

# Ekserpering av leksikalske oversettelseskorrespondanser fra parallelltekst<sup>1</sup>

## Innhold

- 1 Bakgrunn
- 2 Ekserperingsprinsipper
  - 2.1 Innledning
  - 2.2 Definisjon av oversettelsesmessig korrespondanse
- 3 Diskusjon av prinsippene
  - 3.1 Utfyllende kommentarer til definisjonen
  - 3.2 Noen egenskaper ved søkeord og korrespondent
  - 3.3 Når en korrespondent tilhører en lukket ordklasse
  - 3.4 Ekserpering av flerordsuttrykk
  - 3.5 Formlikhet og avledningsforhold mellom adjektiv og adverb
- 4 Noen praktiske spørsmål under ekserpering

Noter

Referanser

## 1 Bakgrunn

I prosjektet “Fra parallellkorporus til ordnett” blir oversettelseskorrespondanser ekserpert fra parallelltekster etter bestemte prinsipper. Ekserperingsarbeidet er knyttet til metoden som er presentert i Helge Dyviks manuskript *Semantic Mirrors*. Metoden avleder enspråklige betydninger, og relasjoner mellom disse betydningene, fra et tospråklig materiale, nemlig parallelltekster. Forutsetningen for å få dette til er at oversettelsesmessige korrespondanser er identifisert mellom leksemene i de to gitte språkene, og det vil si at for gitte leksemer i det ene språket har vi funnet settet av deres mulige oversettelser i det andre språket.

Det aktuelle språkparet i prosjektet er norsk og engelsk, og prosjektet har som delmål å utvikle automatisk ordparallellstilling for dette språkparet, siden parallellstilling på ordnivå gir oversettelseskorrespondanser mellom de to språkernes inventarer av ord. Det er også nødvendig å foreta manuell ekserpering av oversettelseskorrespondanser for å ha en

---

<sup>1</sup> Notatet er ført i pennen av Martha Thunes, men bygger også på diskusjoner med, og bidrag fra, Helge Dyvik, Gunn Inger Lyse og Kjersti Drøsdal Vikøren.

“gullstandard” som resultatet av den automatiske ordparallelstillingen kan sammenlignes med. I arbeidet med manuell uthenting av oversettelseskorrespondanser fra English-Norwegian Parallel Corpus (ENPC) har strategien ikke vært å gå gjennom tekstsamlingen fra begynnelse til slutt for å parallellestille tekstene ord for ord. Isteden har vi startet med å finne alle forekomster av noen få utvalgte leksemer i norsk for å finne deres korrespondenter i de parallelle engelske tekstene. I neste trinn har de engelske korrespondentene vært gjenstand for søk for å hente ut deres motparter i de norske tekstene, og det har gitt et større sett av norske leksemer enn vi startet med. I tredje trinn har vi hentet ut settet av korrespondenter i de engelske tekstene for de nye medlemmene i settet av norske leksemer, og så i fjerde trinn søkt på norske korrespondenter for de engelske leksemene som kom til under tredje trinn. Av praktiske grunner har vi måttet begrense antall bevegelser frem og tilbake mellom tekstene, men prinsipielt er det slik at desto flere bevegelser som gjøres, jo mer fullstendig får vi kartlagt de oversettelsesmessige korrespondansene som foreligger i parallellkorpuset.

## 2 Ekserperingsprinsipper

### 2.1 Innledning

For å ivareta konsekvens i ekserperingsarbeidet er det nødvendig å ha prinsipper som kan avgjøre hva som faller inn under begrepet oversettelsesmessig korrespondanse. Samtidig er det vanskelig å definere prinsipper som kan gi svar i ethvert tvilstilfelle som måtte oppstå under ekserpering, og det er uunngåelig at ekserperingsarbeidet får heuristiske innslag og involverer skjønnsmessige avgjørelser. Vi har prøvd å formulere prinsipper som er presise nok til å la seg anvende, men samtidig ligger nær opp til det vi intuitivt mener er riktig.

Våre prinsipper er satt opp for å tjene et bestemt formål, nemlig å samle inn data som metoden i *Semantic Mirrors* skal anvendes på. Metoden produserer et semantisk nettverk, et ordnett, og det som da er relevante data, er oversettelseskorrespondanser mellom leksikalske ord. Prinsippene er derfor først og fremst definert med tanke på ord og uttrykk som tilhører de åpne, leksikalske ordklassene — substantiv, verb, adjektiv og adverb. Vi sier ikke dermed at det er uinteressant å prøve metoden på lukkede klasser og grammatiske enheter som f.eks. preposisjoner, modale hjelpeverb o.a., men i prosjektet “Fra parallellkorpuset til ordnett” prioriterer vi åpne klasser.

Siden vi i arbeidet med å hente ut oversettelseskorrespondanser søker frem og tilbake mellom de parallelle tekstene, vil ekserperingen skje uavhengig av retningen i den oversettelsesprosessen som gav opphav til den konkrete parallellteksten søket gjøres i. Dermed skal vi i det følgende ikke referere til søkeordets oversettelse, men til søkeordets *motpart*, eller *korrespondent*, i den parallelle teksten.

Ved ekserperingen antar vi at tekstene er lemmatisert, og at vi søker på lemmer. Det innebærer for det første at vi må søke på alle former av lemmer som har bøyning. Dernest

betyr det at vi må ta hensyn til fenomenet kategoritvetydighet, altså tilfeller der en bestemt ordform kan representere mer enn et lemma, og der disse lemmaene tilhører ulike kategorier. I slike tilfeller må vi skille mellom kategoriene og søke etter den kategorien som er relevant. Hvis et søk for eksempel har gitt oss den norske presensformen *stoler* som korrespondent for et gitt engelsk søkeord, så skal vi i neste runde søke på alle former av verbet *stole*, men ikke på noen former av substantivet *stol*.

## 2.2 Definisjon av oversettelsesmessig korrespondanse

I dette avsnittet skal vi presentere hvilke betingelser som må være oppfylt for å si at noe faller inn under begrepet ‘oversettelsesmessig korrespondanse’, og vi gir prototypiske eksempler på hvordan betingelsene kan oppfylles. I avsnitt 3.1 diskuterer vi betingelsene i lys av mer perifere eksempler for å illustrere hvor vidt definisjonen er ment å favne.

Ved søk på et ord (eller uttrykk) *a*, registrerer vi ord og uttrykk som kan sies å *korrespondere med a* i de parallelle setningene. Gitt et par korresponderende setninger, kildeprakssetningen *K* og målprakssetningen *M*. *K* inneholder søkelemmaet *a*: *K* = [...*a*...]. *a* korresponderer med et uttrykk *b* i *M* hvis følgende betingelser er tilfredsstillt:

(a) *a* er inneholdt i en maksimal projeksjon eller en eksosentrisk frase, som en setning, kalt  $XP_k$ , og *b* er inneholdt i en maksimal projeksjon eller en eksosentrisk frase, kalt  $XP_m$ .  $XP_k$  og  $XP_m$  korresponderer med hverandre fordi de åpenbart fyller samme rolle i det som er felles i interpretasjonene av *K* og *M*.  $XP_k$  kan være lik *K*, og  $XP_m$  kan være lik *M*. Skjematisk fremstilt:  $K = [...[_{XP_k} \dots a \dots] \dots]$ ,  $M = [...[_{XP_m} \dots b \dots] \dots]$ . Både *a* og *b* kan være flerordsuttrykk.

(b)  $XP_k$  og  $XP_m$  må ha tilstrekkelig like argumentstrukturer til at uttrykkene i *a*s omgivelser står i de samme semantiske relasjonene til hverandre og til *a* som de korresponderende uttrykkene i *b*s omgivelser gjør til hverandre og til *b*. Dette skal utdypes:

- Hvis *a* og *b* uttrykker relasjoner, må disse relasjonene ha likt antall semantiske argumenter, og tildele samme typer roller til de respektive argumentene. Derimot er det ikke nødvendig at lenkningen<sup>2</sup> mellom syntaktiske konstituenten og semantiske roller er identisk for de korresponderende relasjonene, og det må ikke nødvendigvis være likt antall syntaktiske ledd omkring henholdsvis *a* og *b*. Det er heller ikke et krav at relasjonene uttrykt av *a* og *b* er nær-synonyme; de kan være presiseringer eller depresiseringer av hverandre. Et eksempel fra ENPC viser en korrespondanse mellom verbene *buy* og *kjøpe*:

---

<sup>2</sup> Lenkning mellom syntaktiske konstituenten og semantiske roller er en type leksikalsk informasjon som fastlegger begrensningene for hvordan de semantiske roller som deles ut av en bestemt relasjon, kan realiseres syntaktisk.

(1a) It had been specially made, that bed, for the couple they had bought the house from.  
(DL1)<sup>3</sup>

(1b) Den var blitt spesiallaget for det paret de hadde kjøpt huset av. (DL1T)

I eksempel (1) er søkeuttrykket verbet *buy*. Dermed er *a* lik *had bought*, og i henhold til skjemaet i (a) ovenfor er (1a) = K og (1b) = M. Videre kan vi identifisere  $XP_k$  som nominalfrasen *the couple they had bought the house from*, og  $XP_m = \text{det paret de hadde kjøpt huset av}. Verbfrasen *hadde kjøpt* peker seg ut som korrespondent for *a*-uttrykket, og betingelsene for å si at *hadde kjøpt* korresponderer med *had bought* er oppfylt på følgende måte: Relasjonene ‘buy’ og ‘kjøpe’ tar begge (minst) to argumenter, som tildeles rollene agens og patiens. (Et tredje argument, som får rollen benefaktiv, er fakultativt.) Innenfor  $XP_k$  i (1a) tildeler verbet *buy* agensrollen til NPen *they*. Sistnevnte korresponderer på selvstendig grunnlag med NPen *de* innenfor  $XP_m$  i (1b), som der er tildelt agensrollen av verbet *kjøpe*. Videre tildeler *buy* patiensrollen til NPen *the house* i (1a), og denne korresponderer med *huset* i (1b), som også er tildelt patiensrollen, av verbet *kjøpe*.$

- Hvis *a* og *b* uttrykker argumenter, må disse være tildelt samme type semantisk rolle, og hvis *a* og *b* er referensielle uttrykk, må de være koreferente, altså forankret i samme referent i diskursen. Det er ikke nødvendig at *a* og *b* er nær-synonyme uttrykk.

Hvis vi igjen ser på eksempel (1), er disse betingelsene oppfylt for argumentene innenfor de strengene vi ovenfor identifiserte som  $XP_k$  og  $XP_m$ . De korresponderende NPene *they* og *de* er koreferente, og begge er tildelt agensrollen; de koreferente NPene *the house* og *huset* er begge tildelt patiensrollen.

- Hvis *a* og *b* uttrykker ikke-argumenter, som f.eks. adverbielle ledd, må også disse være koreferente, i den forstand at de refererer til samme situasjon, lokasjon eller egenskap og modifierer korresponderende enheter, f.eks. verb. Det er ikke nødvendig at *a* og *b* er nær-synonyme uttrykk. Ikke-argumenter kan være særtilfeller av relasjoner: et adverb som modifierer noe, kan sees som en en-plass relasjon som tar det som modifieres, som argument. Korrespondanse mellom ikke-argumenter kan illustreres med eksempel (2), der vi utpeker adverbet *quite* i (2a) som *a*-uttrykk.

---

<sup>3</sup> Slike koder i parentes refererer til tekstene som finnes i ENPC. En oversikt over hvilke tekster de forskjellige kodene peker til, finnes på <http://www.hf.uio.no/iba/prosjekt/>. Generelt kan vi si at “T” før høyreparentesen i en kode avslører at den aktuelle teksten er en oversatt tekst. Hvis et eksempel ikke er ført opp med en slik kode, er det et konstruert eksempel.

(2a) Luck, quite simply, had come his way, that was all. (AB1)

(2b) Hellet hadde ganske enkelt funnet veien til ham, det var det hele. (AB1T)

I (2b) finner vi adverbet *ganske*, og betingelsene for å si at *quite* korresponderer med *ganske* er oppfylt på følgende måte: I (2a) uttrykker *quite* en en-plass relasjon som tar som argument den relasjonen som er uttrykt av adverbet *simply*, mens *ganske* i (2b) uttrykker en en-plass relasjon som tar som argument den relasjonen som er uttrykt av adverbet *enkelt*. De enhetene som er argumenter for hhv. 'quite' og 'ganske', korresponderer på selvstendig grunnlag.

### 3 Diskusjon av prinsippene

#### 3.1 Utfyllende kommentarer til definisjonen

Noen eksempler skal ytterligere illustrere hvordan betingelsene for “tilstrekkelig like argumentstrukturer” kan oppfylles.

Når *a* og *b* uttrykker relasjoner:

Først vil vi vise at betingelsene for oversettelsesmessig korrespondanse kan være oppfylt når det er en aktiv-passiv-motsetning mellom  $XP_k$  og  $XP_m$ .

(3a) The food was eaten by John.                    *a = was eaten*

(3b) John spiste maten.                                *b = spiste*

I eksempel (3) er søkeordet verbet *eat*; passivsetningen (3a) tilsvarer  $XP_k$  (= K), og aktivsetningen (3b) tilsvarer  $XP_m$  (= M). Relasjonen ‘eat’ i (3a) tar to semantiske argumenter, som fyller rollene agens og patiens, og samme argumentstruktur er knyttet til den synonyme relasjonen ‘spise’ i (3b). I (3a) er agensrollen tildelt uttrykket *John*, og i (3b) er også uttrykket *John* bærer av denne rollen. I (3a) er patiensrollen tildelt nominalfrasen *the food*, og i (3b) er samme rolle tilordnet det korresponderende uttrykket *maten*. Dermed er betingelsene oppfylt for å si at *was eaten* i (3a) korresponderer med *spiste* i (3b), og følgelig blir lemmaet *spise* ekserpert som korrespondent til søkeordet *eat*.

Når vi, som i ekserperingsarbeidet, behandler par av setninger som er løsrevet fra sine respektive kontekster, finner vi ofte at ulike mengde av informasjon er uttrykt i de parallelle setningene. Det kan skyldes at noe informasjon er blitt utelatt i oversettelsen, eller at noe er lagt til, eller det kan skyldes at en viss informasjonsbit er uttrykt på ulike steder i de parallelle tekstene. I sistnevnte tilfeller kan denne informasjonen være inneholdt i setninger som ikke ellers korresponderer oversettelsesmessig, men likevel være tilgjengelig fra begge tekstene hvis vi ser på en videre kontekst. Eksempel (4) illustrerer hvordan betingelsene for oversettelseskorrespondanse kan oppfylles når parallelle setninger inneholder ulike mengde informasjon:

(4a) Da de tok ham igjen, fant de ham ved elven. Han stod på kne og drakk, og han brukte hendene til å øse opp vannet.

(4b) When they caught up with him, they found him by the river. He was on his knees drinking water, using his hands as a cup.

I eksempel (4) er søkeordet verbet *drikke*, og vi identifiserer  $XP_k$  som setningen *Han stod på kne og drakk* i (4a) og  $XP_m$  som setningen *He was on his knees drinking water* i (4b). Intuitivt vil vi mene at forekomsten av *drikke* i (4a) korresponderer med forekomsten av *drink* i (4b),

men hvis vi ser på det som her er  $XP_k$  og  $XP_m$  i isolasjon, oppstår spørsmålet om disse verbene har tilstrekkelig like argumentstrukturer til at betingelsene i (b) i avsnitt 2.2 er oppfylt. I dette tilfellet inneholder  $XP_m$  et argument mindre enn  $XP_k$  gjør, men det synes urimelig av den grunn å forkaste korrespondansen mellom *drakk* i (4a) og *was drinking* i (4b). De synonyme relasjonene ‘drikke’ og ‘drink’ deler begge ut en agensrolle og en patiensrolle. Verbene *drikke* og *drink* har forøvrig til felles at de kan brukes både transitivt og intransitivt. I (4a) opptrer *drikke* som et intransitivt verb; det tildeler agensrollen til *han*, som korresponderer med *he* i (4b). Verbet *drink* er transitivt i (4b), der det tildeler agensrollen til uttrykket *he*, og patiensrollen til *water*. Når verbene *drikke* og *drink* opptrer som intransitive, deler de ut agensrollen til et syntaktisk uttrykt argument, mens patiensrollen går til et underforstått argument som ikke realiseres syntaktisk. I dette tilfellet korresponderer patiensleddet i  $XP_m$  med noe som ikke er uttrykt, men underforstått, i  $XP_k$ , og vi kan forsvare å betrakte argumentstrukturene i henholdsvis  $XP_k$  og  $XP_m$  som tilstrekkelig like til å si at *drakk* i  $XP_k$  korresponderer med *was drinking* i  $XP_m$ . Følgelig ekserperes lemmaet *drink* som korrespondent til søkeordet *drikke*.

Dersom verbet i den ene teksten er obligatorisk transitivt, og motparten i den parallelle teksten er et obligatorisk intransitivt verb, er det derimot vanskelig å argumentere for at verbene har tilstrekkelig like argumentstrukturer til å oppfylle betingelsene for oversettelsesmessig korrespondanse. Mer generelt kan vi si at hvis to relasjoner ikke i noen kontekster kan dele ut samme sett av semantiske roller, har de ikke tilstrekkelig like argumentstrukturer til å kunne utgjøre en oversettelsesmessig korrespondanse.

Når *a* og *b* uttrykker argumenter:

På samme måte som i tilfellet korrespondanse mellom relasjoner, er det ikke nødvendig at de argumentstrukturene som *a* og *b* inngår i, er identiske m.h.t. lenkning mellom semantiske roller og syntaktiske ledd, eller m.h.t. antall syntaktisk uttrykte argumenter. Hvis en situasjon lignende den i eksempel (4) oppstår, som f.eks. i korrespondansen (5), vil ikke ulikheter m.h.t. antall argumenter i hhv.  $XP_k$  og  $XP_m$  nødvendigvis utelukke korrespondanser mellom realiserede argumenter.

(5a) Barna holdt på å spise.

(5b) The children were eating their food.

I eksempel (5) er det slik at setningen (5a) ikke inneholder noe objekt, mens den parallelle setningen (5b) har et objekt. De korresponderende verbene *spise* og *eat* fungerer henholdsvis intransitivt og transitivt i (5a) og (5b). Vi antar at søkeordet er *barna* i dette eksempelet, og vi vil fastslå at *the children* i (5b) er korrespondent for *barna* i (5a). Betingelsene for å si at *barna* og *the children* korresponderer, er oppfylt i og med at disse korresponderende

nominalfrasene er koreferente, og at begge er tildelt agensrollen av de respektive relasjoner de står som argument til.

### 3.2 Noen egenskaper ved søkeord og korrespondent

En begrensning som gjelder i ekserperingsarbeidet, er at en korrespondent ikke kan registreres hvis den gitte forekomsten av søkeordet ikke kan isoleres som en selvstendig oversettelsesenheter. Sagt på en annen måte: Vi ekserperer ikke et uttrykk  $b$  i  $XP_m$  hvis det korresponderer med søkeordet  $a$  pluss et uttrykk i  $a$ s omgivelser, og ikke med  $a$  alene. I eksempel (6) er søkeordet det norske verbet *gå*:

(6a) Han gikk frem.

(6b) He proceeded.

I (6) korresponderer ikke *gå* alene med *proceed*; det er *gå frem* og *proceed* som korresponderer. Dermed kan ikke *proceed* registreres som korrespondent for *gå* på grunnlag av dette eksempelet.

Både søkeordet  $a$  og motparten  $b$  kan, som nevnt, være flerordsuttrykk, og de behøver ikke være kontinuerlige i setningene. I følgende eksempel er verbet *sovne* søkeord:

(7a) Han sovnet med en gang.

(7b) He fell instantly asleep.

I eksempel (7) skal flerordsuttrykket *fall asleep* ekserperes som korrespondent for *sovne*. Dermed blir *fall asleep* søkeordet i påfølgende runde i ekserperingsarbeidet.

En konsekvens av at søkeord og motpart kan være diskontinuerlige strenger, er at det ikke kan være noe krav at  $a$  og  $b$  må være selvstendige syntaktiske fraser. I eksempel (8) finner vi adjektivet *formidable* i den engelske teksten, og det korresponderer med den inkomplette nominalfrasen *et berg av* i den norske teksten:

(8a) It had almost been a relief when a formidable female novelist, vigorously corseted in a florid cretonne two-piece which made her look like a walking sofa, had borne him off to pull out a crumple of parking-tickets from her voluminous handbag and angrily demand what he was proposing to do about them. (PDJ3)

(8b) Han ble nesten lettet da et berg av en romanforfatterinne, kraftig innsnørt i en draktkjole av blomstret kretong som fikk henne til å ligne en vandrende sofa, feide ham med seg, trakk en krøllete bunt røde lapper med parkeringsbøter opp av den digre håndvesken og rasende spurte hva han aktet å gjøre med dem. (PDJ3T)

Som motpart for adjektivet *formidable* i (8a) ekserperer vi uttrykket *et berg av*. Vi skal i avsnitt 3.4 nedenfor komme tilbake til registrering av flerordsuttrykk.

Vi tillater også at *b* kan være en semantisk og morfologisk utskillbar del av et sammensatt ord. Det kan forekomme at et søkeord har en motpart som bare er en del av et ord, og denne ekserperes hvis den kan isoleres som en selvstendig enhet semantisk og morfologisk. F.eks. har søkeordet *basic* bl.a. motparten *grunn-*, som vi kan se i følgende korrespondanse:

- (9a) For the purposes of this Article, “pay” means the ordinary basic or minimum wage (AEEA1)
- (9b) Ved lønn skal i denne artikkel forstås den alminnelige grunnlønn eller minstelønn (AEEA1T)

### 3.3 Når en korrespondent tilhører en lukket ordklasse

I innledningsavsnittet 2.1 påpekte vi at det først og fremst er leksikalske enheter fra de åpne ordklassene substantiv, verb, adjektiv og adverb som behandles her. Imidlertid hender det at ord fra lukkede klasser inngår i oversettelseskorrespondanser, gjerne med nominale uttrykk, men også med modifierende uttrykk (adjektiver og adverb).

En situasjon som kan oppstå under ekserpering av substantiver, er at søkeordet opptrer i en referensiell nominalfrase (f.eks. *dyrlegen*) som korresponderer med en anafor (f.eks. *she*) i den parallelle teksten. Det kan være på sin plass å presisere hva vi mener med anafori her. Anafori er det fenomen at ulike typer uttrykk kan referere tilbake til en entitet som allerede er introdusert i diskursen av et annet refererende uttrykk. Ord av en rekke syntaktiske kategorier kan opptre som anaforiske: pronomen (*han*), demonstrativer (*denne*), determinativer (*min*), kvantorer (*mange*), noen adjektiver (*slik*), og dessuten substantiver. De mest typiske anaforiske uttrykk er pronomen og demonstrativer, som er grammatikaliserte anaforer: Pronomen og demonstrativer har anaforiske egenskaper i kraft av sine grammatiske egenskaper. Det gjelder ikke for de øvrige kategoriene; der kan medlemmer ha anaforisk funksjon i bestemte kontekster, men altså ikke gjennom sin kategoritilhørighet. I tilfellene substantiver, adjektiver og kvantorer er det pragmatisk betinget anafori vi har med å gjøre.

Substantiver kan opptre anaforisk når det foreligger synonymi-, hypo- eller hyperonymirelasjoner mellom dem. F.eks. kan NPene *veterinæren*, *kvinnen* eller *mannen* opptre anaforisk i forhold til NPen *dyrlegen*. Hvis ekserperingsarbeidet avdekker oversettelseskorrespondanser mellom to leksikalske enheter der den ene har enten et mye snevrere eller mye videre sett av denotata enn den andre (f.eks. hvis no. *dyrlege* korresponderer med eng. *woman*), så skal slike korrespondanser registreres, selv om de ikke vanligvis finnes i tospråklige ordbøker. Korrespondanser av denne type gir informasjon om hypo- og hyperonymirelasjoner mellom leksikalske enheter, og det er informasjon vi ønsker å få representert i ordnettet.

Derimot, når grammatikaliserte anaforer, som pronomen og demonstrativer, opptrer som korrespondenter for (leksikalske) søkeord, registrerer vi ikke korrespondansene, siden vårt formål er å kartlegge oversettelsesforbindelser mellom leksikalske enheter. Altså vil vi ikke registrere pronomenet *she* som korrespondent for søkeordet *dyrlege*, selv om betingelsene for å si at forekomster av dem korresponderer oversettelsesmessig, ellers skulle være oppfylt. Eksempel:

(10a) jeg tror ikke på rykter

(10b) I don't believe in that

I eksempel (10) korresponderer nominalfrasen *rykter* med anaforen *that*. Anaforen tilhører klassen av demonstrativer og blir ikke registrert som korrespondent til substantivet *rykte*.

Generelt kan vi si at vi registrerer ingen oversettelsesforbindelse når korrespondenten *b* for et søkeuttrykk *a* består av en grammatikalisert anafor alene, men det kan tenkes at der kan være tilfeller der en anafor inngår som en del av uttrykket *b*, og da vil det likevel være aktuelt å registrere korrespondenten. Det kan da bli nødvendig å vurdere skjønnsmessig om en korrespondent skal ekserperes eller ikke.

Denne problemstillingen berører til en viss grad skillet mellom åpne og lukkede ordklasser. Det er imidlertid for enkelt å generalisere til å si at når vi søker etter korrespondenter til medlemmer i åpne ordklasser, skal vi unngå å registrere korrespondenter som tilhører lukkede klasser. Der er typer av lukkede klasser som det kan være motivert å ta med i nettverket; en av dem er kvantorer, f.eks. *mange, alle, noen, få, ingen*. F.eks. har vi funnet at adjektivet *heavy* kan korrespondere med kvantoren *massevis av*:

(11a) Hun hadde massevis av sminke på seg (RD1T)

(11b) She wore heavy makeup (RD1)

### 3.4 Ekserpering av flerordsuttrykk

Som vi har sett, tillater ekserperingsmetoden at både søkeuttrykk og dets motpart kan være flerordsuttrykk, og ordene i slike flerordsuttrykk behøver ikke forekomme i sekvens, men de må være klart avgrensbare fra sine omgivelser i  $XP_k$  og  $XP_m$ . De samme krav som er beskrevet i avsnitt 2.2 gjelder ved flerordsuttrykk m.h.t. likhet mellom argumentstrukturene i  $XP_k$  og  $XP_m$ .

Når et søkeord korresponderer med et flerordsuttrykk, blir flerordsuttrykket i neste runde det uttrykket som det skal søkes etter korrespondenter til. Når flerordsuttrykk brukes som søkekriterium ved manuell ekserpering, kan det oppstå konflikt mellom anvendelsen av ekserperingsmetoden og dens teoretiske forutsetninger fordi praktiske hindringer kan gjøre det vanskelig å følge prinsippene bak metoden. Slike konflikter er gjerne resultat av at den

bakenforliggende teori angår oversettelseskorrespondanser mellom de to aktuelle språkernes inventarer av tegn, mens ekserperingsarbeidet utføres på et konkret parallellkorpus. I forhold til de to språkene er korpuset begrenset i den forstand at det ikke vil kunne inneholde alle mulige korrespondanser mellom språkene.

Et praktisk problem for ekserperingsmetoden er dermed at det å registrere flerordsuttrykk kan resultere i “fattige” data. Når flerordsuttrykk opptrer som søkeord, gir de opphav til at nye flerordsuttrykk registreres som oversettelseskorrespondenter, og etter en del bevegelser frem og tilbake mellom de parallelle tekstene, risikerer vi å få registrert nokså komplekse uttrykk. Flerordsuttrykk blir mer og mer sjeldne når de vokser i omfang, og da kan antall forekomster bli så lavt at data blir for fattige til å være brukbare. Dessuten er slik uthenting av kollokasjoner fra korpus ikke så sentralt i forhold til vår hensikt, som er å kartlegge leksikalsk semantikk.

Et annet praktisk problem som kan oppstå, er at det kan være svært arbeidskrevende å søke manuelt etter flerordsuttrykk i en stor tekstmengde, særlig hvis et eller flere av ordene som inngår i uttrykket opptrer i forskjellige bøyingsformer.

Et tredje praktisk problem er at et flerordsuttrykk (i likhet med enkeltord) kan ha mer enn en tolkning. Dette kan sees som et homonymiproblem, der én ordform er assosiert med flere betydninger. F.eks. har sekvensen *stå i* to ulike betydninger i setningene *Hun har det travelt med å stå i med guttene* og *Hun stod i vann til knærne*. Dette illustrerer at noe som overfladisk kan se ut som en instans av et flerordsuttrykk, ikke trenger å være eksempel på vedkommende flerordsuttrykk, eller på et flerordsuttrykk overhodet. Samtidig er det i praksis ikke alltid så klart som i disse eksemplene å avgjøre om en gitt sekvens av ord er et flerordsuttrykk av idiomatisk karakter eller en komposisjonell syntaktisk forbindelse som ikke representerer noe flerordsuttrykk. Slike avgjørelser vil kreve skjønnsmessige vurderinger, og dermed oppstår også et metodisk problem, siden vår metode for å utvikle ordnett er basert på at oversettelseskorrespondanser mellom to språk skal være intersubjektivt tilgjengelige, og ikke måtte bero på vurderinger gjort av et bestemt tolkende individ.

De praktiske problemer som har vist seg i forbindelse med ekserpering av flerordsuttrykk, motiverer å legge begrensninger på registreringen av dem. Noen flerordsuttrykk ønsker vi å ta med, og det er slike som det er naturlig å liste i leksikon (f.eks. *fall asleep*), og som vi betrakter som helt klare idiomer med ikke-komposisjonell semantikk. Idiomkriteriet er, som nevnt, problematisk, siden det i konkrete tilfeller kan være vanskelig å avgjøre om et verb subkategoriserer en preposisjon (som *move on* i betydningen ‘fortsette’), eller om det bare foreligger en syntaktisk komposisjonell forbindelse (som i *move on the floor*).

For å unngå slike tolkningsspørsmål legger vi følgende begrensning på registrering av flerordsuttrykk: Ved søk på oversettelseskorrespondanser frem og tilbake mellom parallelle tekster blir flerordsuttrykk registrert så sant det korresponderer med minst et enkeltord i den parallelle teksten. Hvis enkeltordet  $e_k$  i teksten  $k$  korresponderer med flerordsuttrykket  $f_m$  i den

parallele teksten  $m$ , skal  $f_m$  registreres. I neste runde blir  $f_m$  søkeord i motsatt retning. Da finner vi nødvendigvis at  $f_m$  korresponderer med  $e_k$ , og vi kan også finne at  $f_m$  korresponderer med flere enkeltord i  $k$  og med et eller flere flerordsuttrykk i  $k$ . For hvert av disse nye flerordsuttrykkene i  $k$  må vi så undersøke om det, i tillegg til å korrespondere med  $f_m$ , korresponderer med minst et enkeltord i  $m$ . Hvis det er tilfelle, blir det registrert; hvis ikke, unnlater vi å registrere det. På denne måten begrenser vi adgangen til å innføre flerordsuttrykk i ordnettet.

Regelen kan illustreres med et eksempel. Vi antar at søk etter norske korrespondenter for det engelske adjektivet *asleep* gir preposisjonsfrasen *i søvn* som resultat. Dette uttrykket blir registrert, siden det korresponderer med minst et enkeltord i den engelske teksten (*asleep*). I neste runde søker vi etter korrespondenter i den engelske teksten for flerordsuttrykket *i søvn*, og vi antar at i tillegg til korrespondenten *asleep* finner vi preposisjonsfrasen *to sleep*, som i eksempel (12):

(12a) Hun gråt seg i søvn.

(12b) She cried herself to sleep.

Spørsmålet er da om PPen *to sleep* skal registreres. Ved søk på dette uttrykket finner vi ikke at det korresponderer med noe enkeltord i den norske teksten; det blir bare funnet som korrespondent til *i søvn* og *til å sove*. Dermed registrerer vi ikke preposisjonsfrasen *to sleep* som korrespondent for *i søvn*.

La oss se på hvordan begrensningen vi har beskrevet ovenfor, vil virke for uttrykket *fall asleep*, som er et idiom vi ønsker å ha med i ordnettet. Vi kan enkelt begrunne registreringen av dette flerordsuttrykket med at det korresponderer med ettordsuttrykket *sovne* i norsk. Når vi søker i ENPC etter øvrige korrespondenter for *fall asleep*, finner vi at idiomet korresponderer med *falle i søvn*, *sove*, *sovne* og *sovne inn* i norsk. Blant disse skal enkeltordene *sove* og *sovne* registreres som korrespondenter; deretter er spørsmålet om flerordsuttrykkene *falle i søvn* og *sovne inn* korresponderer med noe enkeltord i engelsk. Søk i ENPC viser at i henhold til de betingelser vi setter for å si at uttrykk korresponderer oversettelsesmessig, har hverken *falle i søvn* eller *sovne inn* noe enkeltord som motpart i engelsk, og dermed registrerer vi dem ikke som korrespondenter for *fall asleep*.

I ekserperingsarbeidet er det særlig erfaringer med konstruksjoner som består av verb og partikkel (f.eks. *move on*, *gå frem*) som har ledet oss til å legge begrensninger på registrering av flerordsuttrykk. Uttrykket *pick up* gir opphav til et noe mer omfattende ekserperingsarbeid enn vi hadde i tilfellet *fall asleep*. Verb+partikkel-konstruksjonen *pick up* dukker opp som korrespondent for det norske verbet *hente*. Siden uttrykket *pick up* korresponderer med enkeltordet *hente*, skal det registreres. I neste runde blir da *pick up* gjenstand for søk, og i ENPC kan vi finne at *pick up* korresponderer med følgende ettordsuttrykk i norsk: *betale*,

*finne, få, gripe, hente, høre, kjøpe, lære, løfte, oppfatte, plukke, samle, sjekke* og *ta*. Disse registreres. Dessuten finner vi at *pick up* korresponderer med 28 forskjellige flerordsuttrykk i norsk, alle med et verb som kjerne.<sup>4</sup> I prinsippet skal vi da undersøke, for hvert av de 28 norske flerordsuttrykkene, om det i tillegg til å være motpart til *pick up*, også korresponderer med minst et ettordsuttrykk i engelsk, som jo er betingelsen for registrering. Dermed aner vi en praktisk barriere, siden dette blir et svært tidkrevende arbeid, og derfor velger vi å takle slike tilfeller med å legge en ytterligere begrensning på registreringen av flerordsuttrykk. Når søkeordet er et flerordsuttrykk med et verb som kjerne, og vi finner at det korresponderer med et flerordsuttrykk som også har et verb som kjerne, registrerer vi bare korrespondansen mellom kjernene. Når vi finner at en forekomst av *pick up* korresponderer med f.eks. *snappe opp*, registrerer vi forbindelsen mellom *pick* og *snappe*.

Å følge denne regelen vil føre til at vi registrerer en del oversettelseskorrespondanser som bare vil forekomme i spesielle syntaktiske kontekster, altså kontekster der verbet er etterfulgt av et subkategorisert uttrykk. For vårt formål vurderer vi dette som uproblematisk, fordi oversettelseskorrespondansene mellom verbene er av større interesse for ordnettet enn de kollokasjonene som verbene opptrer i. Dessuten er disse kollokasjonene gjenvinnbare fra korpuset. Eksemplene (13)–(14) illustrerer noen slike korrespondanser:

- (13a) They quickly picked up a kind of pidgin terminology of revolutionary rhetoric that was the period's replacement for schoolboy slang, (NG1)  
 (13b) De snappet fort opp en slags revolusjonær retorikk som fortrenget skolesjargongen, (NG1T)

Uttrykket *picked up* i (13a) korresponderer med *snappet ... opp* i (13b). Bare forbindelsen mellom *pick* og *snappe* blir registrert.

- (14a) His father was a dustman, and Danny had early picked up the art of sorting through rubbish for saleable items. (MD1)  
 (14b) Faren var søppelkjører, og Danny hadde startet tidlig med kunsten å sortere søppel og plukke ut det som kunne selges. (MD1T)

Uttrykket *picked up* i (14a) korresponderer med *startet ... med* i (14b). Bare forbindelsen mellom *pick* og *starte* blir registrert.

---

<sup>4</sup> Uttrykkene er: *danne seg, fange opp, få tak i, legge merke til, løfte av, løfte opp, løfte opp i armene sine, nappe til seg, pakke sammen, plukke opp, puffe opp, pådra seg, raske til seg, renske opp i, samle sammen, snappe opp, starte med, ta av, ta for seg, ta frem, ta i hendene, ta med seg, ta opp, ta ut, taue inn, trekke frem, velge seg ut, øke på*.

### 3.5 Formlikhet og avledningsforhold mellom adjektiv og adverb

Både i norsk og engelsk forekommer det at en og samme ordform kan opptre både som adjektiv og som adverb (f.eks. no. *hurtig* og eng. *fast*). Dessuten er det både i norsk og engelsk grupper av adjektiver og adverb som har en felles leksikalsk rot, og der adverbene er morfologisk avledet fra adjektivene. I norsk gjelder dette adverb som er identiske med nøytrumsformen av tilsvarende adjektiv, f.eks. adjektivet *tørr* og det avledede adverbet *tørt*. I engelsk gjelder dette særlig adverb som er dannet fra adjektiver ved tillegg av suffikset *-ly*, og det gjelder også danning av adverb ved de mindre frekvente suffiksene *-ward*, *-ways* og *-wise* (kfr. Miller 1998: 60). Disse to fenomenene, formlikhet og avledningsforhold mellom adjektiver og adverb, har fått oss til å gjøre et unntak fra det vi slår fast i avsnitt 2.1 om hva kategoritvetydighet har å si for ekserperingsarbeidet.

I 2.1 sier vi at når en bestemt ordform kan representere mer enn et lemma, og disse lemmaene tilhører ulike kategorier, må vi skille mellom kategoriene og søke bare etter den kategorien som er relevant. Vi avviker fra dette når vi i ekserperingsarbeidet støter på adjektiver og adverb som står i avledningsforhold til hverandre, som f.eks. *hurtig* – *hurtig*, *tørr* – *tørt*, og *nice* – *nicely*. Hvis medlemmene i et slikt par synkront sett har samme leksikalske rot, velger vi å oppheve kategoriskillet mellom adjektivet og adverbet, med hensyn til både søkeordet og dets korrespondenter. M.a.o., hvis søkeordet er *hurtig*, vil vi se etter både adjektiviske og adverbelle forekomster av det, og registrere motparter for begge typer under ett. Vi vil heller ikke registrere noe kategoriskille blant de korrespondentene som finnes i den parallelle teksten. Hvis søkeordet er *nice* eller *nicely*, vil vi lete etter forekomster av begge formene, og registrere både adjektiviske og adverbelle korrespondenter. Derimot, hvis søkeordet er *hardly*, har det en annen leksikalsk rot enn både adjektivet *hard* og adverbet *hard*, og vi må da ta hensyn til dette skillet.

### 4 Noen praktiske spørsmål under ekserpering

#### • Registrering av spesielle bøyingsformer:

Vi har slått fast i avsnitt 2.1 at hvis søkeordet har bøyingsformer, må det søkes etter forekomster av alle bøyingsformene. Korrespondentene i den parallelle teksten registreres på leksemnivå, som grunnformer. Der er imidlertid et viktig unntak fra det siste: når en bestemt bøyingsform av et leksem, men ikke lekset selv, opptrer som motpart for et gitt søkeord, da registreres kun den aktuelle bøyingsformen. (15) er et eksempel på dette; søkeordet er substantivet *klasse*:

(15a) Fritidsordningen kan være en del av en skolefritidsordning for 1.-3. klasse. (S11)

(15b) According to the Kindergarten Act, schools can offer 6-year-olds a combination of pedagogical and day-care activities; the latter can be organized jointly with the school's day-care centre for 1st to 3rd-year pupils. (SIIT)

Uttrykket *klasse* i (15a) korresponderer med *pupils* i (15b). Betydningen 'klasse' kan vanskelig tilsvare betydningen av en entallsform av leksemet *pupil*, og derfor registrerer vi i dette tilfellet bare bøyningsformen *pupils* som korrespondent for søkeordet *klasse*.

- Normalisering av skrivemåte:

Parallelltekstene vi søker på, vil av og til by på mer enn én skrivemåte for bestemte ord. Dette skjer når tekstmaterialet inneholder både britisk og amerikansk engelsk, og både bokmål og nynorsk, og vi må ta hensyn til det ved å søke på alle mulige skrivemåter av de gitte søkeord. Variasjon i skrivemåte blant ekserperte korrespondenter kan normaliseres til en variant, så sant de ulike variantene tilhører samme leksem.

- Registrering av ordmellomrom og bindestrek:

Vi har sett ovenfor at hvis søkeordet korresponderer oversettelsesmessig med en streng av ord, ekserperes hele denne strengen. Ordene registreres da med strek ( \_ ) imellom, f.eks.: *splitter\_ny* (motpart til søkeordet *brand-new*). Streken \_ representerer ordmellomrom i slike tilfeller. Når søkeordet korresponderer med en sammensetning der leddene er forbundet med bindestrek, registreres motparten med denne bindestreken. Når søkeordet korresponderer med en klart avgrensbar del av et ord i den parallelle teksten, registreres kun denne delen av ordet som korrespondent, og bindestrek brukes for å angi hvordan den har inngått i en sammensetning (eks. *grunn-* registreres for søkeordet *basic*). Men bindestrek brukes bare dersom uttrykket ikke ellers forekommer selvstendig som motpart for søkeordet. M.a.o. skal segmentet *grunn-* ikke registreres med bindestrek hvis leksemet *grunn* ellers forekommer som oversettelsesmessig korrespondent for søkeordet *basic*.

## Noter

### Referanser

- Dyvik, Helge. 2001. *Semantic Mirrors. A Translational Basis for Linguistic Semantics*. Draft manuscript. University of Bergen.
- Fellbaum, Christiane (ed.). 1998. *WordNet: an Electronic Lexical Database*. Cambridge, Massachusetts, and London: The MIT Press.
- Miller, Katherine J. 1998. Modifiers in WordNet. In: Fellbaum, Christiane (ed.). 1998. Pp. 47–67.

