahoos forskjellige søketjenester (som rene websøk, nyheter, bilder og video) på Internett. I denne oppgaven er det klassen for web-søk som ble brukt for å rangere norske komposita etter forekomst.

Modulen tillater en rekke parametere hvorav 3 ble brukt i søkene i eksperimentene fra kapittel 6. Disse var *country=NO*, som begrenser søket til å gjelde sider som er kodet som tilhørende Norge, og *language=NO*, som forteller at språket på sidene søket velges fra skal være norsk, ved siden av *type=phrase*, som sikrer at søkene som blir gjort er frasesøk. Ved siden av disse valgene er det mulig også å begrense søket etter en rekke andre kriterier som blant annet lisens (som Creative Commons-lisensen[5]), gyldige domener, antall resultater som returneres.

Yahoo tilbyr 5000 gratis søk hver dag, som gjøres tilgjengelig etter en registrering på deres hjemmesider, og ble valgt av denne praktiske årsak, det var enkelt å komme dithen at søkene kunne utføres (uten manuell søknadsbehandling).

4.8 Oslo-Bergen-taggeren

Oslo-Bergen-taggeren (OBT) (Hagen, Johannessen & Nøklestad, 2000) er en tre-fase språkanalysator bestående av en preprosessor, en *multitagger*, dvs en modul som tilegner ordene et antall ordklasser de kan ha med tilhørende egenskaper, og til sist syntaktisk og morfologisk disambiguator som velger ut den foretrukne analysen blant disse. OBT kan også merke ord som sammensetninger, ved siden av deres ordklasser med sin spesialmodul for dette. Denne prosessen er sentral i forberedelsene til eksperimentene beskrevet i kapittel 6, og er omtalt

Oslo-Bergen-taggeren kan er gjort tilgjengelig på `http://decentius.hit.uib.no:8005/cl/cgp/test.html`.

[5] `http://creativecommons.org/`.

4.8.1 Sammensetningsanalysatoren

Sammensetningsanalysatoren som er brukt i eksperimentene i denne oppgaven og diskusjonen i avsnitt 3.5.2 er basert på algoritmene presentert av Johannessen og Hauglin (1996), og er implementert av Paul Meurer ved Aksis-senteret ved Universitetet i Bergen. Den er gjort tilgjengelig på WWW på adressen `http://decentius.hit.uib.no:8005/cl/cgp/ranked-analyses.xml`.

Ved siden av dette direkte grensesnittet, er den også implementert som en modul OBT.

Kapittel 5

Metodologi

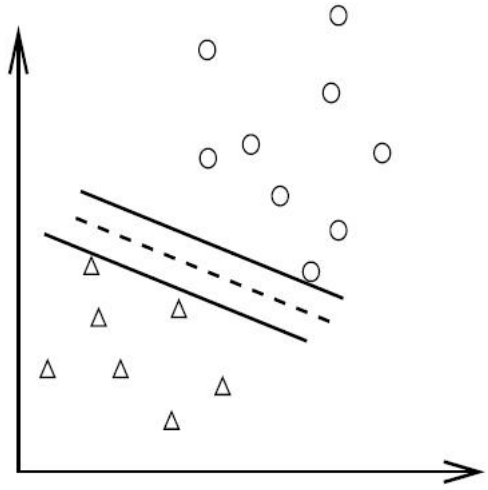
I dette kapittelet gis en beskrivelse av metoder og verktøy som er brukt til eksperimentene i denne oppgaven, ved siden av omtale av *Support Vector Machines* som har en rolle i diskusjonen av tidligere arbeider og problemet. Fordi måten Baldwin og Tanaka (2004) anvender denne metodikken blir omtalt i avsnitt 2.4.3, blir en kort presentasjon av den gitt under. Videre blir den alternative maskinlæringsteknikken basert på maksimal entropi som ble brukt i eksperimentene i kapittel 6 presentert, etterfulgt av en gjennomgang av de praktiske sidene ved hvordan eksperimentene ble utført.

5.1 Support Vector Machines

Support Vector Machines (norsk: *Støttevektormaskiner*), forkortet SVM, er en betegnelse på en maskinlæringsteknikk som utfører *kontrollert læring*[1], som betyr at de kan utføre klassifikasjon og sortering av data, hvor treningsdata er forberedt i noen grad. I SVM-enes tilfelle brukes manuelt definerte *etiketter*[2] på observasjonene, som maskinen kan brukes til å klassifisere etter. SVM-tilnærmingen til maskinlæring har et geometrisk utgangspunkt, hvor en er ute etter å finne et lineært hyperplan som skiller

[1] Engelsk: *supervised learning*.

[2] Engelsk: *labels*.



Figur 5.1: *Observasjoner som kan skilles lineært. Figur fra (Lin, 2006).*

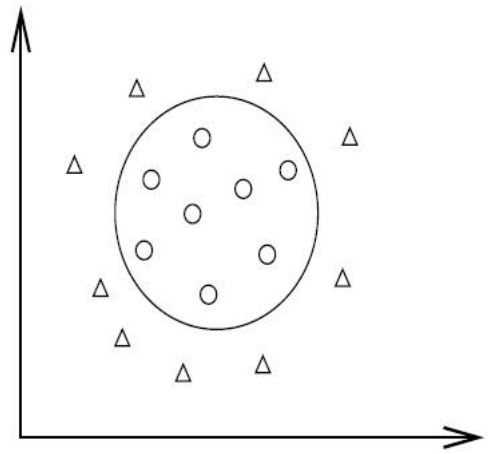
observasjonene. Når man da gjør en ny observasjon, kan den klassifiseres ved å regne ut på hvilken side av planet den havner.

Hvis man da har en samling observasjonsvektorer $x_i ... x_n$, hvor hver x_i kan være en vektor som forteller om de numeriske verdiene i vektoren, så opptrer en en kategori y_i sammen med observasjonen. Hvis y_i kan anta verdiene {-1,1}, så er formålet å finne et hyperplan som skiller disse punktene fra hverandre. Ethvert hyperplan kan beskrives med punktene som tilfredsstiller $w \cdot x + b = 0$, der w er en vektor som inneholder vektingen av de ulike trekkene, og b angir avstanden fra origo. Et eksempel på et slikt plan vises i figur 5.1. Det er mange mulige plan, og hyperplanet som skiller punktene vil ha egenskapene (Lin, 2006):

$$y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1, \quad \text{for alle } 1 \leq i \leq n \tag{5.1}$$

Klassifikatorfunksjonen blir $f(x) = sgn(w \cdot x_i + b)$, altså fortegnet på vektorproduktet mellom vektoren w og observasjonene.

Men som det går frem av figur 5.1 er det mange plan som skiller de to settene med observasjoner. Og det planet som søkes er det med størst mulig margin til observasjonene, i figur 5.1 illustrert ved den stiplede linjen. En størst mulig margin vil gjøre modellen mer generell, fordi flest



Figur 5.2: *Observasjoner som ikke kan skilles lineært. Figur fra (Lin, 2006).*

mulig datasett vil kunne passe med separeringen når marginen til hver side er størst. Avstanden mellom $w \cdot x_i + b = 1$ og $w \cdot x_i + b = -1$ er $2/|w| = 2/\sqrt{w \cdot w}$, representert ved avstanden mellom de to heltrukne linjene i figuren. For å gjøre denne avstanden størst mulig må dermed *normen*, $|w|$ til vektvektoren minimeres. Denne avstanden finnes med følgelde optimeringsproblem:

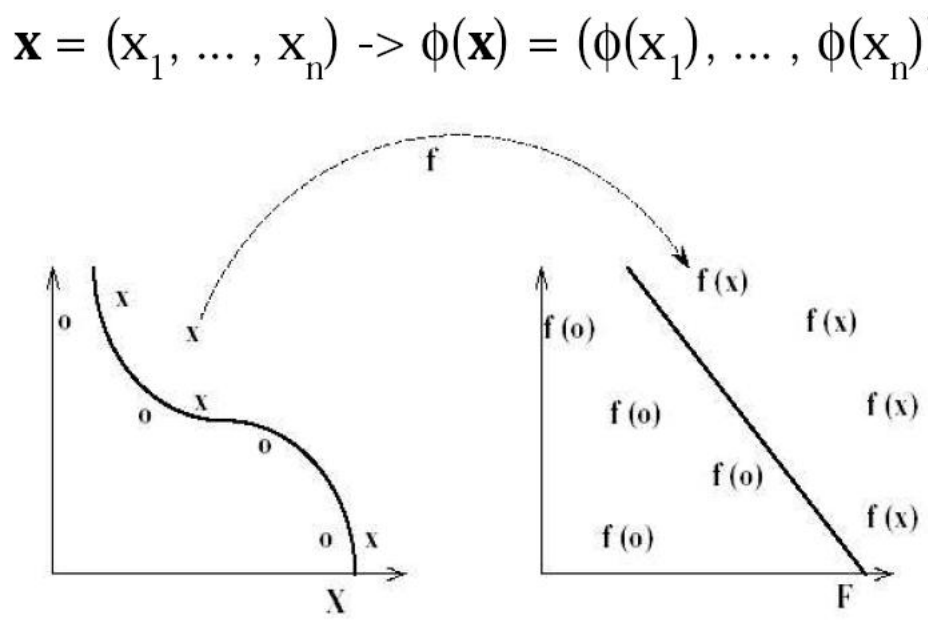
$$\min_{w,b} \frac{1}{2} w \cdot w \tag{5.2}$$

med bibetingelsene

$$y_i(w \cdot x_i + b) \geq 1$$
$$i = 1.....n$$

Det er imidlertid ikke nødvendigvis slik at observasjonene lar seg skille ad på en så ren og enkel måte. Hvis fordelingen av observasjoner istedet ser ut som i 5.2, så er det ikke lenger mulig å trekke en rett linje for å skille dem.

Dette kan avhjelpes ved å tolerere feil når vektvektoren utarbeides. Dersom toleransen for feil skrus høyt nok, vil det også alltid være mulig å finne et hyperplan som skiller observasjonene uansett hvordan de fordeler seg, når det kan godta at et ubegrenset antall observasjoner ligger «feil».



Figur 5.3: *Punkter som kan skilles etter å ha blitt projisert med en funksjon.*

Et annet alternativ er å gjengi informasjonen i observasjonsvektorene på i en høyere dimensjon, ved eksempelvis å lage en ny vektor som inkluderer verdiene multiplisert med seg selv. Selv om det ikke vil gå å skille observasjonsvektorene ad med et hyperplan i sin opprinnelige form, kan det likevel være mulig å gjøre dette i en høyere dimensjon. Sirkelen som er tegnet i figur 5.2 antyder en måte å skille punktene på, eksempelvis ved at punktene de sirkulære punktene i en tredje dimensjon ligger over de andre, og det kan etableres et plan mellom punktene. Jo høyere dimensjon man projiserer punktene til, jo større sjanse er det for at de kan skilles lineært.

Projeksjonen til av vektoren til en høyere dimensjon ved betegnes med en funksjon $\phi(x)$. Figur 5.3 viser hvordan en slik funksjon kan fungere. Disse to metodene blir kombinert i Cortes og Vapnik (1995) sin SVM-formulering fra 1995, hvor feilmarginen representeres ved slakkvariabelen ξ og projeksjonen ved en funksjon ϕ.

$$\min_{w,b,\xi} \frac{1}{2} w \cdot w + C \sum_{i=1}^{n} \xi_i$$

med bibetingelsene:

$$
\begin{aligned}
&y_i(w \cdot \phi(x_i) + b) \geq 1 - \xi_i \\
&\xi_i \geq 0 \\
&i = 1.....n
\end{aligned}
$$

Fordi at dette problemet er konvekst, med lineære bibetingelser, har det et Langrange-dual. Og formuleringen over kan ses på som et SVM-primal, med tilhørende dualformulering:

$$\min_{\alpha} \sum_{i=1}^{n} \alpha_i - \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \phi(x_i) \cdot \phi(x_j)$$

med bibetingelsene:

$$
\begin{aligned}
&y \cdot \alpha = 0 \\
&0 \leq \alpha_i \leq C \\
&i = 1.....n
\end{aligned}
$$

slakkvariabelen ξ finnes ikke lenger i dualformuleringen, hvor bibetingelsene fra primalformuleringen er bragt inn som koeffisienter. Problemet finner sitt optimum i $w = \sum_{i=1}^{n} \alpha_i y_i \phi(x_i)$, som dermed gir en ny klassifikatorfunksjon, $sgn(\sum_{i=1}^{n} \alpha_i y_i \phi(x_i)\phi(x))$.

Det er kostbart å kalkulere $\phi(x_i)\phi(x_j)$ fordi trekkvektorene må løftes opp til en høyere dimensjon ved hjelp av ϕ før indreproduktet kan regnes ut. Ved å definere dette indreproduktet som en *kjerne*, $K(x_i, x_j)$, kan en verdi regnes ut med en annen funksjon uten å måtte gjøre transformeringen til et høyere nivå. Det er forskjellige måter å regne ut denne verdien på, kalt forskjellige kjerner. Kjernen kan også brukes i klassfikatorfunksjonen, som alternativt formuleres $sgn(\sum_{i=1}^{n} \alpha_i y_i K(x_i, (x)))$. Dette *kjernetrikset* brukes i SVM-implementasjoner for å komme frem til et skille-hyperplan.

5.2 Maksimalentropi

Maksimalentropi-prinsippet har gitt opphav til en familie av maskinlæringsteknikker som forsøker å etablere en modell for et sett observasjoner med tilhørende etiketter, hvis sannsynlighetsfordeling har høyest *entropi*. Maksimalentropi blir i det videre forkortet med MaksEnt. Entropi som begrep i informasjonsteorien, såvidt nevnt i avsnitt 2.4, kan føres tilbake til (Shannon, 1948), og vil bli behandlet kort under. En MaksEnt-modell kan brukes til å predikere hvilken etikett en ny observasjon burde ha, etter å ha blitt trent opp på observasjoner som allerede er merket korrekt.

Hvis SVM-teknikken i forrige avsnitt har en geometrisk tilnærming til klassifikasjon, så har maksimalentropi-tilnærmingen en probabilistisk innfallsvinkel til problemet, slik at det kan regnes ut en sannsynlighet for at en observasjon tilhører en gitt klasse, når en MaksEnt-modell er brukt. Formålet er å gjøre så rede for treningsdataene, uten å anta noe mer om dataene ut over det som er absolutt nødvendig.

5.2.1 Informasjonsentropi

Med utgangspunkt i kommunikasjonsteori er informasjonsentropien et forsøk på å kvantifisere informasjonsverdien som ligger i et utsagn for å kunne overføre det på en enklest måte, uten tap av informasjon.

Wikipedia[3] gir en totrinns definisjon på entropien til en stokastisk variabel X som kan anta verdiene $x_1, ...x_n$:

$$H(X) = E(I(X)) = -\sum_{i=1}^{n} p(x_i) \log p(x_i) \tag{5.3}$$

I(X) er her informasjonsgehalten til X. Informasjonsgehalten til en et utfall ω_n er definert som $-\log(p(\omega_n))$, slik at dens forventing ser ut som i likning 5.3[4]. Informasjonsgehalten er også omtalt som *surprisal* på engelsk,

[3] http://en.wikipedia.org/wiki/Information_entropy besøkt 30. juli 2008.

[4] http://en.wikipedia.org/wiki/Self-information besøkt 30. juli 2008.

forstått som et uttrykk for hvor overrasket man blir. Dermed kan entropien tolkes som hvor overrasket over utfallet av en stokastisk variabel man kan forvente å bli. Dermed kan det virke kontraintuitivt å ønske å maksimere denne verdien, fordi at det kan virke ugunstig at en modell som etableres for et sett observasjoner skal være overrasket over hvordan den opptrer, dermed befengt med usikkerhet. Men dette kan ses på analogt til ønsket fra SVM-ene diskutert i avsnitt 5.1 om å finne et hyperplan med mest mulig margin, for å være så generell som mulig. Tilsvarende vil en modell med høyest mulig forventet «overraskethet» som samtidig passer med treningsdataene være mer generell, og dermed presumptivt bedre rustet til å gjøre predikasjoner på usette data.

Entropien vil være på sitt maksimale dersom samtlige utfall er like sannsynlige, og i så tilfelle vil den øke med utfallsrommets størrelse. Entropi kan dermed også betraktes som et mål på hvor uniform modellen er.

5.2.2 Et eksempel, å oversette *in*

Berger, Della Pietra og Della Pietra (1996) bruker en eksemplifisert probabilistisk oversettelse av det engelske ordet *in* til fransk til å motivere sentrale egenskaper ved MaksEnt-tankegangen. Et ønske om å modellere hvordan en ekspert hadde oversatt den engelske preposisjonen *in*, gir hver potensielle franske oversettelse, f av ordet en probabilitet $p(f)$. For å utvikle modellen samles det inn eksempler på eksperters språkutøvelse (eksempelvis fra et parallellkorpus), og målet er å samle inn fakta om beslutningsprosessen, som kan gjøre det mulig å lage en modell av denne prosessen.

Hvis det viser seg at ekspertene velger mellom følgende fem fraser, *dans, en, à, au cours de, pendant*, kan den første skranken til modellen p etableres:

$$p(dans) + p(en) + p(à) + p(au\ cours\ de) + p(pendant) = 1$$

Denne beskrankningen sørger for at p virkelig er en probabilitet. Men det er åpenbart mange modeller som tilfredsstiler skranken, og spørsmålet

er da hvilken som skal velges. Dersom ikke mer informasjon er kjent, fremstår det som naturlig å gi alle frasene lik sannsynlighet, $\frac{1}{5}$. Dette vil også være den mest uniforme modellen, som dermed gir den høyeste entropien. Intuisjonen stemmer overens med MaksEnt-prinsippet.

Men hvis en ny beskrankning introduseres, som at ekspertene i 3 av 10 tilfeller velger *dans* eller *en*, så kompliseres bildet. Og introduseres enda en beskrankning som sier at *dans* eller *à* ble valgt i $\frac{1}{2}$ tilfeller er det ikke lenger trivielt å slutte hva som er den mest uniforme modellen. En MaksEnt-modell forsøker å finne nettopp denne modellen.

5.2.3 MaksEnt-modeller

For å lage en modell over eksempelet i forrige avsnitt, betraktes de forskjellige oversettelsene som mulige utfall y av en stokastisk prosess, som opptrer i en kontekst x som eksempelvis kan være setningen oversettelsen står i, eller et gitt antall ord foran og bak. Treningsdata kan dermed organiseres som par av kontekster $x \in \mathcal{X}$ og $y \in \mathcal{Y}$ dvs. $(x_1, y_1), (x_2, y_2).....(x_N, y_N)$. Den empiriske fordelingen av parene, $\tilde{p}$ er definert som

$$\tilde{p}(x, y) \equiv \frac{1}{N} \times \text{frek(x,y)}$$

Kontekstuell kunnskap fanges opp som *trekk*[5] f_i ved konteksten som inntreffer samtidig med et resultat. Et trekk er en funksjon fra kryssproduktet av kontekster og kategorier, $\mathcal{X} \times \mathcal{Y}$, til $\mathcal{R}$. Ofte er trekkene binære, slik at mulige verdier begrenser seg til $\{0, 1\}$. Et eksempel på et slik trekk kan være:

$$f(x, y) = \begin{cases} 1 & \text{hvis y er } \textit{en} \text{ og det engelske ordet } \textit{April} \text{ kommer etter } \textit{in}, \\ 0 & \text{i andre tilfeller} \end{cases}$$

[5] engelsk: *features*.

Forventingsverdien til f med hensyn på den empiriske fordelingen $\tilde{p}(x, y)$ er:

$$E_{\tilde{p}} f_j = \sum_{x,y} \tilde{p}(x, y) f_j(x, y) \tag{5.4}$$

Forventingsverdien til f med hensyn på den modellerte fordelingen p er:

$$E_p f_j = \sum_{x,y} p(x, y) f_j(x, y) \tag{5.5}$$

Teoretisk er det mulig å ha problemer hvor likning 5.5 kan regnes ut slik den står. I mange NLP-applikasjoner kan det være svært vanskelig å summere over (X), fordi antall kontekster (f.eks grammatiske setninger) kan være enormt eller uendelig. Ved Bayes' lov, så kan forventingen uttrykkes:

$$E_p f_j = \sum_{x,y} p(x) p(y|x) f_j(x, y) \tag{5.6}$$

Derfor anslås modellens sannsynlighetsfordeling for kontekster $p(x)$ til å være lik den empiriske, $\tilde{p}(x)$. Ratnaparkhi (1997) nevner i sin redegjørelse for bruk av MaksEnt-modeller til POS-tagging at dette er en approksimasjon, mens Berger et al. (1996) definerer modellens forventningsverdi til å være utregnet med den empiriske, og ikke den teoretiske fordelingen.

Dette gir følgende modellforventing:

$$E_p f_j = \sum_{x,y} \tilde{p}(x) p(y|x) f_j(x, y) \tag{5.7}$$

Denne modellforventingen blir begrenset til å være lik den empiriske i modellen, $E_{\tilde{p}} f_i = E_p f_i$, eksplisitt

$$\sum_{x,y} \tilde{p}(x) p(y|x) f_j(x, y) = \sum_{x,y} \tilde{p}(x, y) f_j(x, y) \tag{5.8}$$

for alle trekk f_j.

Maksimalentropiprinsippet

Hvis $\mathcal{P}$ er mengden av sannsynlighetsfordelinger, så er $\mathcal{C}$ en delmengde av $\mathcal{P}$ definert ved

$$\mathcal{C} \equiv \{p \in \mathcal{P} | E_{\tilde{p}} f_i = E_p f_i \text{ for } i \in \{1, 2, ...n\}\} \tag{5.9}$$

Berger et al. (1996) presenterer maksimalentropiprinsippet som følger:

For å velge en modell fra settet $\mathcal{C}$ av gyldige sannsynlighetsfordelinger, velg modellen $p_* \in \mathcal{C}$ med maksimal entropi $H(p)$:

$$p_* = \arg\max_{p \in \mathcal{C}} H(p) \tag{5.10}$$

Optimeringsproblemet

Dermed er det definert et optimeringsproblem, som søker å maksimere entropien $H(p)$ under bibetingelsene som velger ut delmengden $\mathcal{C}$ vist i likning 5.9. For å løse optimeringen settes en Lagrange-funksjon opp:

$$\Lambda(p, \lambda) = H(p) - \sum_i (\lambda_i (E_{\tilde{p}}(f_i) - E_p(f_i)) \tag{5.11}$$

Klein og Manning (2003) viser at i Langrange-funksjonens optimum er $p(x, y) \propto exp \sum_i \lambda i f_i$ ved å derivere Langrange-funksjonen med hensyn på p og sette lik 0. Den betingede sannsynligheten for $p(y|x)$ har denne parametriske formen:

$$p(y|x) = \frac{1}{Z(x)} exp(\sum_i \lambda_i f_i(x, y)) \tag{5.12}$$

Hvor $Z(x)$ er en normalingseringsfunksjon over alle y, som følger av å innføre en beskrankning som sier at $\sum_y p(y|x) = 1$.

Optimeringsproblemet og estimering av maksimal sannsynlighet

Maximum Likelihood-metoden er en statistisk metode, også omtalt som en sannsynlighetsmaksimeringsestimator på norsk, som ønsker å finne frem til hvilken probabilitet som passer best til et sett av observerte data. Begrepet *likelihood* defineres som hvor sannsynlige de observerte resultatene er gitt modellene som brukes til å komme frem til sannsynlighetsfordelingen. I det følgende vil begrepet *sannsynlighet* bli brukt på denne måten. Et enkelt eksempel er 100 myntkast, hvorav 56 er kron og 44 mynt. Dette er en binomisk prosess (slike prosesser vil bli nærmere omtalt i avsnitt 5.4), hvor sannsynligheten for 56 av 100 er gitt ved $\binom{100}{56}p^{56}(1-p)^{44}$. Å finne høyest sannsynlighet for denne funksjonen vil si å finne hvilken p som gir høyest verdi, eller passer best med de observerte dataene. I dette tilfelle er det ikke overraskende 0.56 som er probabiliteten som passer best med de observerte dataene. Men i statistiske modeller som baserer seg på flere parametere er dette ikke like opplagt, eksempelvis der sannsynligheten beregnes ved å summere over et antall antall parametere som i likning 5.12, hvor samtlige λ_i påvirker probabiliteten.

Sannsynlighet-funksjonen har samme maksimum som logaritmen av den, kalt log-sannsynlighet. Det kan vises at log-sannsynligheten til den eksponensielle modellen $p(y|x)$ finner sitt maksimum for de samme verdiene av λ som Lagrange-funksjonen over. Dette betyr at modellen $p^* \in \mathcal{C}$ som har den høyeste entropien også er den modellen som har høyest *likelihood*, eller passer best med de observerte dataene.

Lagrange-multiplikatorene λ_i tolkes som vekter for hvert trekk som forteller hvor mye det bidrar med for å øke sannsynligheten for at en observasjon i en kontekst x har en etikett y.

5.2.4 Parameterestimering

Selv om vi vet hvordan p ser ut i optimum, er det ikke like enkelt å beregne verdiene for λ_i. Malouf (2002) påpeker at det ikke er mulig å komme frem til en «lukket løsning» (i betydningen at en løsning kan finnes på en analytisk måte), og viser til ulike algoritimer som løser problemet iterativt,

ved at de starter med et utgangspunkt og forbedrer den med ved å justere verdier til den oppnådde forbedringen er under en viss terskel.

5.2.5 Anvendelser i NLP

MaksEnt-rammeverket har kommet til anvendelse i en rekke NLP-applikasjoner som eksempelvis tagging (Ratnaparkhi, 1996), å finne setningsgrenser (Mikheev, 2000), parsing (Charniak & Johnson, 2005), og disambiguering (Ratnaparkhi, 1998). Rammeverket har den fordelen at et ubegrenset antall trekk som kommer fra mange forskjellige kilder kan brukes i samme modell. Eksempelvis kan både syntaktisk, semantisk og annen kontekstuell informasjon brukes i samme modell i den grad de lar seg ekstrahere. Når det gjelder kombinasjoner av trekk, skjer dette ikke automatisk i en MaksEnt-modell, slik at dersom samhandling mellom trekk er interessant, må disse legges til når trekkene utferdiges. I NLP-anvendelser av MaksEnt-systemer kan antall trekk komme opp i millionvis, slik at en automatisk kombinasjon av samtlige trekk ikke nødvendigvis er mulig eller hensiktsmessig betinget tilgjengelig maskinpark. Derfor er slike kombinasjonstrekk ofte satt opp med bakgrunn i lingvistisk kunnskap. Et eksempel på dette fra en MaksEnt-tagger kan være at forrige ordklasse, samt suffikset (f.eks de 3 siste bokstavene) i det inneværende ordet (ordet som skal tilegnes en ordklasse) antas å fortelle noe om hvilken ordklasse dette ordet har.

5.2.6 Glatting

Det kan være et behov for å utføre glatting[6] av sannsynlighetsfordelingene som resulterer fra MaksEnt-beregninger i NLP-sammenheng av flere årsaker. Det kan gjøre sannsynlighetsfordelingene «mykere», altså mindre ekstreme fordelinger, det kan gi mer vekt til trekk som forklarer mer, det kan tillate flere trekk brukt, og det kan forbedre ytelse. Toutanova, Klein,

[6] engelsk: *smoothing*.

Manning og Singer (2003) viser til en nøyaktighetsforbedring fra 96.54% til 97.10% i sitt POS-taggings-prosjekt.

Måten glattingen foregår kan være å begrense antall iterasjoner parameterestimeringsalgoritmene tillates, eller ved å sløyfe trekk som ikke forekommer mer enn en viss terskel, eller å gi optimeringsproblemet ytterligere begrensinger, eller å sette en terskel for forbedring i hver iterasjon av estimeringsalgoritmen.

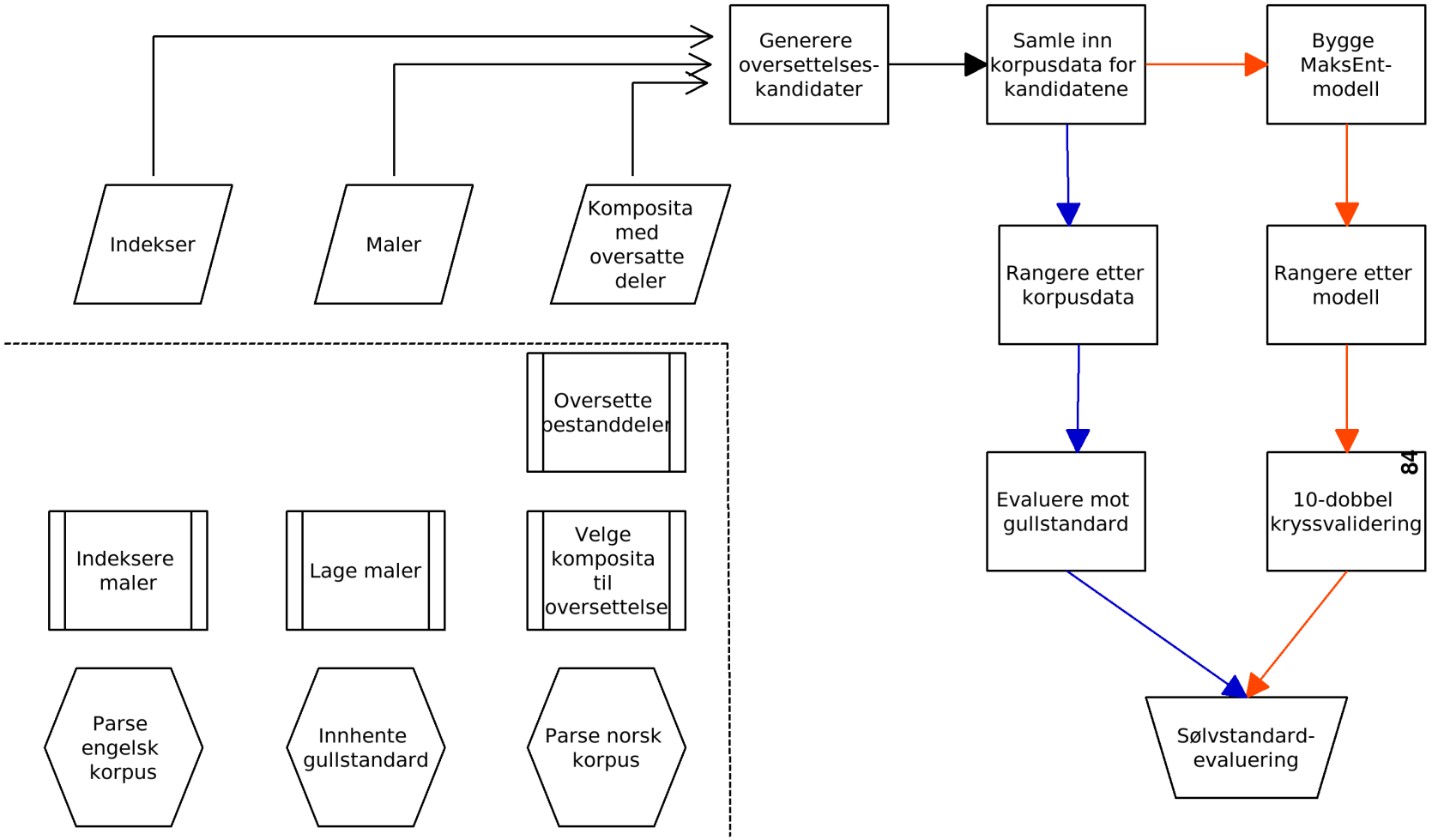
En utbredt slik glatt er å utstyre optimeringsproblemet med en Gaussiansk prior, en forkunnskap som vil få sannsynlighetsfordelingen til å like mer på en normalfordeling, en klokkeformet Gauss-kurve. Med et utgangspunkt der MaksEnt-problemet betraktes som å maksimere log-sannsynlighet til dataene gitt modellen, kan en alternativ *a posteriori*-formulering lages, med en antagelse om at det finnes en *a priori* sannsynlighetsfordeling for λ som skal estimeres. En slik posteriør (der noe kommer i etterkant) formulering, der λ regnes som kjent kan settes opp som summen av logaritmen til denne fordelingen, og log-sannsynligheten. I det nye glattede optimeringsproblemet blir denne *a priori*-fordelingen til λ satt til å være normalfordelt med et gjennomsnitt μ og en varians σ^2.

I kapittel 6 vil bruken av MaksEnt-modeller til å rangere oversettelses-kandidater av oppstykkede komposita bli behandlet. Dette inkluderer en kort diskusjon av glatting for utvalg av modellene som ble brukt.

5.3 Eksperimentmiljø

5.3.1 Oversikt

En oversikt over arbeidsflyten i de gjennomførte eksperimentene kan sees i figur 5.4. Som det fremgår av figuren, så kan arbeidene deles inn i forberedende arbeidsprosesser som ble foretatt i forkant av eksperimentene, og selve utførelsen av oversettelse og evaluering. Som i eksperimentene til Baldwin og Tanaka (2004) så kreves et forprosessert målspråkskorpus til å sanke data for å beskrive hver oversettelseskandidat



Figur 5.4: *Flytdiagram over implementasjon.*

i kvantitative termer, som ble gjort med RASP, som beskrevet i 4.5. Videre ble de parsede engelske korporaene indeksert med Perl-kode, med statistikk for hver engelske konstruksjonstype.

Ordene til oversettelse ble hentet fra et sammensatt norsk korpus, som beskrevet i avsnitt 4.2.1 etter at dette ble tagget med Oslo-Bergen-taggeren, og deretter silet for riktig type komposita, og rangert i tre frekvensbånd. De 750 utvalgte kompositaene ble forelagt en to-språklig informant som ga til beste sin beste oversettelse av ordene. På bakgrunn av disse oversettelsene ble maler for engelske oversettelser laget, der de hadde en komposisjonell oversettelse, i betydningen definert i avsnitt 3.4.5. Fordi norsk skiller seg fra engelsk med henblikk på kompositadannelse, er det også nødvendig å dele opp ordene som skal oversettes for å kunne generere partielle oversettelser til sammensetning. Dette ble gjort med sammensetningsanalysatoren nærmere diskutert i avsnitt 3.5.2 gjennom et XML-grensesnitt som ble behandlet i Python.

En liste med komposita til oversettelse med partielle oversettelser ble så behandlet i Common Lisp, hvor bestanddelene ble satt sammen til oversettelseskandidater, og kvantitative data fra korpus-dataene som var lest inn ble tillagt hver kandidat. Kandidatene ble nå rangert enten etter metoder som kunne beregnes direkte, som CTQ, eller ved at en MaksEnt-modell ble bygget, og rangert etter denne. Avslutningsvis ble resultatene evaluert, enten i koden direkte, eller i MaksEnt-tilfellene via 10-dobbel kryssvalidering gjort i Lisp og bash.

5.3.2 Behandling av engelske korpora

Samtlige korpora forelå i ulike XML-format, og råteksten måtte først ekstraheres, som ble gjort med Pythons PyXML-modul. Resultatfilene ble komprimert med bzip2, og deretter kanalisert til RASP, som parset dem i et format som bestod av Penn Treebank-aktige trær, men annotert med et utvalg av CLAWS-tagsettet (Sampson, 1995). RASP ble eksekvert med en begrensning på 3 sekunders arbeid per setning, og resultatfilene ble igjen komprimert. Under parsingen ble det ikke brukt en begrensning

på antall tillatte ord i hver parse, som viste seg å være en tabbe, fordi at parserscriptet steilet dersom den kom over muntlige passasjer med utradisjonell tegnsetting, og denne prosessen tok unødvendig mye tid. Av og til ble den manuelt avbrutt og gjenopptatt på et nytt stadium, som gjør det vanskelig å nøyaktig fastslå hvor mye tid som både faktisk ble brukt og burde ha blitt brukt på å gjennomføre parsingen. Likevel er det en stor jobb å parse korpora på til sammen 825 millioner ord som krever både stor lagringskapasitet og prosessorkraft, og en måndes tidsbruk på en rask maskin er et omtrentlig anslag. Utdata fra RASP så ut som følger:

```
(|The_AT| |news_NN1| |be+ed_VBDZ| |release+ed_VVN|
|after_ICS| |the_AT| |close_NN1| |of_IO| |stock_NN1|
|market_NN1| |trading_NN1| |yesterday_RT| ._.) 1 ; (-15.041)

(TOP
 (S (NP (AT The) (NN1 news))
  (VP (VP (VP (VBDZ be+ed) (VVN release+ed)) (PP (ICS after)
(NP (AT the) (NN1 close) (PP (IO of)
(NP (NN1 stock) (NN1 market) (NN1 trading))))))
(RT yesterday)))
 (. .))
```

med to typer utdata for hver setning. Det var den nederste tre-analysen som ble beholdt for videre prosessering. Et eksempel på en engelsk konstruksjon som ble indeksert er her *(NP (NN1 stock) (NN1 market) (NN1 trading))*, to substantiv direkte etterfulgt av et annet, inne i en NP-klausul. I forsøkene der parseregenskapene fra RASP ble lagt til grunn ble den omsluttende NP-klasulen forutsatt. Alternativt ble bare riktig tag vurdert som nødvendig for at konstruksjonstypene ble indeksert i forsøkene hvor RASP ble brukt som tagger.

5.3.3 Indeksering av parsede engelske korpora

På tross av at måten til sist ble løst var enkel, viste det seg at det mest tidkrevende med implementasjonen av disse eksperimentene var å finne

en måte å representere korpusdataene på, slik at de kunne hentes ut for hver genererte oversettelseskandidat, som var mer oversiktlig. Først ble en løsning med tgrep2[7] forsøkt, et verktøy for å søke i trær som lagrer indekseringer av trær som vist over i et eget format, men det ble raskt klart at dette var for tidkrevende. For hver oversettelseskandidat behøvdes tre søk, og når hvert av dem kunne ta mer enn et minutt, så var det klart at det gikk for langsomt. For hvert norske kompositum kunne mange hundre kandidater bli generert, avhengig av hvor mange oversettelser som fantes av hvert delledd.

Dernest ble en egen søkealgoritme utviklet som gikk igjennom hvert tre i de parsede korpusene, som kunne tas rett inn som lister i Lisp, og så etter riktige leksikalske kategorier, som NP og NN1, og kontrollerte om det var en godtagbar klasse som etterfulgte avhengig av hvilken mal som ble brukt. NN1 betegner et substantiv i entall, og NN2 et substantiv i flertall. I så fall ble høyresidene av listene med den ønskede etikett lagret i en liste i Lisp. Listen var en liste over lister ord som opptrådte som første ord etterfulgt av mulige andre ord for dette ordet, og hvor ofte det var sett. Selv om dette var raskere en tgrep2-løsningen over, så tok det fremdeles nærmere et døgn å lage hver slik tabell, og oppslagene ble også for kostbare.

Etter dette ble det forsøkt å gjøre søkene ved hjelp av regulære uttrykk i Perl i stedet. Instanser av hver mal ble søkt etter i utdataene fra RASP, og for hver av dem ble forleddet og etterleddet i konstruksjonstypen lagret (de kan av og til bestå av to ord), og en liste ble skrevet ut hvor hver instans av konstruksjonstypen sto oppført sammen med tallfesting av hvor mange ganger den var sett. Når i tillegg parallelisering lik den som er omtalt i avsnitt 5.3.9 ble brukt, så var indekseringen av samtlige maler redusert til noen få minutter. Disse kunne så leses inn i hashtabeller i Lisp, noe som tok få minutter, og deretter lå samtlige i hurtigminnet, og oppslag i disse hashtabellene tok ubetydelig tid.

Mange av malene involverer preposisjonsfraser, og i en situasjon

[7] URL:http://tedlab.mit.edu/~dr/TGrep2/.

der syntaktisk informasjon ikke beholdes, altså der RASP brukes som en tagger, vil det ikke være mulig å skille mellom de to analysene av setningsfragmentet *buy books for children* (norsk: *kjøpe bøker for barn*). Setningen kan enten analyseres som

```
(VP (VB buy) (NP (NNS books)
             (PP (IN for) (NP(NN children)))))
```

hvor preposisjonsfrasen er tilknyttet nominalfrasen, og

```
(VP (VB buy) (NP (NNS books))
    (PP (IN for) (NP(NN children))))
```

hvor preposisjonsfrasen er tilknyttet verbalfrasen. Forskjellen mellom de to analysene er at i det første tilfellet kjøpes bøker ment for barn, og i det andre kjøpe bøkes av ukjent slag for barn. Ved å forutsette den omsluttende NP-etiketten i de regulære uttrykkene der den syntaktiske informasjonen fra RASP ble brukt, ble bare den ønskede typen der preposisjonsfrasen modifiserer nominalfrasen indeksert.

5.3.4 Morfologisk analyse av norske korpora

Morfologisk analyse av norske koropra ble gjort med Oslo-Bergen-taggeren, som også ble disktutert fra en teoretisk synsvinkel i avsnitt 3.5.2. Her var oppgaven langt mindre, og prosessorkraft var ikke en bekymring i parsingen av de 2,7 millioner ordene. Oslo-Bergen-taggeren ble benyttet gjennom et Common Lisp-grensesnitt utviklet i forbindelse med LOGON-prosjektet (Oepen et al., 2004). ≪Multi-tagging≫ ble valgt som analysenivå under parsingen, som betyr at flere mulige ordklasser for hvert ord kom tilbake fra taggeren. Et eksempel på en slik analyse er gjengitt under:

```
"<Tran>"
        "Tran" subst prop <person>
        "tran" subst appell fem ub ent
```

```
        "tran" subst appell mask ub ent
"<kalles>"
        "kalle" verb pres inf pass i1 tr11 pa1 d5 pr6 pr3..
        "kalle" verb inf i1 tr11 pa1 d5 pr6 pr3 pr4 gen
"<fiskeleverolje>"
        "fiskeleverolje" subst appell fem ub ent samset
"<på>"
        "på" prep
"<engelsk>"
        "engelsk" adj ub m/f ent pos
        "engelsk" adj nøyt ub ent pos
        "engelsk" subst appell mask ub ent
        "engelsk" subst appell nøyt ubøy
"<.>"
        "$." clb <<< <punkt>
```

De etterspurte kompositaene ble identifisert ved at de var merket som både «samset» og «subst», som ble gjort med unix-kommandolinjeverktøy. Alle sammensetninger står oppført i entall.

5.3.5 Siling av komposita

37 058 ord var merket som både *samset* og *subst*, hvorav 22 339 var unike. Disse ble så filtrert gjennom sammensetningsanalysatoren, for å velge ut dem som var mulig å stykke opp i to deler, hvorav både forledd og etterledd var substantiv. Dette var tilfellet for 14 442 ord, som dannet grunnlaget for frekvenssortering og tilfeldig utvalg. Samtlige ord med bindestrek ble her forkastet fordi at sammensetningsanalysatoren rapporterer at forleddet alltid er ukjent for slike sammensetninger, uansett hva det er. Listen ble nå rangert etter antall treff i det samlede korpuset de var hentet fra, men denne fremgangsmåten ble forlatt fordi at et for lavt antall hadde mer enn 10 treff, bare 1% lå over. Istedet ble Yahoos pySearch-modul brukt til å søke etter ordene på WWW. Dette grensesnittet er omtalt i avsnitt 4.7. Søkene ble konfigurert slik at de bare søkte sider som var flagget som norske, og fra sider med norske url-er. Dette for å

unngå skandinavisk interferens, slik at ord som også staves identisk på svensk eller dansk fikk en uforholdsmessig høy rangering. 4946 ord hadde flere enn 10 treff fra Yahoo, og av disse kunne 750 trekkes ut, og rangeres etter frekvens. Manuell inspeksjon viste imidlertid at fordi en liten del av det norske korpuset bestod av nynorsk tekst, så ord som *krok·ane* og *ende·lege* ble godkjent som sammensetninger av denne algoritmen, fordi at *krok*, *ane*, *ende* og *lege* er norske substantiv. Feilene var imidlertid enkle å fjerne med regulære uttrykk, fordi at listen over treff for hvert nynorske endelse var så liten at det var lett å se om et ord som skulle vært der som for eksempel *dverg·lege*, stod i fare for å bli fjernet. Noen andre feil ble luket vekk som følge av manuell inspeksjon. Disse kunne eksempelvis være låneord som *akt·ion*, *dom·mare*, men også kuriøse tilfeller som *bed·rite* eller *nedbør·rike* hvis substantiviske betydning ble vurdert som for perifer i forhold til den adjektiviske, som ellers tilfredsstiller kravene over. Fra de gjenværende 4892 ord ble 750 tilfeldige ord trukket ut og sortert etter frekvens. Lavfrekvensordene strakk seg fra 10 (*anretnings·disk*) til 52 forekomster (*natt·skygge*), middelsordene fra 52 (*natur·symbolikk*) til 195 (*bunn·snøre*) forekomster, og det høyeste båndet fra 313 (*glas·skjerm*) til 130 000 treff (*kurs·start*).

5.3.6 Utarbeidelse av fasit og maler

De 750 ordene over tre frekvensbånd ble slått opp i Kunnskapsforlagets *Engelsk Stor Ordbok*, og listen med oversettelse der de fantes ble forelagt informanten. Vedkommende ble bedt om å oversette ordene selv, og i de tilfellene oversettelse allerede fantes enten beholde ordbokens forslag, eller erstatte det med egen oversettelse, eventuelt tilføye denne. Dette fører til at enkelte av ordene hadde flere mulige oversettelser i fasiten. Selv om listen ikke er tolket som en uttømmende liste av alt et norsk kompositum kan oversettes med, ble det likevel vurdert som bedre enn å tvinge informanten til å velge en bestemt uttrykksmåte i samtlige tilfeller. Dette er også i tråd med diskusjonen av komposita fra en teoretisk synsvinkel i kapittel 3. At disse oversettelsene senere brukes til å trene opp

maskinlæringsmotorer understreker dette poenget, fordi en forutsetning om at en oversettelse alltid er den beste ville kunne gi modellen slagside mot en bestemt konstruksjonstype.

Av 750 gullstandardoversettelser var 185 oppført med flere oversettelser, hvorav 81 fra fra det høye frekvensbåndet, 50 fra det mellomste og 54 fra det lave. I gjennomsnitt hadde hvert norske kompositum 1,3 oversettelser.

For noen ord kunne ikke informanten finne passende oversettelser, selv om enkelte av ordene med rimelighet kunne forstås uten sammenheng av andre. Likevel ble slike ord som kom i retur uten oversettelse byttet ut med nye, tilfeldig valgte, uten å gå i dialog informanten om årsakene. Faren for å at en slik dialog kunne bli oppleves ledende ble det vurdert som bedre å bytte dem ut for å få et bilde av den intuitive språkfølelsen der det var mulig. Eksempler på slike ord var *lege·mann*, *kommando·aksjon* og *dusin·tall*. Ikke-oversatte komposita fra informanten avdekket også flere konstruksjoner som systematisk hadde gått igjennom den automatiske silingen beskrevet over. Eksempelvis *løp·erske*, *italiener·inne* og *embryologi·ske* blir godkjent som komposita bestående av to substantiv av sammensetningsanalysatoren fordi at *ske* og *erske* oppfattes som substantiv. Disse står ikke i Bokmålsordboka, og er heller å betrakte som i økende grad historiske bøyningsmorfemer som indikerer at substantivet er av hunkjønn. Tilsammen 82 av 750, eller 11% av ordene ble forkastet av ovenforbeskrevne årsaker. Til sammenlikning byttet Baldwin og Tanaka (2004) ut 6,6% av sine engelske komposita etter manuell inspeksjon.

Oversettelsene ble så gjennomgått, og sortert etter *word alignment*, som diskutert i avsnitt 2.4.3. For de norske kompositaene som hadde en komposisjonell oversettelse, hvor det var tydelig hvilket delledd som pekte på hvilken engelske komponent, og hvilke syntaktiske transisjoner de eventuelt måtte gå igjennom fra sitt substantiviske utgangspunkt, så ble en mal laget for å dokumentere disse endringene. Malene som ble identifisert er vist i tabell 5.1.

Mal	Norsk eksempel	GS-oversettelse
E_1-E_2	*glas·skjerm*	*glass panel*
ADJ(E_1)-E_2	*ansikts·hår*	*facial hair*
E_2-E_1	*gartner·mester*	*master gardener*
E_1-E_2-PL	*lin·bukse*	*linen pants*
E_1-POSS(E_2)	*dags·lønn*	*day's salary*
PL(E_1)-POSS(E_2)	*kvinne·avis*	*women's newspaper*
E_1E_1-E_2	*bydels·senter*	*city district center*
E_2-E_2E_1	*mini·kraftverk*	*mini power plant*
E_2-*of*-E_1	*due·flokk*	*flock of pigeons*
E_2-*between*-PL(E_1)	*bande·krig*	*war between gangs*
E_2-*for*-E_1	*arbeids·utstyr*	*equipment for work*
E_2-*to*-E_1	*forståelses·evne*	*ability to understand*
E_2-*in*-E_1	*jakt·lykke*	*luck in hunting*
E_2-*with*-E_1	*natur·sammenheng*	*connection with nature*
E_2-*of*-*the*-E_1	*hale·spiss*	*tip of the tail*
E_2-*of*-*the*-E_1-PL	*muskel·atrofi*	*wasting of the muscles*
E_2-*of*-*a*E_1	*dronning·krone*	*crown of a queen*
E_2-*on*-*the*E_1	*miljø·syn*	*view on the environment*
E_2-*from*-*a*-E_1	*jente·hår*	*hair from a girl*
E_2-*of*-*the*-E_1E_1	*finger·knips*	*snap of the fingers*

Tabell 5.1: *Identifiserte engelske konstruksjonstyper fra utvalgte norske komposita. N_1 og N_2 betegner henholdsvis første og andre ledd i det norske kompositum til oversettelse, og E_1 og E_2 betegner den engelske oversettelsen av disse. Merknaden* POSS *betyr at substantivet står i possesiv form (med apostrof og genitivs -s-, og* PL *betegner at substantivet i oversettelsen skal stå i flertall.* ADJ *betyr at et adjektiv som tilsvarer substantivet skal brukes. GS står for gullstandard.*

5.3.7 Rangeringsprosessen

Rangeringsprosessen ble implementert ved at hver mal ble definert som en struktur, som hadde informasjon om hvordan oversettelseskandidater skal genereres, og hvordan korpusdata skal hentes inn og hvordan de skulle evalueres mot en gullstandard. På den måten kunne det ved hjelp av to løkker velges ut den oversettelseskandidaten som var høyest rangert. For bygging av MaksEnt-modellene ble de korpusdataene skrevet ut (trekkene ved hver kandidat) til fil i TADMs format, som vil bli vist under.

Hver mal ble implementert i Lisp som en *struct*, slik at hver mal ble navet i et hjul som hadde tilgang til tre hashtabeller, som inneholder de indekserte oversiktene over hvor ofte hver konstruksjonstype representert ved en mal opptrer. Det er behov for tre slike, for å kunne hente ut henholdsvis hvor ofte hver konstruksjonstype opptrer i sin helhet i korpuset, hvor ofte forleddet står som forledd i slike konstruksjonstyper, og tilsvarende for etterleddet. Videre har hver mal en unik id, et filnavn som danner roten for innlasting av hashtabeller (de navngis med denne roten med tallene 1,2,3 som suffiks), et oppbygningsfelt som brukes til å gjenopprette overflateformen til hver konstruksjonstype for å vurdere den mot gullstandarden. Tilhørende hver mal er også en generator som er en funksjon som danner oversettelseskandidater i samsvar med hver konstruksjonstype malen representerer.

Initialiseringen av mal-strukturene foregikk ved at den nødvendige informasjonen om hver mal ble lagret i en tabell. Etter at instansene var opprettet kunne hashtabellene fylles fra filer hadde en rot definert i tabellen, og faste suffikser som identifiserte hver av de tre tabellene.

Prosessen er illustrert i figur 5.3.7.

5.3.8 Bygging av MaksEnt-modell

Innhenting av trekk til MaksEnt-modellen blir gjort etter mønster av metoden som ble brukt for rangeringsmetodene vist over, men i stedet for at en rutine velger ut den oversettelseskandidaten som har høyest rangering og sjekker den mot gullstandarden, så blir verdier for hvert

```
STRUKTUR mal (ID Filnavn Oppbygning
                Hashtabell-1 Hashtabell-2
                Hashtabell-3 Generator)

(.. initier maler fra tabell med verdier..)

PROSEDYRE ranger(mal,komp)
        FOR HVER mal I mal-liste:
                gjeldende-mal =
                        FOR HVER o-kand I mal->Generator(komp):
                                HVIS SCORE(o-kand) > valgt-for-mal
                                valgt-for-mal = o-kand
                        TIL SLUTT RETURNER valgt-for-mal
                HVIS gjeldende-mal > valgt-mal
                        valgt-mal = gjeldende-mal
        TIL SLUTT RETURNER valgt-mal
SLUTT PROSEDYRE
```

Figur 5.5: *Rangeringsprosessen illustrert med pseudo-kode.*

trekk hentet ut og nummerert i henhold til det påkrevde formatet fra TADM.

For hvert kompositum til oversettelse ble en *context* (norsk: *kontekst*) etablert, med en linje for hver oversettelseskandidat, kalt en *event* (norsk: *hendelse*) i TADM-terminologi. Et eksempel på en slik kontekst er vist under:

```
4
0 3 0 1.9316431e-7 1 2.0794415 2 7.816417
1 3 0 2.3011644e-9 1 0.0 2 7.816417
0 3 0 2.4155268e-8 1 0.0 2 10.167504
0 3 0 9.779341e-8 1 1.3862944 2 7.9476786
2
....
```

Tallet som står alene på en linje markerer starten på en ny kontekst, og forteller om hvor mange hendelser som opptrer i den. Det første tallet i hver hendelse forteller om dette var en ønsket hendelse eller ikke (i tilfelle oversettelse om den sto i gullstandarden). Dernest følger hvert trekk nummerert fra 0 og oppover, og trekkets verdi som kan være et reelt tall. I mange MaksEnt-prosjekter brukes binære trekk, men fordi det

brukes ekte verdier (og logaritmiske) i eksperimentene omtalt i kapittel 6, vises slike verdier.

Byggingen av modellen gjøres med å eksekvere TADM med ønskede parametre, som for så små modeller som dette, med opp til 35 trekk er en lite kostbar affære, bygging av en modell tok omlag 30 sekunder på en rask maskin (som beskrives under).

5.3.9 Evaluering

I forbindelse med fininnstilling av parametre og 10-dobbel kryssvalidering må mange modeller bygges, som naturligvis tar lenger tid, men er fortsatt greit håndterbart i forhold til MaksEnt-modeller i andre NLP-anvendelser hvor antall trekk kan være flere tusen. Evaluering er allerede vist ovenfor for rangeringsmetodene som kunne beregnes direkte ut fra hashtabeller som var lastet inn i Lisp. For MaksEnt-modellene er bildet noe mer komplisert, spesielt for beregningen av sølvstandard-ytelse, som er en manuell prosess som blir omtalt under.

10-dobbel kryssvalidering

I K-dobbel kryssvalidering, så deles utgangspunktet opp i K like deler, og for hver kjøring av systemet, så holdes en K ut for validering, mens de resterende K-1 delene brukes som treningsdata. Operasjonen kjøres K ganger, slik at samtlige deler har vært gjenstand for validering. Gjennomsnittet av resultatet blir så innhentet. Det er ønskelig å unngå å trene og teste på samme data, fordi dette kan gi en urimelig høy skår, ved at modellen kan være overtilpasset til treningsdataene. I slike tilfeller kan modellen være godt i stand til å reproduserer rangeringen av treningsdataene, men liten grad i stand til å gjøre gode prediksjoner på usette data.

Måten dette ble gjort på var at hver fil med trekk med formatet som beskrevet over ble delt opp i 10 like deler, og 10 MaksEnt-modeller ble bygget og evaluert som beskrevet over. TADMs egen evaluator som virker direkte på trekkfiler ble brukt til å vurdere hvor mange prosent som er

riktige. Den sjekker om hver trekkvektor multiplisert med sine tilsvarende vekter fra modellen som er den høyest rangerte innenfor hver kontekst også er markert med en 1 foran hendelses-linjen i trekkfilen.

Dette ble gjort ved hjelp av et bash-script som kunne holde n prosesser i gang i bakgrunnen samtidig. Scriptet setter i gang så mange prosesser som det får beskjed om samtidig, og sjekker ved gitte intervaller om prosessene er i live eller ikke. Forsøkene ble utført på en maskin med 8 parallelle Opteron-prosessorer, og 32GB RAM, og enkle forsøk viste at en 10-dobbel kryssvalidering som tok 2m50 (170 sekunder) når alle treninger og evalueringer av modellene foregikk serielt, behøvde 28 sekunder når det ble gjort med 8 parallelle prosesser med skriptet beskrevet ovenfor, som representerer en betydelig effektivitetsgevinst. Dette var spesielt nyttig under eksperimenter med glattingsvariabler, hvor samme 10-doble kryssvalidering måtte utføres for $100 \cdot 50$ kombinasjoner av variable, 50 000 eksekveringer av TADM. I forbindelse med utviklingen av modellene, så kortet det ned ventetiden etter en endring var gjort eksempelvis i måten trekkene ble ekstrahert på, og det er vanskelig å sette tall på denne besparelse.

Sølvstandardevaluering

Som vist ovenfor kan TADMs egen evaluator brukes til å validere en trekkfil ut ifra en gitt modell, men det er ikke mulig å se fra trekkfilen hvilket forslag som blir rangert som det høyeste. For å evaluere etter sølvstandard er dette nødvendig, fordi en informant må vurdere om det er mulig å gjenvinne det norske utgangspunktet fra den høyest rangerte engelske oversettelseskandidaten. For å løse dette ble derfor TADMs evaluator reimplementert i Lisp, slik at en MaksEnt-modell kunne leses inn, og brukes til å evaluere trekkene generert for hver oversettelseskandidat, som da ble skrevet ut sammen med det norske kompositum som var dets utgangspunkt. I tråd med den 10-doble kryssvalideringen beskrevet i forrige avsnitt, ble de MaksEnt-modellene som var bygget uten de 10 evalueringsbolkene brukt til å

generere kandidater for disse. Der MaksEnt-modellene ble brukt til å lage kandidater for de kompositaene hvis gullstandard-oversettelse ikke lot seg generere dynamisk ble en modell tilfeldig trukket ut (den første av de 10 for alle frekvensbånd).

Resultatene av disse gjennomkjøringene ble så skrevet ut og satt sammen. Deretter ble de 250 opprinnelige kompositaene delt inn i tre. Først de som allerede var rangert riktig etter gullstandarden, dernest de som var «mulige» å finne rett oversettelse til for modellen, der den valgte feil, og til sist de overskytende kompositaene hvor den ønskede oversettelsen ikke var en del av mulig å finne frem til. Disse to siste listene ble forelagt en informant (den samme som utarbeidet gullstandarden), for å vurdere om det var mulig. En vurdering tilsvarende den som ble gjort i (Baldwin & Tanaka, 2004) ble brukt, som såvidt diskutert i avsnitt 2.4.3. Kriteriene var forsøksvis de samme som i disse forsøkene, og informanten ble presentert kriteriene i sin opprinnelige form. Selv om utarbeidelse av gullstandarden også er subjektiv, fordi det bare er én persom utarbeider den, er faren enda større for sølvstandarden, fordi at kravene til å rangere noe som *gjenfinnbart* er flere og løse, slik at faren for at flere ord blir godtatt hvis informanten skulle ønske at en av rangeringsmetodene skulle være bedre enn den andre er større. For gullstandarden er jo fasiten den samme mellom alle metoder, slik at denne faren ikke er der. I likhet med hos Baldwin og Tanaka (2004) fikk ikke informanten vite hvilken metode som var brukt for å komme frem til forslagene, for å redusere denne risikoen. Men på grunn av disse risikomomentene er det betimelig å minne om at sølvstandardresultatene må leses for det de er.

Det overhengende konseptet var «gjenfinnbarhet», og syntaksen trenger ikke å være riktig, slik at en frase i plural vil være gjenfinnbar selv om fasiten sier entall, men den skal fange de grunnleggende semantiske egenskapene ved ordet. Et eksempel på en oversettelseskandidat som ble vurdert som gjenfinnbar er *peaseat uprisings* som oversettelse av *bonde·reisning* (i gullstandarden stod *peasant revolt*), mens *cloth countries* ikke vurdertes som en gjenfinnbar oversettelse av *fille·land* (i gullstandarden sto *stupid country*).

5.4 Hypotesetesting

Den såkalte «tegntesten» (engelsk: *sign test* er en binomisk test, som går ut på å beregne kumulative sannsynligheter for binomiske forsøksrekker. En binomisk rekke er en situasjon hvor

1. Hvert forsøk resulterer i enten i suksess eller fiasko.

2. Sannsynligheten for suksess er den samme i alle forsøk.

3. Forsøkene er uavhengige.

Sannsynligheten for x suksesser i løpet av n forsøk i en slik rekke er $\binom{n}{x}p^x(1-p)^{n-x}$, hvor p er sannsynligheten i hvert forsøk.

Hvis en evaluering av to metoder går ut på at de er like gode før og etter en variabel er forandret, så kan avvikene mellom forsøkene betraktes som en binomisk forsøksrekke. Dersom forsøk 1 og forsøk 2 avga forskjellig resultat n antall ganger, så burde det være like stor sannsynlighet for at forsøk 1 var bedre enn forsøk 2 som omvendt, hvis de er like gode. Med en antagelse om at metoden er like gode, så kan sannsynligheten for at avvikene ligger på den ene eller andre siden av hverandre betraktes som en binomisk forsøksrekke med n forsøk og $p = 0.5$.

Testen går ut på å beregne hvor stor sannsynligheten er for at et gitt antall resultater ligger på den ene eller andre siden av den andre metoden, gitt at det er like stor sannsynlighet for hvor de havner. Hvis k^+ er antallet avvik hvor metode 1 var bedre enn metode 2, og k^- er antallet avvik der metode 2 var bedre enn metode 1 og n det totale antall avik, så beregnes sannsynligheten for at enten et avvik kan ligge så høyt som det høyeste av k^+ og k^-, eller så lavt som det laveste, med en nullhypotese som går ut på at sannsynligheten for at de havner på den ene eller andre siden er 0.5. Det vil alltid være en viss sjanse for at samtlige forsøk havner på oversiden eller undersiden av samme forsøk med den andre metoden, men det denne minsker jo flere det er snakk om.

Dette beregnes gjennom kumulative binomiske sannsynligheter, som er en summering over formelen over. Hvis man går ut ifra at en mynt er rettferdig, og kaster den 10 ganger, og 7 av resultatene er kron, blir sannsynligheten for at 7 eller flere skal kron gitt at det er like stor sjanse for mynt og kron $\sum_{x=7}^{10} \binom{10}{x} 0.5^x (1-0.5)^{10-x} = 0.17$, og dermed er sannsynligheten for at 3 eller færre er mynt $\sum_{x=0}^{3} \binom{10}{x} 0.5^x (1-0.5)^{10-x} = 0.17$. For å gjennomføre en tosidig test, beregnes summen av disse verdiene, eller en dobling av en av dem. Nullhypotesen forkastes til fordel for en alternativ hypotese som sier at mynten er urettferdig, hvis sannsynligheten for de empiriske resultatene er tilstrekkelig lav, kalt p-verdien til testen.

Kapittel 6

Eksperimenter

I dette kapittelet vil de gjennomførte eksperimentene med automatisk oversettelse av komposita bli omtalt. Eksperimentene ble utført langs tre dimensjoner, rangeringsmetode, analysedybde for korpuset som ble brukt til å danne grunnlag for rangering samt størrelsen på dette.

Med metodene som ble beskrevet i avsnitt 5, ble det gjennomført eksperimenter i tre omganger. Først ble de norske komposita holdt ut, hvor det var mulig å komme frem til en oversettelse som stod i gullstandarden. Hvis det ikke er mulig å komme frem til denne konstruksjonen, enten fordi at et av delleddene ikke står i ordboken, eller at den ønskede oversettelsen ikke står der, så har ingen av rangeringsmetodene muligheten til å komme frem til riktig svar. Listen av komposita som lot seg oversette *komposisjonelt* for hvert frekvensbånd ble så rangert på bakgrunn av bruk av ett korpus, to korpora og tre korpora, for å se om ytelsen ble forbedret ettersom datagrunnlaget vokste. I denne sammenhengen betyr at de kunne oversettes komposisjonelt at det var teknisk mulig å sette sammen et riktig svar etter gullstandarden med den ordboken og gullstandarden som ble brukt, og ikke som et uttrykk for at disse ordene er *komposisjonelle* i forhold til diskusjonen av komposisjonalitet i avsnitt 3.2.

De tre korporaene det er snakk om er British National Corpus (BNC), Aquaint-korpuset (AQ), og The North American News Text Corpus

(NAN), som er nærmere beskrevet i avsnitt 4.2.2. Indeksering av de tre foreligger i filer etter en rangering som beskrevet i avsnitt 5.3.3. Resultatet var to sett av indeksfiler for hver av korporaene, hvor RASP blir brukt som henholdsvis tagger og parser. Når tabellene ble lastet inn i systemet ble de lagt oppå hverandre, slik at de omtalte korpora blir lagt til det forrige i rekkefølgen som blir presentert, vist med en + i tabellene under.

6.1 Rangeringsmetoder

Eksperimentene tok utgangspunkt i (Baldwin & Tanaka, 2004), slik at å måle CTQ-rangering mot en maskinlæringsbasert rangering var hovedformålet. Men fordi oppgaven ønsket å undersøke hvordan analysedybde påvirket resultatet, måtte to forsøk gjøres for hver rangering, hvor resultatene fra RASP ble brukt som *parser* eller *tagger*, hvor i det siste tilfellet utelukkende den ordklasse som ble tilskrevet hvert ord ble lagt til grunn for å identifisere en konstruksjonstype. Se avsnitt 5.3.3 for videre omtale av dette.

6.1.1 Heuristiske metoder

Fordi det var tvil rundt hva som ble brukt som nevner i utregningen av CTQ, som beskrevet i avsnitt 2.4.3, ble begge metoder forsøkt. Verdiene for α og β var 0.9 og 0.1 i alle eksperimenter. CTQ utregnet på bakgrunn av en en CTQ som er betinget av den aktuelle malen er vist som CTQ-B i tabellene under, og den reviderte utgaven som regner ut sannsynligheten for at hver konstruksjonstype opptrer i forhold til samtlige konstruksjonstyper kalles CTQ-S.

Videre ble en enkel rangering basert utelukkende på hvor mange ganger hver konstruksjonstype opptrådte i det analyserte korpuset i sin fulle form brukt som referanse, kalt REF i tabellene, i tillegg til en metode kalt BIDIR, som er det samme som CTQ-S, men i tillegg med enkel bruk av tospråklig informasjon. Dette ble gjort slik at antall ganger hele oversettelseskandidaten sto oppført som oversettelse av kompositumet

som skulle oversettes ble lagt til CTQ-S-summen. I enkelte tilfeller kan en ordbokoppslag ha flere betydninger, og samme oversettelse kan stå oppført flere ganger.

6.1.2 Maskinlæringsmetoder

Ved siden av de fire heuristiske metodene ble også to forskjellige MaksEnt-modeller bygget, og brukt til rangering. Den ene, MaksEnt$_{korp}$ baserer seg på korpusbaserte og mal-trekk, mens MaksEnt$_{full}$ bruker tospråklige trekkene i tillegg.

6.2 Trekk

Valg av trekk til å bygge MaksEnt-modellen ble foretatt etter mønster av dem som ble brukt av Baldwin og Tanaka (2004), vist i avsnitt 2.4.3, men de måtte endres noe som en følge av at forutsetningene for prosjektet var noe annerledes. Eksempelvis var det bare en ordbok tilgjengelig, slik at summen av oversettelsene av et norsk ord til engelsk alltid ville bli én. Videre ville det blitt vanskelig å fange opp ledd-til-ledd-korrspondanse, altså hvor mange ganger forleddet i *fjøs·katt*, *fjøs*, ble oversatt til *barn*, forleddet i gullstandardoversettelsen *barn cat*, og tilsvarende for etterleddet. Dette ville ha forutsatt en analyse av samtlige oppslag i ordboken med sammensetningsanalysatoren, og på grunn av utfordringene med den omtalt i avsnitt 5.3.3 måtte disse ha blitt verifisert manuelt før slik statistikk kunne innhentes.

6.2.1 Korpusbaserte trekk

Korpustrekkene som ble brukt var de samme variablene som ble brukt til å regne ut CTQ-B og CTQ-S, og er vist i tabell 6.2.1. Med unntak av CTQ-S ble logaritmen av korpusfrekvensene brukt. Opprinnelig ble verdiene brukt direkte, men dette resulterte i numeriske problemer under eksekvering av TADM ved bygging av MaksEnt-modellene. Ved å istedet

Trekk	Beskrivelse
CTQ-S	CTQ-S regnet ut med alle konstruksjoner som nevner.
log(frek(E_1,E_2,t))	Frekvens av konstruksjonen i sin fulle form.
log(frek(E_1,_,t))	Frekvens av første engelske ord, som første del.
log(frek(_,E_2,t))	Frekvens av andre engelske ord, som andre del.
log(frek(E_1,t))	Frekvens av første engelske ord, uansett plassering.
log(frek(E_2,t))	Frekvens av andre engelske ord, uansett plassering.

Tabell 6.1: *Korpusbaserte trekk. E_1 og E_2 betegner her oversettelsen av henholdsvis forledd og etterledd av det norske kompositumet. Det kan bestå av to ord ved bruk av enkelte maler.*

bruke logaritmen ble spredningen mellom verdiene flatet ut, og TADM konvergerte. I tidlige forsøk ble også verdiene forsøkt erstattet av sin prosentvise andel av den høyeste observerte verdi for et trekk, men dette ble forlatt fordi logaritme-verdiene lot til å gi bedre resultater.

6.2.2 Tospråklige trekk

På grunn av bare en tilgjengelig ordbok, var det ikke var mulig å bruke akkurat de samme trekkene som i (Baldwin & Tanaka, 2004), så ble det forsøkt noen andre varianter, hvor ordboken ble brukt i begge retninger. De tospråklige trekkene som ble brukt er gjengitt i tabell 6.2.2.

Den kanskje mest iøynefallende fordelen med å bruke tospråklige trekk, er at informasjon om at den engelske oversettelseskandidaten eventuelt står oppført som en oversettelse av det norske kompositumet blir tilgjengelig, som er en klar indikasjon på at en riktig oversettelse er funnet. Summeringen over hvor mange ganger en oversettelse av forledd og etterledd sto oppført kommer fra at hvert ordbokoppslag kan ha flere betydninger av ordet. Eksempelvis kan *hus* ha 6 forskjellige betydninger, og *house* står oppført som en oversettelse i 5 av dem. Dermed er det en indikasjon på at *house* er en vanligere oversettelse av ordet enn *firm* som sto oppført en gang.

Trekk	Beskrivelse
$frek(E_1, E2\|N_1, N2)$	Frekvens av at oversettelseskandidaten i sin fulle form som oversettelse av det norske kompositumet.
$frek(N_1, N2\|E_1, E2)$	Frekvens av det norske kompositumet som oversettelse av oversettelseskandidaten.
$frek(E_1, E2, \text{n-e})$	Frekevens av oversettelseskandidaten i ordboken med retning norsk-engelsk.
$frek(E_1, E2, \text{e-n})$	Frekevens av oversettelseskandidaten i ordboken med retning engelsk-norsk.
$frek(E_1\|N_1)$	Frekvens av at det første engelske ordet var en oversettelse av forleddet.
$frek(E_2\|N_2)$	Frekvens av at det andre engelske ordet var en oversettelse av etterleddet.
$frek(N_1\|E_1)$	Frekvens av at forleddet var en oversettelse av det første engelske ordet.
$frek(N_2\|E_2)$	Frekvens av at etterleddet var en oversettelse av det andre engelske ordet.

Tabell 6.2: *Tospråklige trekk. N_1 og N_2 betegner forledd og etterledd i det norske kompositumet, og E_1 og E_2 betegner her oversettelsen av henholdsvis forledd og etterledd av det norske kompositumet. Det kan bestå av to ord ved bruk av enkelte maler.*

6.2.3 Mal-trekk

I tillegg ble det lagt et binært trekk for å identifisere hver mal. Trekket hadde verdien 1 dersom malen ble brukt til å lage den aktuelle oversettelseskandidaten og 0 ellers. Dette er fordi at malene ikke fordeler seg likt, den desidert vanligste malen i gullstandarden var E_1-E_2, to substantiver ved siden av hverandre uten noe imellom, som i *print shop*. Ved å gi modellen informasjon om hvilken mal som er brukt for oversettelseskandidatene med oppføring i gullstandarden, kunne den gi preferanse til å velge kandidater fra den vanligste malen.

6.2.4 Eksempler

Kompositumet *konge·krabbe* ble delt opp i *konge* med oversettelsene *king, king, king, boss, king, king* og *King* og *krabbe* med oversettelsene *true crab, common crab, edible crab* og *crab crocket*. Tilsammen ble det generert 125 oversettelseskandidater til engelsk ut ifra disse deloversettelsene og maler.

Trekkvektorene til bruk i TADM for to av kandidatene, *king crab* og *boss crab* blir vist i det følgende. Den første linjen forteller TADM at linjen representerer en korrekt oversettelse (1) eller en gal (0), og tallet 35 forteller at 35 trekk følger. Videre følger trekkene med stigende nummerering og verdi.

```
king crab:
1 35
0 4.3037136e-7 1 2.8903718 2 5.5012584 3 4.9767337 4 7.278629 5 5.771441
6 1.0 7 1.0 8 1.0 9 4.0 10 5.0 11 2.0 12 3.0 13 3.0
14 1.0 15 0.0 16 0.0 17 0.0 (....) 32 0.0 33 0.0 34 0.0

boss crab:
0 35
0 6.5432846e-12 1 0.0 2 5.273 3 4.9767337 4 7.771067 5 5.771441
6 0.0 7 0.0 8 0.0 9 0.0 10 1.0 11 0.0 12 3.0 13 3.0
14 1.0 15 0.0 16 0.0 17 0.0 (....) 32 0.0 33 0.0 34 0.0
```

I den første linjen står de korpusbaserte trekkene, med CTQ-S etterfulgt av logaritmen av de øvrige korpusdataene. I andre linje står de tospråklige trekkene, og i siste linje de binære mal-trekkene, som er helt likt for de to kandidatene, begge generert av den samme malen. Trekk 1 viser at *king crab* ble funnet i korporaene $exp(2.8903718) = 18$ ganger, i

Metode	REF		CTQ-B		CTQ-S	
Korpora	*Tagger*	*Parser*	*Tagger*	*Parser*	*Tagger*	*Parser*
BNC	19,77	18,20	16,18	19,77	25,62	29,65
+AQ	24,50	23,59	15,28	18,88	29,66	32,12
+NAN	24,50	24,55	15,96	19,37	27,42	27,70

Tabell 6.3: *Oppsummering av gullstandardresultater for REF, CTQ-B og CTQ-S, i prosent av komposita som blir oversatt med en oversettelse som stod i gullstandarden.*

forhold til *boss crab* som ikke ble funnet i det hele tatt. Modellen som ble bygget når *konge·krabbe* ble holdt utenom klarte å rangere dette *king crab* øverst.

6.3 Resultater

Baldwin og Tanaka (2004) gjorde forsøk på fjernt beslektede språk, japansk og engelsk, og for å vurdere rangeringsmetodes egnethet over lingvistiske grenser. Eksperimentene i denne oppgaven ønsket å gi et bilde av hvordan liknende teknikker virker også for nær beslektede språk. Ved siden av dette forskningsspørsmålet var det et mål for oppgaven å prøve ut hvilke utslag sammensetning og analyse av grunnlagskorpora gjorde på resultatene. Resultatene presenteres med disse forskningsspørsmålene som utgangspunkt.

En oppsummering av resultatene for rangeringsmetodene målt mot gullstandarden finnes i tabellene 6.3 og 6.4. Komplette resultater for samtlige eksperimenter finnes i appendiks A på side 126. Tallene er gjennomsnittsresultater for de tre frekvensbåndene innenfor hvert korpus. Det går frem av resultatene at det er en gradvis økning i nøyaktighet fra den enkleste rangeringsmetoden, REF, til MaksEnt-modellen som benytter seg av samtlige trekk. I lys av at bare 1 (0,6%) av kompositaene i det lave frekvensbåndet, 7 (4,5%) i det mellomste og 10 (7,5%) fra det høye sto oppført med oppslag i ordboken, er økningen i nøyaktighet påtagelig. Det er vanskelig å sammelikne resultatene direkte med dem fra (Baldwin

Metode	BIDIR		$\mathbf{ME}_{korp}$		$\mathbf{ME}_{full}$	
Korpora	*Tagger*	*Parser*	*Tagger*	*Parser*	*Tagger*	*Parser*
BNC	27,19	30,55	34,81	35,42	49,3	50,49
+AQ	30,78	33,70	37,9	40,5	51,31	53,18
+NAN	28,31	29,28	37,28	38,54	52,51	52,03

Tabell 6.4: *Oppsummering av gullstandardresultater for BIDIR og MaksEnt-metodene, i prosent av komposita som blir oversatt med en oversettelse som stod i gullstandarden. Evaluering for MaksEnt-eksperimentene er foretatt ved hjelp av 10-dobbel kryssevaluering.*

& Tanaka, 2004) gjengitt i tabell 2.3, til det er premissene for ulike, men de minner om hverandre på den måten at både CTQ-rangering og maskingslæringsteknikkene gir en vesentlig ytelsesforbedring fra referanseforsøkene.

6.3.1 Rangeringsmetoder

REF-rangeringen er ikke direkte sammenliknbare med forsøkene til Moa (2006) og Grefenstette (1999), fordi den baserer seg på hvor ofte en oversettelseskandidat forekom som en konstruksjonstype representert ved malene brukt i forsøkene, men de er likevel beslektet, fordi at den bare baserer seg på hvor ofte oversettelseskandidaten opptrer i sin fulle form. Selv om korpusgrunnlaget er stort når alle tre er lastet inn, er det ikke sammenliknbart med WWW i størrelse, og er dessuten prosessert gjennom RASP og indeksert etter maler. Likevel viser resultatene at metoden gir en forbedring fremfor å bare holde seg til en ordbok.

Videre går det frem at CTQ-S, hvor flere korpusdata brukes i en lingvistisk motivert kombinasjon gir bedre resultater enn referanseforsøket. Således gir eksperimentene støtte til (Baldwin & Tanaka, 2004) som også fant at en slik rangering gir bedre resultater enn referansen. CTQ-S gir gjennomgående også bedre resultater enn CTQ-B, som kan tyde på at det er denne varianten av CTQ som ble brukt i den siste formuleringen av CTQ. Se avsnitt 2.4 for en diskusjon av CTQs utvikling.

Forsøket BIDIR viser at også CTQ-rangeringen får en liten forbedring

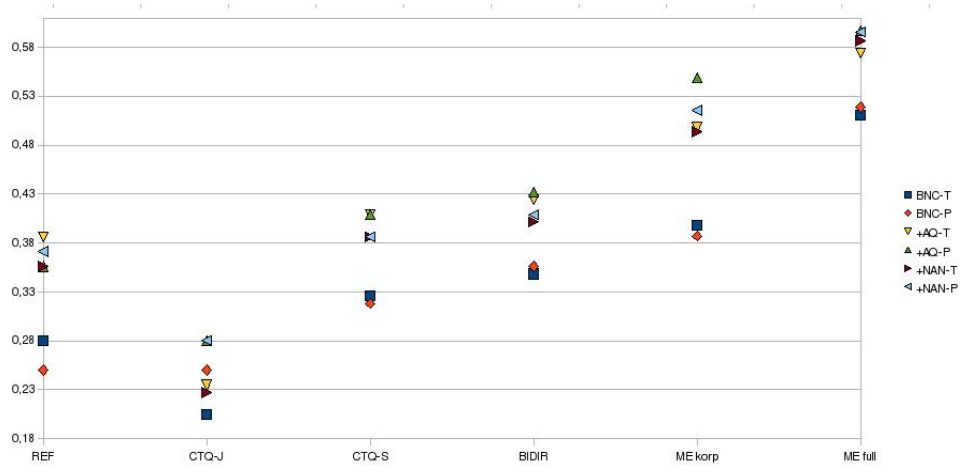
(fra 32,12% til 33,70%) dersom oppslag av oversettelseskandidatene i ordboken ble tatt med. Kontrasten til forbedringen mellom ME_{korp} og MaksEnt_{full} er likevel stor, der er forbedringen gikk fra 40,5% til 53,2%.. Dette lar seg naturlig forklare i det at ikke bare det rene ordbokoppslaget fra BIDIR var tilgjengelig for ME_{full}, men hele spekteret av tospråklige trekk diskutert over. Men det viser at også den heuristiske CTQ-metoden fikk en forbedring fra tospråklig informasjon, selv om måten denne informasjonen ble brukt i formelen kunne vært utvidet og raffinert for å oppnå ytterligere forbedringer.

Det er ikke mulig å lese ut av nøyaktigheten til MaksEnt-modellene hvor bra denne teknikken er i forhold til SVM-en brukt i (Baldwin & Tanaka, 2004), eller endatil hvordan problemet med oversettelse mellom norsk og engelsk skiller seg fra japansk til engelsk, men når omlag halvparten av kompositaene ble rangert riktig sett i forhold til det lave antall ord som står oppført i ordboken, er det rimelig å peke på en praktisk nytte av teknikkene. Forbedringen fra de heuristiske teknikkene til maskinlæringsbaserte metoder viser en økning i nøyaktighet fra 32,1% til 40,5% ved bruk av bare korpusinformasjon, og fra 33,7% til 53,2% når samtlige trekk ble brukt, for eksperimentene av type parser. Det må her minnes om at den heuristiske rangeringen gjorde bruk av mindre tospråklige data enn MaksEnt-modellen, slik at det antagelig er et forbedringsrom her.

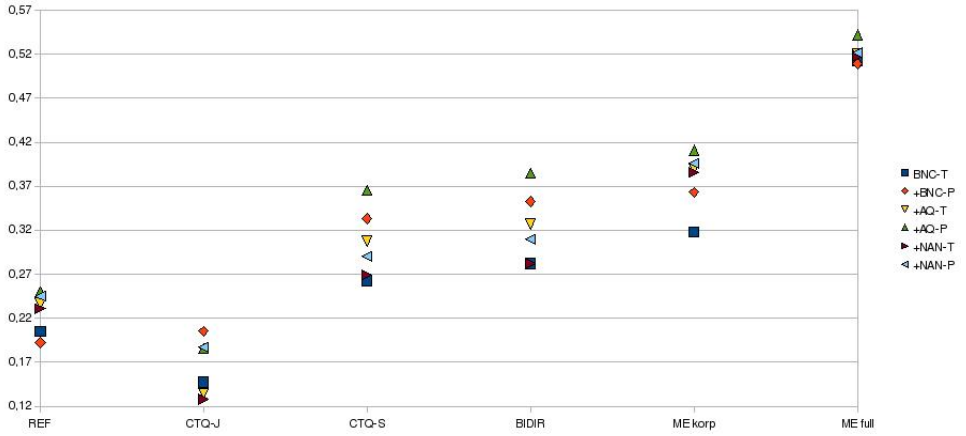
6.3.2 Analysedybde

I tabell 6.5 vises forskjellen i nøyaktighet mellom RASP brukt som parser og tagger for de samme rangeringsmetodene som over, og i figur 6.1 vises samtlige 72 gjennomførte eksperimenter grafisk, sortert etter rangeringsmetode med angivelse av korpusstørrelse og analysedybde (parser eller tagger). Forskjellen er liten, men gjennomgående er parser-resultatene de beste.

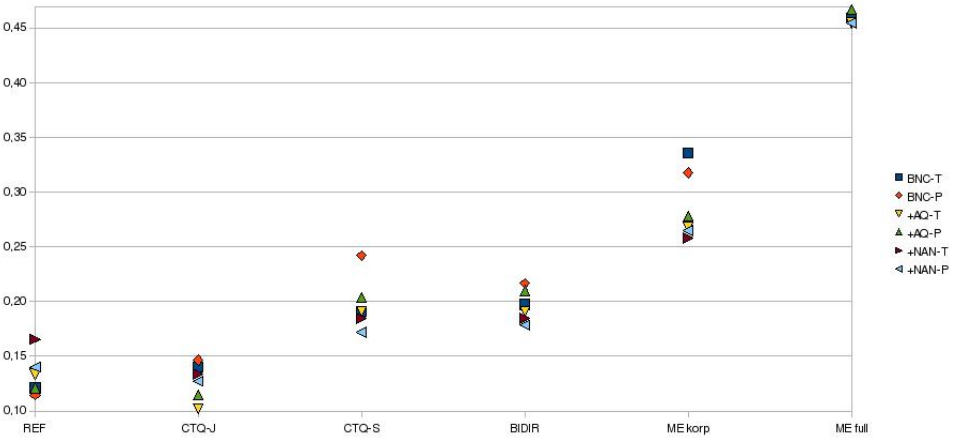
At forskjellene er små mellom å bruke analysen fra RASP som henholdsvis parser og tagger er ikke uventet fordi en dypere analyse (se



(a) Spredning av resultater for høyt frekvensbånd.



(b) Spredning av resultater for middels frekvensbånd.



(c) Spredning av resultater for lavt frekvensbånd.

Figur 6.1: *Plotting av resultater over 3 frekvensbånd. P og T etter korpusidentifikasjon indikerer om RASP ble brukt som parser eller tagger i korpusgrunnlaget.*

Korpus	REF	CTQ-B	CTQ-S	BIDIR	ME_{korpus}	ME_{full}
BNC	-1,65	3,65	3,80	3,24	0,53	1,17
+AQ	-1,01	3,65	2,35	2,81	2,73	1,9
+NAN	0,14	3,52	0,28	0,96	1,3	-0,4

Tabell 6.5: *Differanse i gjennomsnittlig prosentvis nøyaktighet over korpora mellom når RASP er brukt som henholdsvis parser og tagger. Et negativt tall viser at tagger-baserte treningsdata var bedre.*

avsnitt 4.5) er bakgrunnen for hvilken ordklasse enkeltordene tildeles. Dermed vil ikke metodene skille seg i tilfeller der engelske ord kan være både substantiv og verb, som *read*, (norsk: *(å) lese* eller *lesning, noe som blir lest*), *book* (norsk: *(å) booke* eller *bok*), *shop* (norsk: *å shoppe* eller *butikk*), men vil vise tilfeller utelukkende i de tilfeller et nytt ord av ønsket ordklasse (substantiv, eller preposisjon/artikkel) mellom syntaktiske grenser, som ved toverdige verb som *give* (norsk: *gi*) og *bring* (norsk: *bringe*). Eksempelvis er ikke *neighbor cake* (norsk: *nabo kake*) et kompositum i setningen *I brought my neighbor cake* (norsk: *Jeg bragte min nabo kake*), som fanges opp dersom det tas hensyn til at parseren skiller mellom de to nominalfrasene *neighbor* og *cake*, men ikke dersom man utelukkende ser på at det er to substantiv som følger hverandre.

6.3.3 Korpusstørrelse

Resultatene viser en gjennomgående forbedring fra ett til to korpora (BNC til BNC+AQ), for alle metoder med unntak av CTQ-B, mens ytelsen viser en tilbakegang i ytelse fra to til tre korpora (NAN blir tillagt) for metodene CTQ-S, BIDIR, ME_{korp} og ME_{full} (med parser-data). En svakhet ved korpusbasert metodikk som den inneværende i forhold til WWW-baserte modeller som dem brukt av Grefenstette (1999) og Moa (2006) er at om korpuset er for lite eller snevert, vil en ønsket oversettelse kunne være fraværende, og systemet vil aldri få muligheten til å finne frem til den. Oppgaven ønsket å undersøke hvordan korpusstørrelse påvirker ytelse.

På grunn av ovenfornevnte problemstilling var det dermed mulig at at ytelsen ville øke ved å gå ett til to korpora, men ikke like forventet

at ytelsen skulle gå ned igjen når det tredje korpuset, NAN, ble lagt til i grunnlaget. Men det er heller ikke rimelig å forvente at økningen i ytelse vil øke like mye per n tillagte ord i korpusgrunnlaget, men heller at forbedringen ville bli mindre og mindre, og konvergere mot et nivå hvor potensialet i korpus-informasjonen var hentet ut. Som diskusjonen i kapittel 3 var innom, er det ikke nødvendigvis mulig å vite hva et kompositum betyr uten å ha kontroll over en rekke kontekstuelle variabler (som vær, vind og humør), slik at en gradvis økning mot 100% synes urealistisk. Det er også en langt større forskjell i økningsdifferansen i mellom en og to korpora enn den tilsvarende nedgangsdifferansen mellom to og tre, dersom det summeres over alle eksperimentene. Dermed er det mulig å se på dette som et tegn på at resultatene er i ferd med å konvergere rundt nivåene sett for +AQ og +NAN.

Statistisk signifikans

Selv om resultatene viste en nedgang i ytelse ved å gå fra 2 til 3 korpora for flere rangeringsmetoder, var nedgangen mindre for den beste, ME_{full}. Forskjellen ble sjekket for statistisk signifikans med tegn-testen beskrevet i avsnitt 5.4. De 6 forsøkene med samtlige korpora (+NAN) ble sammenliknet med de tilsvarende forsøkene med 2 korpora (+AQ). Ved å sammenlikne listen over hvilken kandidat som ble valgt for kompositalistene med hverandre, viste det seg at metodene hadde divergerende svar for 38 komposita, hvorav halvparten gikk i favør av hver metode. Fordi fremgangsmåten er helt lik, forskjellen i forsøkene er bare korpusgrunnlaget, er det mulig å slutte at forskjellen ikke er signifikant også uten utregning (at halvparten av de divergerende går i hver retning er trivielt godt i samsvar med at sjansen er 50% for at de går i hver retning).

Men dersom resultatene der samtlige korpora ble brukt (+NAN) sammenliknes med forsøkene der bare BNC-korpuset ble brukt, så var det 106 divergerende svar, hvorav 66 gikk i favør av +AQ, og 40 gikk i favør av BNC. Sannsynligheten for at dette skulle skje dersom fremgangsmåtene

	CTQ-S		$\mathbf{ME}_{full}$	
	Samtlige	Mulige	Samtlige	Mulige
BNC	0,59	0,75	0,66	0,83
+AQ	0,60	0,74	0,72	0,85
+NAN	0,62	0,77	0,7	0,83

Tabell 6.6: *Sølvstandard-skår for rangering etter MaksEnt$_{full}$ og CTQ-S*

var likeverdige, er 0,015, slik at fremgangen fra 1 til 3 korpora er statistisk signifikant med $\alpha = 0.02$.

Utregningen gjelder bare for rangeringsmetoden ME$_{full}$, slik at det er ikke mulig å projisere denne signifikanstesten til å gjelde alle forsøkene uten videre, men denne metoden ble valgt ut fordi den hadde høyest ytelse.

6.3.4 Sølvstandard

To av rangeringsmetodene ble også vurdert med sølvstandardevaluering, en løsere, mer subjektiv evalueringsmetrikk omtalt i avsnitt 5.3.9. CTQ-S og ME$_{full}$. Både listen over komposita som rangeringsmetodene hadde valgt ut en oversettelseskandidat som ikke står i gullstandarden, og de resterende blant de 250 opprinnelige ble gjort gjenstand for sølvstandardevaluering.

En oppsummering av resultatene er gjengitt i tabell 6.6. Inndelingen i *mulige* og *samtlige* representerer henholdsvis de kompositaene systemet hadde muligheten til å finne rett oversettelse, og andelen av samtlige kompositum tilfeldig trukket ut fra fra det norske samlekorpuset. Bildet som tegner seg her likner det fra gullstandevalueringen, hvor en svak økning i ytelsen anes når korpuset økes, samt at MaksEnt$_{full}$ har betydelig høyere nøyaktighet enn CTQ-S.

Som nevnt i avsnitt 5.3.9 er det grunn til å lese disse resultatene med en viss varsomhet, og derfor ble heller ikke slik evaluering gjennomført for samtlige 36 eksperimenter. Disse resultatene er jevnere enn fra gullstandardevalueringen, spesielt for kategorien *mulige* fordi at mange

av de oversettelseskandidatene hvor rangeringsmetodene bommet med grunnlag bare i BNC ble akseptert som gjenfinnbare, ofte som følge av syntaktiske feil.

Likevel er evalueringen tatt med, fordi det kan si noe om i hvilken grad en meningsbærende oversettelse kan finnes frem til. Det er mulig å tenke seg eksempelvis oversettelsessystemer hvor en slik oversettelse kan være bedre enn ingen ting, alternativt korrigeres med en språkmodell.

6.4 Glatting

Ulike måter å glatte sannsynlighetsfordelingene i MaksEnt-modeller ble diskutert i avsnitt 5.2.6. I disse eksperimentene ble dataene først tilpasset, ved at bare komposita som hadde flere enn 10 webtreff ble brukt. TADM har tre ulike parametere som kan brukes til glatting, en absolutt og en relativ toleranse for forbedringer for når parameterestimeringen skal stoppe å iterere, og muligheten til å definere en gaussiansk prior med varians σ^2. Den absolutte setter en grense for hvor stor forbedringen skal være mellom hver iterasjon før de skal stoppe, mens den relative forteller om forbedringen i nøyaktigheten i forhold til det den var ved forrige iterasjon.

Det er ingen analytisk måte å komme frem til hvilken varians som er hensiktsmessig å gå ut ifra for å glatte disse modellene, og det må dermed søkes etter verdier som kan forbedre ytelsen. Fordi eksperimentene i denne oppgaven valideres med 10-dobbel kryssvalidering, blir søk langs to dimensjoner, relativ toleranse og gaussiansk prior fort kostbart, fordi antall modell-bygginger og evalueringer øker med 10 ganger antall verdier av den andre variabelen som prøves ut, for hver verdi som legges til i den ene listen..

50 verdier for relativ toleranse fra 3e-8 til 1.01e-6 ble forsøkt (1e-7 er forvalg), og 100 verdier for varians ble forsøkt fra 0.021 til 200.009. Tallene økte med $n * 2e - 8$ for n fra 1 til 50, for toleransen og med $n^2 * 1e - 8$ for n fra 1 til 100 for variansen. Ingen av disse kombinasjonene resulterte

i en forbedring av resultatene fra å bruke TADM uten å endre disse parameterne.

Kapittel 7

Diskusjon og konklusjoner

Eksperimentene utført i kapittel 6 hadde som formål å forsøke en tilsvarende tilnærming til oversettelse av komposita som hos (Baldwin & Tanaka, 2004). Det eksperimentelle designet måtte bli litt annerledes, da norske komposita skrives i ett ord og må deles opp først, som i deres eksperiment bare er tilfelle for de japanske kompositaene, og en MaksEnt-rangering ble brukt til rangering i stedet for en SVM. Forskjellene er dermed så store, ikke minst i språkene som oversettes, at resultatene ikke kan sammenliknes direkte. Likevel viser resultatene i likhet med Baldwin og Tanakas at maskinlæringsteknikken presterer bedre enn de heuristiske tilnærmingene, hvor forskjellige korpusdata for oversettelseskandidatene er satt sammen på en fast måte som de rangeres etter, og støtter således deres funn om at en maskinlærer kombinert med lingvistiske ressurser er en god måte å løse problemet komposita på.

Oppgaven begynner med en diskusjon av tidligere arbeider i kapittel 2, og viser til Rackow et al. (1992) sin delvise implementasjon av et system for å dele opp tyske komposita, som i likhet med norske skrives i ett ord, og oversette dem del for del. Også i dette arbeidet blir korpusinformasjon brukt til å velge mellom flere mulige oversettelser. På grunn av denne likheten får disse arbeidene relevans for eksperimenter på norsk, fordi arbeidsprosessen ligner, selv om måten oppsplittingen foregår blir forskjellig. Videre omtales eksperimentene til Grefenstette

(1999) og Moa (2006), som bruker data fra Internett istedet for et tradisjonelt tekstkorpus, og til avslutningsvis Baldwin og Tanakas mange forsøk på å systematisere bruk av korpora til valg av rett oversettelse. Tanaka og Baldwin (2003) innfører et skille mellom dyp og grunn oversettelse av komposita, mens Baldwin og Tanaka (2004) fokuserer mer på maskinlæring og rangering innenfor rammen av en slik grunn strategi.

I kapittel 3 blir komposita behandlet fra et lingvistisk ståsted. Den svært gamle diskusjonen om hvordan det semantiske forholdet mellom komponentene i et kompositum er bygget opp i forhold til den morfologiske komposisjonen, viser at det ikke er noe entydig bilde. Som tydeliggjort mellom norsk og engelsk varierer måten komposita anskueliggjør seg i skrift både i språk og mellom språk, som gjør det vanskelig å trekke grenser, og derfor eksisterer flere måter å se dette problemet på. I en eksperimentell kontekst som denne er det imidlertid naturlig å se på komposita som en konstruksjon som består av to sammensatte ord det er mulig å identifisere. Nettopp fordi de kan identifiseres ved hjelp av regler er det mulig å lage en metode for å behandle dem systematisk.

Avslutningsvis blir metodene i oppgaven presentert, inkludert Maks-Ent-rammeverket. Her kan informasjon fra forskjellige kilder representeres ved såkalte trekk, og en parameterestimator kan finne frem til en modell basert på disse trekkene, som gir opphav til en sannsynlighet for at en ny observasjon (i denne oppgaven en ny oversettelseskandidat) er en ønsket oversettelse. I oppgaven ble både trekk som var hentet fra korpusfrekvenser og trekk fra en tospråklig ordbok brukt. En av årsakene til at MaksEnt-metoder er egnet til å løse problemer innenfor NLP er nettopp at informasjon kan kombineres fra forskjellige kilder, så lenge de kan representeres som trekk.

I videre arbeider hadde det derfor vært interessant å se forsøk som utvidet horisonten for trekk som ble brukt. Rackow et al. (1992) fant at hvilket verb som sto sammen med et kompositum, eller hvilken sammensetningsformativ den brukte, kunne påvirke hvilken engelsk konstruksjonstype som var den riktige. Det hadde vært interessant å sett dette undersøkt

også for norske komposita, og om slik informasjon kunne gitt MaksEnt-modellene bedret treffsikkerhet. Å inkludere «dyp» informasjon om semantiske egenskaper ved kompositaene kunne også tenkes å samvariere systematisk med sine motstykker i engelsk språk. I kapittel 3 diskuteres inndeling av komposita basert på betydning, informasjon som kunne vært brukt som trekk i en MaksEnt-modell, selv om utfordringene med å innhente slik informasjon er større enn for de grunnere søkene etter korpusfrekvenser og ordbokoppslag. Maskinlæringsrammeverk som brukt i denne oppgaven tilrettelegger for å kombinere dyp og grunn analyse i byggingen av modellene. Det kunne også vært interessant å gjøre en feilanalyse for de indekserte korpusdataene, for å se hvilken andel av instansene som er registrert for hver konstruksjonstype representert ved en mal, som virkelig er en konstruksjon man er ute etter, og hvor stor andel som er støy.

Tabeller

2.1 Webtreff for utvalgte fraser i British National Corpus og på World Wide Web. Kopi av tabell fra Grefenstette, med webtreff fra i dag i egen kolonne. 13

2.2 Enspråklige og tospråklige trekk fra Baldwin og Tanaka (2004). Fordi disse forsøkene ble gjort i begge retninger mellom engelsk og japansk, brukes L_1 og L_2 til å betegne henholdsvis kildespråk og målspråk. ${w_1}^{L2}$ betegner dermed det første ordet (frasen) i oversettelsen til målspråket, L_2, mens ${w_2}^{L1}$ betegner det andre ordet (frasen) kildespråket, del 2 av det kompositum som skal oversettes. 31

2.3 Fra Baldwin og Tanaka. Utvalg av resultater fra oversettelse mellom japansk og engelsk. 32

3.1 Mulige ordklasser som kan stå som forledd i norske komposita. Mulige forledd i rader. 49

5.1 Identifiserte engelske konstruksjonstyper fra utvalgte norske komposita. N_1 og N_2 betegner henholdsvis første og andre ledd i det norske kompositum til oversettelse, og E_1 og E_2 betegner den engelske oversettelsen av disse. Merknaden *POSS* betyr at substantivet står i possesiv form (med apostrof og genitivs -s-, og *PL* betegner at substantivet i oversettelsen skal stå i flertall. *ADJ* betyr at et adjektiv som tilsvarer substantivet skal brukes. GS står for gullstandard. 92

6.1 Korpusbaserte trekk. E_1 og E_2 betegner her oversettelsen av henholdsvis forledd og etterledd av det norske kompositumet. Det kan bestå av to ord ved bruk av enkelte maler. . . . 104

6.2 Tospråklige trekk. N_1 og N_2 betegner forledd og etterledd i det norske kompositumet, og E_1 og E_2 betegner her oversettelsen av henholdsvis forledd og etterledd av det norske kompositumet. Det kan bestå av to ord ved bruk av enkelte maler. 105
6.3 Oppsummering av gullstandardresultater for REF, CTQ-B og CTQ-S, i prosent av komposita som blir oversatt med en oversettelse som stod i gullstandarden. 107
6.4 Oppsummering av gullstandardresultater for BIDIR og MaksEnt-metodene, i prosent av komposita som blir oversatt med en oversettelse som stod i gullstandarden. Evaluering for MaksEnt-eksperimentene er foretatt ved hjelp av 10-dobbel kryssevaluering. 108
6.5 Differanse i gjennomsnittlig prosentvis nøyaktighet over korpora mellom når RASP er brukt som henholdsvis parser og tagger. Et negativt tall viser at tagger-baserte treningsdata var bedre. 111
6.6 Sølvstandard-skår for rangering etter MaksEnt$_{full}$ og CTQ-S 113

A.1 Fullstendige eksperimentresultater. Nøyaktighet målt i prosent. 126

Figurer

5.1 Observasjoner som kan skilles lineært. Figur fra (Lin, 2006). 72
5.2 Observasjoner som ikke kan skilles lineært. Figur fra (Lin, 2006). 73
5.3 Punkter som kan skilles etter å ha blitt projisert med en funksjon. 74
5.4 Flytdiagram over implementasjon. 84
5.5 Rangeringsprosessen illustrert med pseudo-kode. 94

6.1 Plotting av resultater over 3 frekvensbånd. P og T etter korpusidentifikasjon indikerer om RASP ble brukt som parser eller tagger i korpusgrunnlaget. 110

Tillegg A

Oppsummering av resultater

Metode	**REF**		**CTQ**		**CTQ-REV**		**BIDIR-HEUR**		**MaksEnt$_{korpus}$**		**MaksEnt$_{full}$**	
Bånd	*Tagger*	*Parser*	*Tagger*	*Parser*	*Tagger*	*Parser*	*Tagger*	*Parser*	*Tagger*	*Parser*	*Tagger*	*Parser*
						BNC						
Høyt	28,03	25,00	20,45	25,00	32,58	31,82	34,85	35,61	39,80	38.70	51,10	51,90
Middels	20,51	19,23	14,74	20,51	26,28	33,33	28,21	35,26	31,80	36,30	51,20	50,90
Lavt	12,10	11,46	14,01	14,65	19,11	24,20	19,75	21,66	33,60	31,80	45,90	48,90
Gjennomsnitt	19,77	18,20	16,18	19,77	25,62	29,65	27,19	30,55	34,81	35,42	49,30	50,49
Bånd	*Tagger*	*Parser*	*Tagger*	*Parser*	*Tagger*	*Parser*	*Tagger*	*Parser*	*Tagger*	*Parser*	*Tagger*	*Parser*
						+AQAUINT						
Høyt	38,64	35,61	23,48	28,03	40,91	40,91	42,42	43,18	49,80	54,90	57,40	59,70
Middels	23,72	25,00	13,46	18,59	30,77	36,54	32,69	38,46	39,00	41,10	52,00	54,20
Lavt	13,38	12,10	10,19	11,46	19,11	20,38	19,11	21,02	26,80	27,80	45,50	46,70
Gjennomsnitt	24,50	23,59	15,28	18,88	29,66	32,12	30,78	33,70	37,90	40,5	51,31	53,18
Bånd	*Tagger*	*Parser*	*Tagger*	*Parser*	*Tagger*	*Parser*	*Tagger*	*Parser*	*Tagger*	*Parser*	*Tagger*	*Parser*
						+NANTC						
Høyt	35,61	37,12	22,73	28,03	38,64	38,64	40,15	40,91	49,40	51,60	58,70	59,60
Middels	23,08	24,52	12,82	18,71	26,92	29,03	28,21	30,97	38,60	39,60	51,80	52,20
Lavt	16,56	14,01	13,38	12,74	18,47	17,20	18,47	17,83	25,80	26,50	48,00	45,50
Gjennomsnitt	24,50	24,55	15,28	19,37	27,42	27,70	28,31	29,28	37,28	38,54	52,51	52,03

Tabell A.1: *Fullstendige eksperimentresultater. Nøyaktighet målt i prosent.*

Referanser

Akø, J.-O. (1989). *Sammensatte ord: Bruken av s-fuge i moderne bokmål.* Upublisert masteroppgave, Universitetet i Oslo.

Atkinson, K. (2004). *Varcon (variant conversion info).*

Bakken, K. (1998). Leksikalisering av sammensetninger. en studie av leksikaliseringsprosessen belyst ved et gammelnorsk diplommateriale fra 1300-tallet. I *Acta humaniora 38.* Oslo: Universitetsforlaget.

Bakken, K. (2006). Lexicalization. *Encyclopedia of Language and Linguistics, 2nd Edition*.

Baldwin, T. & Tanaka, T. (2004). Translation by machine of complex nominals: Getting it right. I *Proceedings of the acl04 workshop on multiword expressions: Integrating processing.*

Bauer, L. (1983). *English word-formation*. Cambridge University Press.

Bauer, L. (2006). Compounds. *Encyclopedia of Language and Linguistics, 2nd Edition*.

Bergenholtz, H., Cantell, I., Vatvedt Fjeld, R., Gundersen, D., Jónsson, J. H. & Svensén, B. (1997). *Norsk leksikografisk ordbok* (vol. 4). Universitetsforlaget. (ISBN 82-00-22901-7)

Berger, A. L., Della Pietra, S. A. & Della Pietra, V. J. (1996). A maximum entropy approach to natural language processing. *Computational Linguistics*, *22*(1), 39–71.

Breen, J. (1995). *Building an electronic japanese-english dictionary.*

Briscoe, E. & Carroll, J. (2002). Robust accurate statistical annotation of general text. I *Third international conference on language resources and evaluation (LREC 2002)* (s. 1499–1504). Las Palmas, Canary Islands.

Briscoe, T., Carroll, J. & Watson, R. (2006). The second release of the rasp system. I *Proceedings of the coling/acl 2006 interactive presentation sessions.*

Charniak, E. & Johnson, M. (2005). Coarse-to-fine n-best parsing and maxent discriminative reranking. I *In acl* (s. 173–180).

Cortes, C. & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273-297.

Eek Øystein (red.). (2001). *Engelsk stor ordbok: engelsk norsk/norsk-engelsk*. Kunnskapsforlaget. (ISBN 82-573-1288-6)

Endresen, R. T., Simonsen, H. G. & Sveen, A. (2000). *Innføring i lingvistikk* (A. Sveen, red.). Universitetsforlaget.

Faarlund, J. T., Lie, S. & Vannebo, K. I. (1995). *Norsk referansegrammatikk*. Universitetsforlaget.

Fodor, J. A. & Lepore, E. (2002). *The compositionality papers*. Oxford: Clarendon Press.

Grefenstette, G. (1999). The World Wide Web as a resource for example-based machine translation tasks. I *Translating and the computer 21: Proceedings of the 21st international conference on translating and the computer.*

Hagen, K., Johannessen, J. B. & Nøklestad, A. (2000). A constraint-based tagger for norwegian. I *17th scandinavian conference of linguistics.*

Hodges, W. (1998). Compositionality is not the problem. *Logic and Philosophy*(6), 7-33.

Ikehara, S., Shirai, S., Yokoo, A. & Nakaiwa, H. (1995). Toward an mt system without pre-editing — effects of new methods in alt-j/e —.

Johannessen, J. B. & Hauglin, H. (1996). *An automatic analysis of norwegian compounds.*

Johansson, K. A. S. (2004). Multilingual corpora: Models, methods, uses. *TradTerm*, 59-82.

Karlsson, F. (1990). Constraint grammar as a framework for parsing running text. I *Proceedings of the 13th conference on computational linguistics* (s. 168–173). Morristown, NJ, USA: Association for Computational Linguistics.

Klein, D. & Manning, C. (2003). *Maxent models, conditional estimation and optimization.* (ACL 2003)

Koehn, P. & Knight, K. (2003). Feature-rich statistical translation of noun phrases. I *Proceedings of the 41st annual meeting of the association for computational linguistics.*

Langer, S. (1998). Zur morphologie und semantik von nominalkomposita. I *Tagungsband der konvens 1998* (s. 83-97).

Lin, C.-J. (2006). *Support vector machines.* videolectures.net.

Malouf, R. (2002). *A comparison of algorithms for maximum entropy parameter estimation.*

Matsumoto, J., Kitauchi, A., Yamashita, T. & Hirano, Y. (1999). *Japanese morphological analysis system chasen version 2.0 manual* (Teknisk rapport nr. NAIST-IS-TR99009). NAIST.

Mikheev, A. (2000). Tagging sentence boundaries. I *Proceedings of the first conference on north american chapter of the association for computational linguistics* (s. 264–271). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Miller, G. A. (1995). Wordnet: a lexical database for english. *Commun. ACM*, *38*(11), 39–41.

Moa, H. (2005). Compounds and other oddities in machine translation. I *15th nordic conference of computational linguistics.*

Moa, H. (2006). *Search engines and linguistics - with a case study of an automated compound translator using serach engines*. Upublisert masteroppgave, NTNU.

Munthe, S. K. (1972). *Sammensatte ord. en kvantitativ undersøkelse av norsk litteratur and sakprosa.* Upublisert masteroppgave, Universitetet i Oslo/Bergen. (Ikke offentlig tilgjengelig)

Ngai, G. & Florian, R. (2001). *Transformation-based learning in the fast lane.*

Oepen, S., Dyvik, H., Lønning, J. T., Velldal, E., Beermann, D., Carroll, J. et al. (2004). Som å kapp-ete med trollet? towards mrs-based norwegian – english machine translation. I *In proceedings of the 10th international conference on theoretical and methodological issues in machine translation* (s. 11–20).

Pawley, A. (1986). Lexicalization. I *Georgetown universal round table on languages and linguistics 1985. languages and lunguistics: The inderdependence of theory, data and application.* (s. 98-120). Washington DC: Georgetown University Press.

Pelletier, F. J. (2006). Compositionality: Philosophical Aspects. *Encyclopedia of Language and Linguistics, 2nd Edition*.

Rackow, U. (1992). On the treatment of compounds in machine translation.

a study. *IWBS Report*, *221*.

Rackow, U., Dagan, I. & Schwall, U. (1992). Automatic translation of noun compounds. I *Proceedings of the 14th conference on computational linguistics* (s. 1249–1253). Morristown, NJ, USA: Association for Computational Linguistics.

Ratnaparkhi, A. (1996). A maximum entropy model for part-of-speech tagging. I E. Brill & K. Church (red.), *Proceedings of the conference on empirical methods in natural language processing* (s. 133–142). Somerset, New Jersey: Association for Computational Linguistics.

Ratnaparkhi, A. (1997). *A simple introduction to maximum entropy models for natural language processing* (Teknisk rapport). nstitute for Research in Cognitive Science, University of Pennsylvania.

Ratnaparkhi, A. (1998). *Maximum entropy models for natural language ambiguity resolution* (Teknisk rapport). University of Pennsylvania.

R Malouf, M. O., J Baldridge. (2006). *The toolkit for advanced discriminative modeling (tadm).* WWW.

Sampson, G. (1995). *English for the computer*. Oxford University Press.

Sandu, G. & Salo, P. (2006). Compositionality: Sematnic Aspects. *Encyclopedia of Language and Linguistics, 2nd Edition*.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell system technical journal*, *27*.

Svanlund, J. (2002). Lexicalization. *Språk & Stil. Tidskrift för svensk språkforsking.*

Tanaka, T. & Baldwin, T. (2003). Noun-noun compound machine translation: A feasibility study on shallow processing. I *Proceedings of the acl 2003 workshop on multiword expressions: Analysis, acquisition and treatment.*

Toutanova, K., Klein, D., Manning, C. D. & Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. I *In proceedings of hlt-naacl 2003* (s. 252–259).

Velldal, E. (2008). *Empirical realization ranking*. (ISSN 0806-3222)

Wangensteen, B. (red.). (2005). *Bokmålsordboka*. Kunnskapsforlaget. (ISBN 9788257316297)