

LEXICONORDICA

17 · 2010

LEKSIKOGRAFI OG
SPRÅKTEKNOLOGI
I NORDEN

SÆRTRYK

NORDISK FORENING FOR LEKSIKOGRAFI

LexicoNordica 17 · 2010

Leksikografi og språkteknologi i Norden

Hovedredaktører

Henrik Lorentzen (ansvarshavende)

Ruth Vatvedt Fjeld

Nasjonale redaktører

Sturla Berg-Olsen

Ken Farø

Jón Hilmar Jónsson

Nina Martola

Emma Sköldberg

© 2010 LexicoNordica og forfatterne

Omslag og sats: Laurids Kristian Fahl

Trykk: Rosendahls – Schultz Grafisk A/S

LexicoNordica trykkes med støtte fra

Ekspertgruppen Nordens språkråd

ISSN 0805-2735

INNHold

Ruth Vatvedt Fjeld & Henrik Lorentzen

Leksikografi og språkteknologi i Norden.....9

Tematiske bidrag

Eckhard Bick

DeepDict – et korpusbaseret relationelt leksikon..... 17

Lars Borin

Med Zipf mot framtiden – en integrerad lexikonresurs
för svensk språkteknologi 35

Kristin Hagen & Anders Nøklestad

Bruk av et norsk leksikon til tagging og andre
språkteknologiske formål 55

Jakob Halskov

Halvautomatisk udvælgelse af lemmakandidater
til en nyordsordbog..... 73

Viggo Kann

KTHs morfologiska och lexikografiska verktyg
och resurser99

Krister Lindén & Lauri Carlson

FinnWordNet – WordNet på finska via översättning.....119

Anna Björk Nikulásdóttir & Matthew Whelpton

Lexicon Acquisition through Noun Clustering.....141

Bolette Sandford Pedersen
Semantiske sprogressourcer – mellem sprogteknologi
og leksikografi163

Eiríkur Rögnvaldsson
Sprogteknologiske ressourcer for islandsk leksikografi.....181

Christian Sjögreen & Emma Sköldberg
Svenska ordboksredigeringsystem – med fokus
på Cronoma..... 197

Trond Trosterud
Felles leksikalske ressursar for språkteknologi
og leksikografi 211

Ikke-tematiske bidrag

Loránd-Levente Pálfi, Erzsébet Stokholm & Sven Tarp
Bilingvale ordbøger med dansk og ungarsk..... 227

Bo-A. Wendt
En SAOB-artikel växer fram249

Anmeldelser

Ilse Cantell
Ordbok över karelskan på Internet 277

Cathrine Fabricius-Hansen
Lexicography in the 21st Century. In honour of
Henning Bergenholtz.....289

Ruth Vatvedt Fjeld & Sven-Göran Malmgren
Värd ett besök – om DSL:s nya webbsida ordnet.dk..... 297

<i>Jan Terje Faarlund</i>	
Norsk Ordbok, band VIII	313
<i>Anna Helga Hannesdóttir</i>	
”Ordaboken moste tryckias”	321
<i>Riina Klemettinen</i>	
En deskriptiv finsk frasordbok	337
<i>Kristina Nikula</i>	
Svensk ordbok – en guldgruva för språkintresserade	351
<i>Loránd-Levente Pálfi</i>	
Finn Stefánsson: Symbolleksikon.....	377
 Kommentarer til tidligere bidrag	
<i>Christian Becker-Christensen</i>	
Nogle bemærkninger til Henning Bergenholtz: “Hurtig og sikker tilgang til informationer om ordforbindelser” i LexicoNordica 16.....	395
 Konferanser	
<i>Marcin Overgaard Ptaszynski</i>	
Rapport fra den 10. Konference om Leksikografi i Norden.....	407
Inbjudan till 11:e Konferensen om lexikografi i Norden	415
Redaksjonelt	419

FinnWordNet – WordNet på finska via översättning

Krister Lindén & Lauri Carlson

FinnWordNet is a WordNet for Finnish that conforms to the framework given in Fellbaum (1998) and Vossen (1998). FinnWordNet¹ is open source and currently contains 117,000 synsets. A classic WordNet consists of synsets, or sets of partial synonyms whose shared meaning is described and exemplified by a gloss, a common part of speech and a hyperonym. Synsets in a WordNet are arranged in hierarchical partial orderings according to semantic relations like hyponymy/hyperonymy. Together the gloss, part of speech and hyperonym fix the meaning of a word and constrain the possible translations of a word in a given synset. The Finnish group has opted for translating Princeton WordNet 3.0 synsets wholesale into Finnish by professional translators, because the translation process can be controlled with regard to quality, coverage, cost and speed of translation. The project was financed by FIN-CLARIN at the University of Helsinki. According to our preliminary evaluation, the translation process was diligent and the quality is on a par with the original Princeton WordNet.

1. Inledning

Ett WordNet är en tesaurus som består av grupper av synonymer, dvs. ord som hör till samma ordklass och som är utbytbara i en given kontext. Sådana grupper av synonymer kallas även *synset*. I WordNet är synonymgrupperna partiellt hierarkiskt ordnade enligt semantiska relationer såsom hyperonymer, hyponymer, meronymer, antonymer, osv. Varje synset (med några få undantag) har en förklaring som exemplifierar eller beskriver dess betydelse.

1 FinnWordNet: <http://www.ling.helsinki.fi/cgi-bin/fiwn/search>

En typisk WordNet-förklaring inkluderar en ordboksdefinition och ett exempel, t.ex. *devilize* ‘förvandla till en djävul eller göra demonisk; “en människa demoniserad av kriget”’. Somliga förklaringar innehåller bara en hyperonym (exempelvis ett namn på en djurart).

FinnWordNet är ett WordNet för finska som följer den struktur som beskrivs i Fellbaum (1998) och Vossen (1998) för Princeton WordNet, som är öppen källkod och för närvarande innehåller 117 000 synset. FinnWordNet är en översättning av synseten i Princeton WordNet 3.0.

Härnäst ger vi en överblick över existerande WordNet på andra språk. Längre fram går vi igenom fördelarna med manuell översättning, arbetsflödet och några teoretiska och praktiska problem som vi stötte på innan vi utvärderar arbetet och nämner något av det som återstår nu när grovjobbet är gjort.

2. WordNet

WordNet på andra språk har skapats med olika metoder. Vi börjar med en kort genomgång av huvudalternativen. Dessutom försöker vi uppskatta hur mycket arbete det innebär att göra ett nytt WordNet utgående från Princeton WordNet.

2.1. WordNet på andra språk

WordNet har också skapats för andra språk än engelska, men generellt sett har de inte lika god täckning som Princeton WordNet. I tabell 1 ser vi en jämförelse av storlekarna på WordNet för olika språk och språkfamiljer. Siffrorna har hämtats från källor på internet under våren 2010.

WordNet	Synset
Princeton WordNet (Fellbaum 1998, WordNet 3.0)	~120 000
EuroWordNet (Vossen 2004a)	~10 000–50 000
BalkaNet (Vossen 2004b)	~18 000
Polskt WordNet (Piasecki m.fl. 2009)	~18 000
Danskt WordNet (DanNet 2009, Pedersen 2010)	~41 000–60 000
Svenskt WordNet (Viberg m.fl. 2002)	~15 000
NorNett (Fjeld och Nygaard 2009)	~50 000 synset? (80 000 relationer)

Tabell 1: WordNet på olika språk och för olika språkfamiljer och deras storlekar i antal synset

2.2. Att skapa ett WordNet för ett nytt språk

Om man vill skapa ett WordNet för ett nytt språk, kan man välja mellan tre olika alternativ: skapa ett WordNet från grunden, översätta ett annat WordNet eller använda en toppontologi och utvidga den med en lokal synonymordbok.

Om man väljer det första alternativet och skapar ett WordNet från grunden, måste synseten utvinnas med olika automatiska metoder ur stora korpusar. Ett bra exempel på detta är polska (Piasecki m.fl. 2009), där synseten utvanns från polska korpusar. För att säkerställa att de mest väsentliga orden och deras synonymer kom med, blev den polska gruppen ändå tvungen att använda en ordlista över det centrala ordförrådet i polska med cirka 30 000 ord. Dessutom måste synseten i ett WordNet som skapas från grunden separat länkas till synseten i ett annat WordNet om man vill skapa en tvåspråkig ordbok.

Det andra alternativet, dvs. att översätta t.ex. Princeton WordNet rubb och stubb, tonar ned argumentet att varje språk är så annorlunda att man måste skapa alldeles egna synonymgrupper

och synonymhierarkier för varje språk. De flesta ord i ett språk är namn på objekt och fenomen i en för mänskligheten gemensam extern verklighet som varje språk har hittat på eller lånat namn för. Det kan finnas en viss skillnad i hur små nyanser som fått egna ord beroende på hur viktigt ett fenomen är inom ett språkområde, men i stort är den grundläggande fysiska verkligheten densamma i alla kulturkretsar: solen går varje dag upp i öster, vattnet är vått och rinner, vi föds, blir hungriga, äter mat och dör, osv. För abstrakta begrepp kan man däremot förvänta sig en större divergens. Ett exempel i Princeton WordNet är specifikt amerikanska begrepp med alla sina moderna sidobetydelser och nyanser som inte finns i ett potentiellt målspråk. Se exempel i tabell 2.

engelska	svenska	finska
hungry	hungrig	nälkäinen
hunger	hungra	nähdä nälkää
Thanksgiving	tacksägelsedag?, skördefest?	kiitospäivä, ≠ kekri

Tabell 2: Exempel på likheter och skillnader mellan konkreta och kulturbundna begrepp i olika språk. Hungern är global men festerna varierar.

Ett tredje alternativ och en medelväg att skapa WordNet för olika språk är därför att översätta 5000 centrala begrepp i Princeton WordNet för att skapa en kärna som sedan utvidgas med en lokal synonymordbok. Detta är den väg man har tänkt sig för DanNet, där man börjat med att skapa den lokala synonymordboken på basen av en lokal ordbok (Pedersen 2010). Avsikten är att sedan länka DanNet till Princeton WordNet. Detta skapar en länk mellan de mest centrala begreppen i de båda språken och underlättar användningen av den nya ordboken vid översättning. Detta användes i t.ex. EuroWordNet (Vossen 2004a). Världsorganisationen för WordNet, The Global WordNet Association (GWA), har

gjort ett urval av 5000 grundläggande begrepp enligt strukturella principer för WordNet i olika språk såsom begreppens position i betydelsehierarkin, antal relationer, frekvens med vilken de förekommer i definitioner eller förklaringar och vilka deras morfologiska komplexitet och beroenden är. I många fall finns det ingen statistik på hur vanlig en viss betydelse är och därför är den lista på 5000 grundbegrepp som GWA föreslår en rätt brokig samling med det amerikanska TV-programmet *60 minutes* men utan ordet *television*.

Å ena sidan fordrar skapandet av en enspråkig resurs från grunden att man upprepar årtionden av lexikografiskt arbete som investerats i Princeton WordNet, men å andra sidan kräver en översättning av hela Princeton WordNet att man även tar ställning till några specifikt amerikanska fenomen som inte nödvändigtvis har en exakt motsvarighet i målspråket. I många språk finns dessutom existerande synonymordböcker och därför har många WordNet-projekt valt den tredje vägen, dvs. en kombination av översättning och en lokalt skapad betydelsehierarki.

Vi anser att den grundläggande enheten i översättning är en betydelseenhet som kan representeras av ett synset med hyperonym, ordklass och förklaring som tilläggsinformation för att styra översättningen av de olika synonymerna i ett synset. Vi valde därför att översätta alla synset. Detta ger oss en parallell uppsättning av synset i de två språken så att vi kan återanvända de flesta semantiska relationer som definierats i Princeton WordNet och dessutom få en tvåspråkig ordbok.

Finska är som språk obesläktat med de andra språken med ofentligt tillgängliga WordNet. Dessutom är vi inte medvetna om att alla synset i Princeton WordNet skulle ha översatts tidigare. Det återstår att utvärdera i ett senare skede hur många specifikt finska ord som saknas och hur många specifikt finska fenomen och betydelser som behöver läggas till.

2.3. Princeton WordNet 3.0

Nedan i tabell 3 och 4 ger vi en uppskattning på mängden av det översättningsarbete som vi valt. I tabell 3 ser vi att Princeton WordNet 3.0 har 117 659 synset eller betydelseenheter. Det finns 155 287 olika ord (strängar) i dessa synset. Några ord förekommer i flera synset. Detta skapar totalt 206 941 synonymer eller ord-synset-par att översätta.

Princeton WordNet	ord	synset	synonymer
substantiv	117 798	82 115	146 312
verb	11 529	13 767	25 047
adjektiv	21 479	18 156	30 002
adverb	4 481	3 621	5 580
Totalt	155 287	117 659	206 941

Tabell 3: Antal ord, synset och synonymer i Princeton WordNet

I tabell 4 ser vi att de flesta ord bara har en betydelse, dvs. de förekommer i bara ett synset. Eftersom det finns färre synset än ord är det naturligt att det finns ett antal synset som har fler än ett ord i sig. I medeltal är antalet synonymer per synset 1,8. Dessutom förekommer somliga ord (strängar) i flera olika synset, dvs. är homonymer för olika betydelser. I medeltal har ett ord 1,3 betydelser. De flesta ord har bara en betydelse, så om man betraktar bara de ord som har flera betydelser, har de flertydiga orden nästan 3 betydelser i medeltal.

Princeton WordNet	monosema ord	homonyma ord	homonymer totalt
substantiv	101 863	15 935	44 449
verb	6 277	5 252	18 770
adjektiv	16 503	4 976	14 399
adverb	3 748	733	1 832
Totalt	128 391	26 896	79 450

Tabell 4: Antal monosema och homonyma ord samt det totala antalet homonymer i Princeton WordNet

3. Manuell översättning av synset

Vi valde professionell manuell översättning av alla synset i Princeton WordNet till finska eftersom den manuella översättningsprocessen är kontrollerad i flera avseenden:

1. Manuell översättning görs av människor och garanterar därför *hög kvalitet* i jämförelse med maskinöversättning och de datautvinningstekniker som normalt används för att upptäcka ords betydelser i stora korpusar.
2. Översättning av alla synset garanterar att vi får en *stor mängd*, dvs. cirka 200 000 betydelser, så att vi får mer än ytterligare en kärnvokabulär med bara de mest centrala ordbetydelserna.
3. Att använda en översättningsbyrå till ett överenskommet pris per ord garanterar en *kontrollerad kostnad*.
4. En översättningsbyrå garanterar även en viss *hastighet*, dvs. 3–6 månader för en första version med upp till 4–5 översättare som jobbar parallellt.

I det följande går vi in på ytterligare några fördelar med ett översatt WordNet och betraktar specifika utmaningar när det gäller att

översätta WordNet jämfört med att översätta vanliga dokument. Vi ger en överblick över översättningsprocessens arbetsflöde och kvalitetskontroll samt bekantar oss med de metadatum som översättaren förväntas bidra med för att hjälpa till med efterbearbetningen för att förbättra kvaliteten.

3.1. Ytterligare fördelar med ett tvåspråkigt WordNet

Översättning gav oss möjlighet att kontrollera processen och producera resultat effektivt, men dessutom ville vi skapa en tvåspråkig resurs som är direkt länkad till Princeton WordNet. Vi skapade ett fritt tillgängligt tvåspråkigt finsk-engelskt WordNet på fyra månader och hoppas att det kan fungera som ett exempel för andra språk.

Den tvåspråkiga ordboken bestående av två parallella WordNet kan användas till informationssökning och med lite modifikation även för maskinöversättning tack vare det stora lexikonet med 100 000–200 000 översättningar med disambiguerad betydelse. Det bör påpekas att nyckelfraser i informationssökning inte är de vanligaste orden utan de mest specifika orden som namn och termer, som ofta saknas i enkla tvåspråkiga ordböcker med enbart kärnvokabulär. Det bör också påpekas att översättningar utan kontext inte är direkt användbara i maskinöversättning, men i WordNet fungerar de engelska förklaringarna och definitionerna som en grund för att utvinna en kontext.

Vi tror att det finns tillräckliga likheter mellan engelska och finska lexikala enheter för att överbygga eventuella olikheter, även om en grundlig utvärdering är nödvändig för att ge en slutlig indikation på hur många specifikt finska kärnbetydelser och kärnord som saknas och hur mycket manuell och halvautomatisk bearbetning som ännu behövs.

3.2. Översättningsutmaningar för WordNet

Vårt dilemma är att vi vill ha alla tänkbara synonymer för en specifik ordbetydelse men samtidigt vill vi ha enbart de relevanta synonymerna. Vårt mål står delvis i konflikt med hur professionella översättare tränas att översätta kommersiella dokument, dvs. att i mån av möjlighet undvika icke-standardiserade ordvarianter. Nu vill vi ha alla varianter i tillägg till standardöversättningen. Vi löste problemet genom att be översättarna hitta finska översättningar till ett engelskt synset som helhet men att översätta varje engelsk synonym i ett synset med sina närmaste särskiljande översättningar. Detta uppmuntrade översättare att leta efter mer än en finsk motsvarighet för att matcha de engelska synonymerna. För att underlätta och stabilisera översättningen bad vi översättaren fästa uppmärksamhet vid²:

1. *förklaring*, dvs. betydelsekontext eller betydelsedefinition
2. ordets *hyperonym*, dvs. begreppets abstraktionsnivå
3. *ordklass*, dvs. den morfo-syntaktiska kontexten

Vi använde en vanlig översättningseditor som stöder XML-format för att skydda de fält som vi ville visa översättarna enbart som tilläggsinformation för att guida deras översättning av synseten.

3.3. Arbetsflöde och kvalitetskontroll

Först tänkte vi dela in WordNet i sektioner enligt ämnesområde, men alla ord har inte en angivelse om ämnesområde och många ämnesområden hade relativt få ord, så vi övergav idén till förmån för en mera genomgripande och lättgenomförbar princip. Vi an-

2 De här tre parametrarna hade i förberedande experiment med studerande i översättning konstaterats vara värdefulla stabiliserande riktlinjer för översättning av synset i WordNet.

vände följande arbetsflöde för att åstadkomma kvalitetskontroll och en någotsånär jämn fördelning av arbetet mellan översättarna:

1. WordNet transformerades till XML-format och delades in i 20 alfabetiskt ordnade synsetsamlingar med enbart de fält som var väsentliga för översättningsarbetet. Cirka 10 000 ortografiskt identifierade egennamns³ synset (med 20 000 varianter) åtskiljdes i en separat egennamnsfil. Många mindre självklara namngivna objekt och fenomen blev trots det kvar i de övriga filerna.
2. En synsetsamling överlämnades till en översättarkandidat som först producerade ett översättningssampel för 1000 engelska synonymer.
3. Översättningsprovet inspekterades av en översättare ansvarig för kvalitetskontrollen.
4. Om det inte fanns några större klagomål, gavs översättaren ytterligare instruktioner på basis av översättningsprovet och han fick hela filen att översätta inom en viss tid.
5. Efter att filen överlämnats, gjordes en slutlig kvalitetskontroll för att bestämma om översättningen föll inom ramen för de uppställda kvalitetskraven.

3.4. Praktiskt översättningsarbete

För att läsaren bättre skall förstå översättningsprocessen visar vi ett förenklat sampl av en XML-kodad fil:

```
<GLOSS>an Egyptian descended from the ancient
Egyptians
<HYPER ID="109700492">Egyptian
<SYNONYM PoS="n">
```

3 Egennamn, dvs. namn på personer och fenomen, har separata synset i WordNet. Synsetets förklaring kan vara väsentlig för att identifiera vem eller vad det är fråga om. Namn på växter och djur har normalt flera varianter bl.a. ett latinskt namn.

```
<Tuv Lang="EN-US">Copt
<Tuv Lang="FI">koptilainen
```

Vi valde att använda SDL Trados som vår översättningsomgivning, eftersom den är vanlig bland professionella översättare. Trados TagEditor stöder översättning av XML-dokument. Trados håller översatta ord i översättningsminnet. Ett översättningsminne erbjuder mycket lite hjälp när man skall översätta synset där en ny förekomst av ett ord i ett annat synset sannolikt kräver en annan översättning. Editorn hjälper till att begränsa editeringen till de fält som skall översättas och tillåter oss att definiera de specifika koder som skall användas under översättningsprojektet. Översättarna använder koderna för att ange metadata om sina översättningar för att styra kvalitetskontrollen och vidareutvecklingen till ställen där det kan finnas brister. Efter en diskussion med översättarna beslöt vi oss för följande koder:

- <or/>
- <approximate/>
- <broader/>
- <narrower/>
- <unconfirmed/>
- <note> ... </note>
- <GEN/>

Med <or/> separerar vi alternativa översättningar av samma term för att undvika tvetydighet mellan objektspråk och metaspråk (interpunktion såsom komma kan förekomma som en del av en objektspråksfras). Med <approximate/> kan översättaren ange att han efter en rimlig insats inte kunde hitta någon bra motsvarighet. Med <broader/> indikerar han att översättningen är mera generell än originaltermen och med <narrower/> att översättningen är mera specifik än originaltermen. Avsikten med koderna var inte att problematisera varje motsvarighet utan att låta översättaren

ange lexikala luckor. I praktiken användes koderna rätt lite (cirka en gång per tusen betydelser). Exempelvis användes <broader/> för översättning med hyperonymen *nahka* 'leather' för *ooze leather*, och <narrower/> för översättning med hyponymen *yläpu-
rentainen* 'overshot (jaw)' för *overshot*. Koden <unconfirmed/> användes om översättaren ansåg att termen var korrekt men inte kunde bekräftas i skrivna källor. Koden <note> ... </note> tillät översättaren att lägga till den information han ansåg lämplig för senare inspektion och vidareutveckling av översättningen, t.ex. för att indikera problem i den engelska källan.

Översättarna fick en manual före uppgiften med exempel på goda översättningar och sådana översättningar som borde undvikas. Kvalitetskontrollanten kunde därför närmast hänvisa till manualen för att ge ytterligare instruktioner. Översättarna ombads tänka ut motsvarigheter som kunde ersätta källtermen i synsetkontexten, undvika nonsensöversättningar, parafraaser och betydelseförklaringar. Vidare ombads de undvika sammanblandning av objektspråk och metaspråk såsom parentetiska tillägg, ad hoc grammatikaliska förklaringar, osv.

Översättarna instruerades att bevara samma ordklass som i källspråket när det var möjligt. Avvikelser från detta borde anges för att underlätta en senare automatisk morfologisk analys av översättningen.

Koden <GEN/> användes i det allmänt förekommande fallet att ett engelskt adjektiv översattes med genitiv av ett finskt substantiv. En mera komplett annotering av grammatiken kunde ha utvecklats men det avstod vi avsiktligt från för att inte komplicera översättarens arbete. En ytterligare genomgång av materialet av en tränad lingvist kan bättre åtgärda den aspekten.

4. Översättningsproblem

Teoretiskt sett finns det många problem förknippade med översättningen av en tesaurus. Vi tittar först på ett prototypexempel med icke-matchande semantiska fält, för att sedan titta på vilka problem som förekom i praktiken. Vi går också igenom hur vi kontrollerade översättningskvaliteten under översättningsprocessen och gör några allmänna observationer om den översättningskvalitet vi fick.

4.1. Teoretisk utmaning – icke-matchande semantiska fält

Från en teoretisk synvinkel har det största översättningsproblemet beskrivits som icke-matchande semantiska fält mellan begreppen i två språk. Beträffande WordNet i synnerhet kan vi ha icke-matchande synset i två språk, eftersom inte alla ord som är synonymer i ett språk har ekvivalenter som uppfattas som synonymer i ett annat språk, om man betraktar ordens kärnbetydelse eller mest frekventa betydelse. Vår intuition är att detta vanligen inte är ett problem vid översättning av WordNet, eftersom ordklass, förklaring och hyperonym ger en klar indikation om vilken översättning som lämpar sig i den givna kontexten. Dessutom har den semantiska rymden i Princeton WordNet under en lång tid granskats ur många synvinklar.

Låt oss ta en titt på ett exempel som har icke-matchande semantiska fält i engelska och finska på grund av det sätt som engelska och finska hanterar nationalitetsord. I båda språken är nationaliteten ofta ett adjektiv härlett från namnet på landet. På finska sammanfaller namnet på språket ofta med landet. På engelska däremot sammanfaller namnet på språket vanligen med nationaliteten.

<GLOSS> **an Egyptian descended from the ancient Egyptians**

<HYPER> **Egyptian**

<POS="n" SENSE="1" COUNT="0">

"EN-US": **Copt**

"FI": **koptilainen** <or/> **kopti**

<GLOSS> **the liturgical language of the Coptic Church used in Egypt and Ethiopia; written in the Greek alphabet**

<HYPER> **Egyptian**

<POS="n" SENSE="1" COUNT="0">

"EN-US": **Coptic**

"FI": **koptin kieli** <or/> **kopti**

<GLOSS> **of or relating to the Copts or their Church or language or art; "the distinctive Coptic art of 6th-century Christian Egypt"**

<HYPER> **Egyptian**

<POS="a" SENSE="1" COUNT="0">

"EN-US": **Coptic**

"FI": **koptilainen**

I detta fall är WordNet tillräckligt finmaskigt för att översättningen av begreppen såsom de har underindelats i WordNet inte skall ställa till problem. Detta kan förstås bero på att det är fråga om en hel grupp av ord som har liknande icke-matchande fält i många språk så problemen har för länge sedan upptäckts och åtgärdats. I exemplet kompliceras översättningen av att ordet även råkar vara en benämning på en religiös inriktning, men under hyperonymen *Church* finns det även ett separat synset för *Coptic Church* som då översätts med *koptilaisuus*.

Vi ser det som en styrka hos Princeton WordNet att det har utvecklats och uppdaterats lexikografiskt under många år. Detta arbete tas tillvara vid översättning. Dock återstår vissa problemområden vid närmare granskning.

Ett framträdande problem är behandlingen av idiom i Word-

Net. Ibland förekommer idiomerna exemplariskt dokumenterade som separata betydelseenheter med egna synonymer – dock inte alltid, t.ex. finns *kick the bucket* som underbetydelse av *kick*. I störande många fall låter WordNet en icke-kompositionell betydelseenhet utgöra ett separat synset, t.ex. *tartaric* i *tartaric acid* eller *take* i *take in*. Detta är ett återfall till det traditionellt lemma-baserade tänkandet i pappersordböcker där lexikografen även förutsattes hitta ett lämpligt uppslagsord för flerordiga uttryck. Sökningen sköter datorn numera alldeles utmärkt, så lexikografen kan koncentrera sig på innehåll och betydelse. Det bästa en översättare kan göra är att ändå översätta hela idiomerna och anteckna svårigheten i en kommentar.

4.2. Problem i praktiken

I praktiken visade det sig att det ibland är svårt att hitta rätt eller ens någon översättning alls för:

- medicinska termer såsom varunamn på mediciner
- växter och organismer specifika för Amerika
- kemiska substanser såsom varunamn på aktiva ingredienser
- olika ismer såsom religiösa rörelser
- lagtekniska termer
- affärstermer
- andra termer som är specifika för amerikansk kultur

Ett exempel på icke-översättbar amerikanska är *hanging chad*, dvs. en icke-fullföljd perforering av en röstsedel på grund av en felaktig röstningsmaskin, vilket var förstasidesnyheter under valåret 2000 i USA. Det finns även specifika WordNet-ord som är svåra att hitta på webben utanför själva WordNet (och dess kloner). Ett exempel är *spiritize* 'att förse med ande eller andlighet'. Andra exempel är *spouter* 'en frustande val', *tapper* 'en person som slår lätt, ofta upp-

repat, på en yta', och *Tareekh e Kasas* 'en organisation av muslimer i Indien som dödade hinduer i september 2002 och som tros ha samröre med muslimska terrorister i Pakistan'. En uppdaterad version av WordNet på nätet kunde kanske göra upp användningsstatistik på orden om man vill ta bort mera efemära dagsländor och *hapax legomena*. Å andra sidan är de här orden mest ett problem vid översättning av WordNet, vilket ju görs rätt sällan. I daglig användning av Wordnet söker ingen upp de udda orden och därför stör de inte heller. I praktiken kan det kosta mer att ta bort dem än att betala för lite mer dataminne.

De udda orden är exempel på kulturbundna fenomen eller *realia* som har utvecklats i det amerikanska samhället under de senaste tvåhundra åren och som därför är specifika för amerikansk kultur. Många är namn på ideologier och sociala fenomen som är okända i finländsk kultur och i det finländska samhället. Detta verkar bekräfta vår ursprungliga hypotes att bara relativt nya och kulturspecifika fenomen orsakar verkliga problem vid översättning. Lyckligtvis har de problematiska orden sällan synonymer. Dessutom är de en liten minoritet av alla ord i WordNet och motiverar inte varför man skulle göra om 95 % av de icke-problematiska synseten från grunden.

En liten motsatt brist i WordNet är att det ibland görs distinktioner där samma ord används i flera olika synset utan att det har betydelse vid översättningen, dvs. alla nyanserna har samma ord även i målspråket. Ibland kan detta bero på en tillfällig likhet mellan språken.

4.3. Indikatorer på översättningskvalitet

För att kontrollera översättningskvaliteten skapade vi olika indikatorer. Vi följde upp varje indikator genom att jämföra en översättares produktion med medeltalet för alla översättare. En stor avvikelser fäster kvalitetskontrollantens uppmärksamhet på ett spe-

cifikt synset. Förutom statistiken får kvalitetskontrollanten exempel på avvikande synset sorterade i sjunkande grad av avvikelse.

De två huvudsakliga indikatorerna är antalet olika översättningar per synset, och antalet oöversatta eller identiskt översatta ord. Andra indikatorer följer översättarnas användning av annotationskoder.

Indikatorn för antalet olika översättningar per synset anger hur många synonymer det finns i ett synset på finska jämfört med antalet synonymer på engelska. Man kan använda denna indikator för att upptäcka den variationsbrist som uppstår t.ex. om en översättare använder samma eller få översättningar för alla synonymer i ett synset.

Vi tillåter att ett ord efter ett lämpligt övervägande kanske inte har någon känd finsk översättning, vilket kan signaleras genom att kopiera den engelska strängen som sådan. Vi följde användningen av denna icke-översättning med en indikator på hur många strängar som är identiska på engelska och finska. I teorin borde det vara tillämpligt endast på sådana namn som inte behöver translittereras men för ett litet förbestämt pris per ord kan en översättare inte förväntas använda många dagar på gediget terminologiarbete. I stället skall de indikera sin osäkerhet med en lämplig kod, så att vi kan återkomma till detta ställe i ett senare skede. Om procenten oöversatta ord är hög för en översättare måste kvalitetskontrollanten ge filen extra uppmärksamhet för att uppdaga orsaken.

Liknande indikatorer för hur ofta koder används i medeltal gjordes också upp. Ett överdådigt antal varianter syns i en onödigt stor användning av <or/>. För många varianter är ibland en indikation på att översättaren är osäker på den korrekta varianten och garderar sig genom att generera flera möjligheter. Det samma gäller även för onödigt liberala översättningar vilket syns i att koderna <approximate/> och <unconfirmed/> används mer än väntat.

Inom översättningsverksamhet är det praxis att kvalitetskont-

roll får ta högst 20 % av översättningstiden. För att höja hastighe-
ten på granskningen, beräknade vi processindikatorerna som me-
deltal på alla synsetsamlingar för att bestämma väntevärden för
olika koder samt för storleken på synset för engelska och finska.
Synsetsamlingar där indikatorerna avvek från medeltalet granska-
des mera utförligt. Alla synset jämfördes med medeltalet för sina
synsetsamlingar. De 500 mest avvikande synseten i en samling in-
spekterades ytterligare, men vi kunde inte upptäcka några försök
att medvetet missbruka processen. De misstag som kunde hän-
föras till enstaka tillfällen av bristande uppmärksamhet rappor-
terades som generell feedback till alla översättare.

4.4. Observationer angående den levererade översättningskvaliteten

Några generella observationer om översättningsprocessen kan
göras. Först och främst bör vi notera att översättarna var samvets-
granna och gjorde sitt bästa som professionella översättare.

I början tenderade några översättare att vara alltför försiktiga
och oftare använda koderna <unconfirmed/> och <approximate/>, men användningen minskade mot slutet. En annan strategi
för att dölja tillfällig osäkerhet i början var att hitta på opropor-
tionerligt många synonymer. För andra än egennamn, kopiera-
des mindre än 10 % av strängarna direkt. Oftast kopierades den
engelska strängen till finska för egennamn, varunamn eller andra
namngivna enheter som inte behövde translittereras till finska.

I början försökte översättarna generera tämligen kompletta
grupper av synonymer. Mot slutet av sina synsetsamlingar tende-
rade de ändå att falla tillbaka på den normala rutinen att förse
varje synonym i synsetet med bara en översättning. Vi kan bara
spekulera i orsaken till detta men en är säkert behovet att ta igen
förlorad tid om alltför mycket tid användes på början av filen.

Detta kom inte som någon överraskning eftersom vi egentligen

inte räknade med att få flera översättningar per synonym i synset. Översättarna fick betalt för att översätta varje synonym minst en gång och vi antog att detta borde ge oss en tillräcklig mängd betydelse per ord och synset.

Även om vi först nyligen har påbörjat en grundligare evaluering av resultatet, kan vi erbjuda några preliminära siffror. I tabellen inbegriper *finska* endast icke-identiska översättningar till finska (vilket utesluter ca 26 000 egennamn och andra identiskt översatta strängar som finns med i siffrorna för engelska).

Jämförelsetal	engelska	finska
synonymer (token)	206 723	181 753
ord (typer)	148 556	99 639
synonymer / ord	1,39	1,81
synonymer: eng./fin.	1,15	
ord: eng./fin.	1,50	

Tabell 5: Jämförelsetal för engelska och finska synonymer per ord

Vi tolkar siffrorna i tabell 5 på följande sätt. De 181 753 finska synonymerna utgör översättningar till 206 723 engelska synonymer. Skillnaden i antal består mest av ovan nämnda identiskt översatta engelska namn och andra översättbara ord. Endast 287 synonymer förblev helt översatta. De flesta av dem var siffror som översättningsredskapet inte tillät översättas.

De 99 639 olika finska orden översätter 148 556 olika engelska ord. De engelska siffrorna inkluderar identiskt översatta strängar. När deras antal dras av från antalet engelska ord i jämförelsen tyder siffrorna på att minskningen i lexikal variation är liten, dvs. översättarna har genererat ungefär lika många olika och specialiserade ord på finska som det fanns olika ord i den engelska förlokan. Det är mest fördelningen som är olika.

Om man jämför fördelningen av ord på synset i det engelska

WordNet med fördelningen i översättningen, var de finska översättningarna 1,3 gånger mer flertydiga än det engelska originalet med 1,81 betydelser per finskt ord i medeltal jämfört med originalets 1,39 betydelser per engelskt ord. Om man även här lägger till de ca 26 000 identiskt översatta strängarna till antalet finska ord är antalet betydelser per ord bara 1,61 på finska mot 1,39 på engelska. Man kan dra slutsatsen att något fler finska än engelska ord används i två eller flera synset.

För att ge ett exempel på den lexikala variationen på finska, jämför vi synonymerna för *wave* på engelska med synonymerna för *aalto* på finska. I WordNet är *wave* en hyponym till *motion*. Ordet *wave* får två direkta finska översättningar: *aalto* och *laine*. Bland sina WordNet-hyponymer har *wave* ord såsom *billow*, *surge*, *comber*, *fluctuation*, *ripple*, *rippling*, *riffle*, *wavelet*, *roll*, *roller*, *rolling wave*, *seiche*, *surf*, *breaker*, *swell*, *crestless wave*, *whitecap*, *white horse*. Dessa 18 engelska ord fick 10 finska översättningar (inte uppräknade här). En jämförelse med en finsk synonymordbok på nätet ger ytterligare nio synonymer. Av dessa nio förekommer sex i FinnWordNet som översättningar av ord under närliggande men andra hyperonymer: *spray*, *breeze*, *wind*, *flow*, *foam*. Man kan argumentera för att några av dem hör till begreppet 'vågor' men metonymt såsom *pärsk* 'spume'. Bland de föreslagna orden i synonymordboken hade de poetiska orden såsom *pärsky*, *lakkapä* and *vaahtopää* 'whitecap' inte alls nämnts av översättarna (*whitecap* hade översatts med *vaahtopäinen aalto* 'a white-capped wave'). Å andra sidan har WordNet översättningarna en del faktaorienterade hyponymer till *wave* som inte finns med i synonymordboken såsom *tsunami* och *backwash*.

5. Sammanfattning och fortsättning

Tidtabellen för att skapa en första version av FinnWordNet var strikt. Översättningsprocessen började i november 2009 och råöversättningen var klar i slutet av mars 2010. Det tog alltså 100 arbetsdagar. Om översättarna hade översatt med en jämn hastighet på tusen synonymer om dagen, skulle detta grovt räknat ha inneburit två heltidsöversättare. I praktiken var genomsnittshastigheten mellan 500 och 1000 betydelser per dag för en översättare.

Behovet att förbättra FinnWordNet kommer att utvärderas genom att titta på vilka finska grundformer som inte förekommer i FinnWordNet med speciell vikt på saknade högfrekventa ord. Dessutom behöver vi identifiera saknade betydelser av existerande grundformer, vilket antagligen måste upptäckas med halv-automatiska metoder i stora korpusar så som i arbetet med det polska WordNet (Piasecki m.fl. 2009). En separat studie kommer att göras manuellt för att bestämma om det finns delar av det översatta WordNet som inte har en adekvat ontologisk struktur för finska.

För att korrigera misstag i översättningarna och lägga till nya översättningar, kommer den finsk-engelska WordNet-ordboken att göras offentligt tillgänglig på nätet med en möjlighet även för allmänheten att föreslå korrigeringar och förbättringar.

Litteratur

- DanNet 2009 = *DanNet – det danske wordnet*. <http://wordnet.dk/> (mars 2010)
- Fellbaum, Christiane (red.) 1998: *WordNet: An Electronic Lexical Database*. Cambridge/London/England: The MIT Press.

- Fjeld, Ruth V. och Lars Nygaard 2009: NorNett — a monolingual wordnet of modern Norwegian. I: *NEALT Proceedings Series 7*, 13-16. <http://hdl.handle.net/10062/9837>
- Pedersen, Bolette Sandford 2010: Semantiske sproressourcer – mellem sprogteknologi og leksikografi. I: *LexicoNordica 17* (i dette bind).
- Piasecki, Maciej m.fl. 2009: *A Wordnet from the Ground Up*. Wrocław: Oficyna Wydawnicza Politechniki Wrocławskiej.
- Viberg, Åke m.fl. 2002: *The Swedish WordNet Project*. <http://www.lingfil.uu.se/personal/viberg/SwedishWordNet2.pdf> (mars 2010)
- Vossen, Piek (red.) 1998: *EuroWordNet: a multilingual database with lexical semantic networks for European Languages*. Dordrecht: Kluwer.
- Vossen, Piek 2004a: EuroWordNet: a multilingual database of autonomous and language-specific wordnets connected via an Inter-Lingual-Index. I: *International Journal of Linguistics 17(2)*, 161-173
- Vossen, Piek 2004b: Introduction. I: *Romanian Journal of Information Science and Technology 7(1-2)*, 5-6.
- WordNet 3.0 = *WordNet – A lexical database for English*. <http://wordnet.princeton.edu/> (mars 2010)

Krister Lindén
docent, FD
Helsingfors universitet
Unionsgatan 40
FIN-00014 Helsingfors universitet
krister.linden@helsinki.fi

Lauri Carlson
professor, FD
Helsingfors universitet
Unionsgatan 40
FIN-00014 Helsingfors universitet
lauri.carlson@helsinki.fi

LexicoNordica
Utgitt av
Nordisk forening for leksikografi
(NFL)
med støtte fra Ekspertgruppen
Nordens språkråd

Tidsskriftet kommer ut hvert år
i november og koster NOK 250
for ikke-medlemmer av
Nordisk forening for leksikografi

Tidsskriftet kan bestilles hos
NFL v/Rikke Hauge
Språkrådet
Postboks 8107 Dep
NO-0032 Oslo
Tlf. +47 22 54 19 73
e-post: rikke.hauge@sprakradet.no

Dette nummer av LexicoNordica har leksikografi og språkteknologi i Norden som hovedtema. Temaet er svært aktuelt i moderne leksikografisk praksis og teoriutvikling, både med hensyn til valg av hjelpemidler i redigeringen av ordbøker og til leksikalsk beskrivelse i språkteknologiske programmer. Artikkelen i dette nummer dekker begge temaene og gir en oversikt over status for denne fagutviklingen i de forskjellige nordiske landene. I tillegg inneholder tidsskriftet noen ikke-tematiske artikler og en rekke anmeldelser av utgitte ordbøker og andre leksikografiske produkter.

Bøker man ønsker anmeldt i tidsskriftet, sendes til til en av hovedredaktørene eller til en nasjonal redaktør.