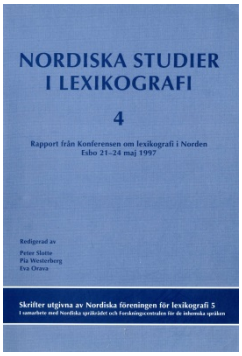
NORDISKE STUDIER I LEKSIKOGRAFI

Titel:	ERLEKS: en estnisk-svensk lexikalisk databas
Forfatter:	Signe Cousins
Kilde:	Nordiska Studier i Lexikografi 4, 1997, s. 37-44 Rapport från Konferens om lexikografi i Norden, Esbo 21.-24. maj 1997
URL:	http://ojs.statsbiblioteket.dk/index.php/nsil/issue/archive



Betingelser for brug af denne artikel
Denne artikel er omfattet af ophavsretsloven, og der må citeres fra den. Følgende betingelser skal dog være opfyldt:

- Citatet skal være i overensstemmelse med „god skik“
- Der må kun citeres „i det omfang, som betinges af formålet“
- Ophavsmanden til teksten skal krediteres, og kilden skal angives, jf. ovenstående bibliografiske oplysninger.

Søgbarhed
Artiklerne i de ældre Nordiske studier i leksikografi (1-5) er skannet og OCR-behandlet. OCR står for ’optical character recognition’ og kan ved tegngenkendelse konvertere et billede til tekst. Dermed kan man søge i teksten. Imidlertid kan der opstå fejl i tegngenkendelsen, og når man søger på fx navne, skal man være forberedt på at søgningen ikke er 100 % pålidelig.

Signe Cousins

ERLEKS: en estnisk-svensk lexikalisk databas

There is an Estonian-Swedish parallel corpus currently under construction at the Department of Scandinavian Studies, University of Tartu. It is being built according to the example of some already existing translation corpora. The corpus was first and foremost meant as a completion to a new and comprehensive Swedish-Estonian dictionary, but there have appeared other fields of application and research in connection with it: an authentic lexical source material for different glossaries and dictionaries, an active study aid for the students of Swedish, a source of research material for the study of collocations and constructions as well as of differences in sentence construction and word order of the two languages.

Bakgrund

Förberedelser för ett forskningsprojekt med syfte att ställa samman **en estnisk-svensk parallellkorpus** började i september 1994 med att avdelningen för skandinaviska språk och litteratur vid Tartu universitet fick en svensk gästprofessor, Stig Örjan Ohlsson, och jag som nybörjare på magisternivå i svenska under hans handledning fick korpusarbetet som uppgift för min magisterexamen. Idén till ett sådant för Estland rätt så innovativt tema fick vi från Martin Gellerstams artikel om produktionsordböcker (Gellerstam 1994:48). Först behövde vi sätta oss in i teorin, för datalingvistisk kunskap var då ingen självklarhet för en lingvist här.

I mars 1995, i och med att projektet blev godkänt för finansiering av Magnus Bergwalls stiftelse, började det praktiska arbetet. Under tiden hade även en annan projektidé mognat vid avdelningen; behovet av **en ny omfattande svensk-estnisk ordbok** gjorde sig känt mer och mer. Projektansvarig är prof. Ohlsson, och sex lexikografer har skrivit var sin del av ordboken. I september 1997 var det första skedet så gott som färdigt och redigeringen har påbörjats. De två projekten tillsammans verkar numera under namnet ERLEKS – *Eesti-Rootsi LEKSikaalne Andmebaas* (Estnisk-svensk lexikalisk databas). Arbetet har möjliggjorts med materiellt stöd från Humanistisk-Samhällsvetenskapliga Forskningsrådet.

Korpusbeskrivning

Vi har använt begreppet **parallellkorpus** i meningen av en **översättningskorpus**, alltså har vi inte (än) tagit med texter med likartad struktur och liknande innehåll. Antalet estniska texter med översättningar till svenska och antalet svenska texter med översättnigar till estniska har vi försökt hålla lika stort. De största **textkategorierna** är skönlitteratur och populärvetenskap, mest historia. Det har redan kommit ut ett antal historieböcker som är parallelltexter från början, så att vi har fått både originalet och översättningen med detsamma.

I några fall har vi haft svårigheter med att avgöra vad som är original och vad som är översättning, för några författare är exilester eller Sverigeester som ju är mer eller mindre bilingvala

och skriver på båda språken. Dessa texter utgör en intressant källa för jämförelser mellan Sverigeesternas estniska och estniskan som den talas och skrivs nu, eller talades eller skrevs för några tiotal år sedan i Estland.

Det är ju inte alltid man får en maskinläsbar text från ett förlag. En del material behövde läsas in i dator med hjälp av scanning och efteråt tog genomläsningen och rättandet av fel sin tid. Maskinläsbara versioner av både original och översättningar av några skönlitterära böcker fick vi med benäget tillstånd från korpusen Studia Comparativa Linguarum Orbis Maris Baltici i Åbo, Finland.

En framtida källa för material med stilvariationer kan vara en tidskrift som heter "Ronor" och kommer ut på båda språken i Nuckö, ett gammalt svenskt område i nordvästra Estland.

Taggning och ihopkoppling

Textfiler ska vara i ASCII-format, tankstreck och språkspecifika bokstäver som 'å', 'õ', 'ä', 'ö' och 'ü' behöver ersättas med koder för att texterna skall kunna användas i olika program, t.ex. 'õ' blir 'õ'. Även tryck- och andra fel behövde kodas med motsvarande taggar som innehåller den rätta versionen.

Texterna är taggade enligt Text Encoding Initiative (TEI) bestämmelser med hjälp av Standard Generalized Markup Language (SGML) (Sperberg–McQueen & Burnard 1994). TEI valdes eftersom det verkade vara ett flexibelt system med ett brett urval av taggar och även eftersom de skandinaviska korpuslag vi har kontakt med använder samma system, vilket gör det möjligt att byta texter mellan olika korpusar. Varje text har en s.k. **TEI-header** som innehåller allmän information om publiceringen, språk och textklass samt om den redigerings-, taggnings- och ihopkopplingsprocess texten i fråga har genomgått.

Taggar som kan sättas in automatiskt är t.ex. <s>-taggar som markerar gränserna på en **S-enhet** (en ortografisk sats) och de sätts in automatiskt med hjälp av ett dataprogram skrivet av Knut Hofland, Humanistisk datacentrum i Bergen, Norge (Johansson, Ebeling & Hofland 1996). Vi har tillhört ett nordiskt nätverk Språk i kontrast och lärt oss mycket av korpusverksamheten vid Universitetet i Oslo och Lunds universitet. Därför är vi även tacksamma användare av flera datalingvistiska program (med funktioner för taggning, ihopkoppling och sökning) som ursprungligen var utarbetade för den engelsk-norska korpusen och i några fall blev modifierade för den estnisk-svenska korpusen. Ändå har mycket möda lagts ner på handarbete.

Det finns olika principer för ihopkoppling (*eng.* alignment) av texter i en parallellkorpus. Vår är ihopkopplad på satsnivå, men arbete är på gång för att göra det på ordnivå.

Ett kopplingsprogram tar som sin utgångspunkt en s.k. **ankarlista**, alltså ett speciellt lexikon. Dåvarande svenskstuderanden, numera studeranden på magisternivå Mari Aidla hjälpte mig sätta ihop ett svenskt-estniskt basordförråd där de estniska orden är **trunkerade**, d.v.s. böjnings- och avledningsändelser är borttagna och ersatta med en asterisk, för om man skulle ta med alla olika ordformer och sammansatta ord i estniskan skulle ankarlistan bli oanvändbart stor.

Ett exempel på en del av ankarlista:

kamp* / võitlus*, heitlus*
kamrat* / kaasla*, sõber*, sõbra*
kan / oska*, suut*
kanal* kanal*

kanske / võib-olla, vahest, ehk
kapitel*, kapitel* / peatükk*, peatüki*
kapitulera* kapituleeru*
kappa*, kappor* / mantel*, mantli*, palitu*

Snedstreck används för att skilja åt de svenska och estniska ordstammarna, om det finns fler än en i något av språken. Ofta stämmer trunkeringen inte med morfemgränser utan är endast praktiskt motiverad. På grund av de komplicerade böjningsreglerna i estniskan blev de flesta estniska ekvivalenter trunkerade och ett svenskt ord har ofta flera estniska ekvivalenter, vilket betyder att fall av ett till ett-betydelseekvivalenter är mycket vanligare mellan norska och engelska än mellan svenska och estniska. Resultatet blev en förhållandevis lång ankarlista vilket ibland påverkar arbetshastigheten och framgångsgraden hos kopplingsprogrammet.

Varje S-enhet får en förbindelselänk till motsvarande S-enhet i den parallella texten. Programmet flyttar ett fönster med 15 S-enheter genom texten och försöker matcha en svensk sats med en estnisk. Det tar hänsyn till gemensamma ankarord och antalet bokstäver i satsen, räknar ut ett ankarpoängvärde samt antalet gemensamma ankarord, och hittar den bästa motsvarigheten. Andra kännetecken som programmet tittar på är fråge- och utropstecken, egennamn och markering av framhävda textpartier, för sådana saker finns oftast kvar även i parallelltexten. Resultatet blir en matris av ett textfönster och ihopkopplade S-enheter. En diagonal genom matrisen maximerar summan av gemensamma enheter i ankarlistan.

Ett exempel på utmatningen från kopplingsprogrammet: en matris och några ihopkopplade S-enheter.

		115	100	40	80	88	147	31	55	97	68	57	106
		1	2	3	4	5	6	7	8	9	10	11	12
1	154 I	6	1	0	2	1	2	0	3	4	1	1	0
2	210 I	2	13	1	5	1	4	0	3	5	2	2	0
3	164 I	2	4	1	6	9	4	1	3	4	2	2	1
4	216 I	2	2	1	1	3	15	5	1	3	2	2	1
5	156 I	3	7	0	5	2	4	0	8	5	3	1	0
6	118 I	2	5	0	3	2	4	1	3	10	2	1	0
7	75 I	1	1	0	1	1	2	0	0	2	6	1	0
8	74 I	0	1	0	2	1	2	0	1	0	1	3	0
9	112 I	0	0	1	0	1	2	1	0	0	0	0	5
10	201 I	2	2	0	0	1	2	0	0	1	1	2	0
11	167 I	2	1	0	0	2	4	0	0	0	2	2	0
12	60 I	2	0	0	0	1	1	0	1	0	1	1	0

Sum=105/0.93: 1,1 2,2+3 3,4+5 4,6+7 5,8 6,9 7,10

1: <s>När man däremot på 1670-talet uppförde ett nytt rådhus i staden Kuressaare följde man i stort de ritningsförslag som tillkommit på De la Gardies initiativ.</s></p> (HL1.19.30)

1: <s>Kuressaare uue raekoja ehitamisel 1670-ndatel aastatel järgiti seevastu põhiliselt De la Gardie'lt tulnud jooniseid.</s></p> (HL1T.19.31)

2: <pb n=69> <omit desc=photo resp=tag> </div1> <div1 type=part id=HL1.20>

<head>Slottet i Haapsalu</head> <pb n=70> <p><s>Jacob De la Gardie var en av de svenska adelsmän som verkligen fick möjlighet att ånjuta den kungliga generositeten vad gällde förvärv och förläningar av gods i Östersjöprovinserna.</s> (HL1.20.1)

2: <pb n=69> <omit desc=photo resp=tag> </div1> <div1 type=part id=HL1T.20> <head>Haapsalu loss</head> <pb n=70> <p><s>Jacob De la Gardie'l oli meeldiv võimalus nautida kuninglikku suuremeelsust.</s> (HL1T.20.1)
3: <s>Ta omandas Läänemereprovintsides mõisaid.</s> (HL1T.20.2)

3: <s>Utöver att Jacob De la Gardie var befälhavare var han även en duktig affärsman som under de krig han deltog i, mot Ryssland och Polen, skapade sig en stor förmögenhet.</s> (HL1.20.2)

4: <s>Peale selle, et Jacob De la Gardie oli väejuht, oli tal ka majanduslikku mõtlemist.</s> (HL1T.20.3)
5: <s>Ta võttis osa sõdadest Venemaa ja Poola vastu ning kogus sel ajal endale suure varanduse.</s> (HL1T.20.4)

4: <s>År 1623 köpte han med kungligt tillstånd hela ön Dagö och i november 1625 förskaffade han sig kungens försäkringsbrev på att få förvärva staden Haapsalu med omkringliggande landskap för vilket han betalde 66.850 daler.</s> (HL1.20.3)

6: <s>1623. aastal ostis ta Hiiumaa ja novembris 1625 muretses ta endale kuninga toetuskirja, et saada Haapsalu linna ja selle ümbruses oleva maa omanikuks.</s> (HL1T.20.5)
7: <s>Ta ostis maa 66 850 taalri eest.</s> (HL1T.20.6)

Svenska satser jämförs med den svenska ankarfilen och estniska satser med den estniska för att hitta nya ord till ankarlistan. Ifall en S-enhet är mycket kort och har ett litet antal ord kan det orsaka problem, för antalet ordmotsvarigheter förblir litet. Även när översättningen har varit tämligen fri är antalet ordmotsvarigheter obetydlig, och när det gäller sammansatta satser stämmer inte satslängden i bokstäver, vilket ibland får felaktiga ihopkopplingar som följd. Programmet testar längden av den sammansatta satsen mot målsatsen (och godkänner skillnader upp till 20 %).

Utmatningen blir två ihopkopplade texter med S-enheter som har tecken för 'id' (identifikation) och 'link' (identifikation av den motsvarande S-enheten i parallelltexten).

Sättet hur kopplingsprogrammet fungerar kan faktiskt föra fram fall där översättaren har valt en mindre direkt översättningsmöjlighet och på så sätt kasta ljus över hur översättare resonerar.

Statistiska undersökningar

Paketet innehåller även ett enkelt statistikprogram av Knut Hofland (Johansson & Ebeling 1994) vilket räknar antalet förekomster av ett ord i texten, antalet S-enheter där ordet förekommer, antalet förekomster av ordet i två eller i tre S-enheter i rad. En enkel statistisk undersökning har genomförts på korpusen vad gäller jämförelse av antalet stycken i båda språken, antalet ortografiska satser, satser per stycke, antalet ord, ord per sats, bokstäver per sats och bokstäver per ord.

Både i svenska originaltexter och översättningar till svenska överskrider antalet ord detta i estniska texter. Det är då rätt logiskt att den genomsnittliga svenska satsen innehåller fler ord än

den estniska, men procenttalet är överraskande högt – svenska satser innehåller 33,62 % fler ord än estniska satser. Estniska ord visar sig vara 27,11 % längre än de svenska. Jämförelsen av antalet S-enheter visar tendensen att estniska texter innehåller lite fler.

Morfoanalysator

Morfologiska taggar sätts in i texter med hjälp av ESTMORF, en morfologisk analysator som var programmerad för den estniska skriftspråkskorpusen vid Tartu universitet av Heiki-Jaan Kaalep (Kaalep 1996). Några exempel på dess utmatning:

kui	kui+0 //_D_ // kui+0 //_J_ //	– om
Eesti	Eesti+0 //_H_ sg g, sg n, //	– estniska
poolel,	pool+l //_N_ sg ad, // pool+l //_S_ sg ad, //	– (på) sidan

Programmet hittar lemman, delar upp sammansättningar och ger alla möjliga morfologiska tolkningar och markeringar för ett ord, vilket leder till disambiguationskrav. Det finns två datalingvistikstuderande i den estniska korpusgruppen som arbetar med att lösa disambiguationsfrågor.

En annan morfoanalysator för estniska har utarbetats av Lingsoft OY; dessutom har de en morfoanalysator för svenska språket och håller på med en svensk syntaxanalysator, som alla kan komma till användning vid parallellkorpusen, om de första försöken att köra programmen på korpusen är framgångsrika.

Konkordanser

En av de första saker som även en begränsad korpus kan användas för är konkordanser. Eftersom korpusprojektet löper samtidigt som ordboksprojektet, kan ordboken inte helt baseras på korpusen, men den kan fungera som en källa för levande exempel och ge bevis på om ett estniskt ord eller fras verkligen används som ekvivalent till ett visst svenskt ord eller fras. Andra användningsområden är undersökning av kollokationer och bruksanvisningar, samt valensuppgifter.

Konkordanser har gjorts om:

- svenska ord 'ta', 'tar', 'tog', 'tagit' och den estniska direkta motsvarigheten 'võtma' med sina konjugationer (alltså den trunkerade formen 'võt*');
- svenskt 'fick' och estniskt 'sai';
- svenskt 'var' och estniskt 'olid' (tredje person pluralis);
- konjunktioner: estniska 'ja', 'ning', 'kuid', 'ega', 'ehk', 'aga' och svenska 'och', 'men', 'eller'

Det visar sig i ordmotsvarighetsstudierna att i bara 52,6 % fall hade 'ta', 'tar', 'tog', 'tagit' översatts med någon form av den direkta estniska motsvarigheten 'võtma', medan 'võtma' i bara 61,2 % fall hade översatts med en form av 'ta'. Detta tyder på en mängd av kollokationer och idiomatiska uttryck som inte kan översättas direkt.

Satsstruktur och ordföljd

En orsak till att vi började undersöka skillnader i satsstruktur var ihopkopplingsfel som programmet gjorde, för flera av felen berodde på olikt antal satser i originaltexter och översättningar. Antalet fall där en svensk sats motsvaras av två estniska är 2,2 gånger större än omvänt. I några fall motsvaras en svensk sats av tre estniska, och ytterst sällan sker det omvända – som ni kommer att se snart. Svenskan verkar bruka samordnade eller underordnade satser mycket oftare där det i den estniska texten börjar en ny sats. Ofta finns det ett komma eller semikolon i den svenska texten där estniskan sätter punkt och börjar en ny mening.

Några tydligare exempel på olikheter i svensk och estnisk satsstruktur och ordföljd enligt korpustexter:

Originalet är estniska:

[1]Siinkohal<rumsadverbial> ingli käsi<subjekt> peatub<finit verb>, ta vajub mõttesse ning lisab siiski: "Vägivald armastab vabadust, tahab ta võita ja allutada..." ning<konjunktion> [2]jääb<finit verb> suure huviga<sättsadverbial> ootama<infinit verb>, millal ühe endise väikese Balti Riigi luuletaja<subjekt> [3](kelle<personlig pronomen i genitiv> juuksed<subjekt> on võrdlemisi pikad ja kelle raadio on nii rikkis, et ta ei kuule vaba Lääne jaamadest muud kui ainult pahaendelist raginat) [2 forts.]need read<objekt> õhust üles leiab<finit verb> ning paberile kirjutab.

1.

Översättningen är svenska:

[1]Här<rumsadverbial> hejdar sig<finit verb> ängelns hand<subjekt>, han försjunker i tankar, men lägger ändå till: "Våldet älskar friheten, vill besegra och kuva den..."
[2]Med stort intresse<sättsadverbial> avvaktar<finit verb> han<subjekt> när en poet<subjekt> från en av de före detta små baltiska länderna fångar<finit verb> just dessa rader<objekt> och sätter dem<personligt pronomen i objektsform> på pränt.
[3](Poetens hår<subjekt> är ganska långt och hans radio är så trasig, att han bara hör ett illavarslande skrällande från sändarna i det fria Väst.)

2.

Originalet är svenska:

[1]Det<formellt subjekt> spratt<finit verb> till<verbpartikel> i honom<egentligt subjekt> av fröjd<adverbial>, [2]när<konjunktion> han<subjekt> kände<finit verb> myntet<objekt> mellan fingrarna<rumsadverbial>, [3]vad en femöring var stor och präktig i alla fall!

Översättningen är estniska:

[2]Münti<objekt> sõrmede vahel<rumsadverbial> tundes<infinit verb> [1]ta<subjekt> lausa võpatas<finit verb> rõõmust<adverbial>.
[3]Küll on viieööriline ikka suur ja tore!

ren som "l'ensemble des informations ordonnées de chaque article", hvilket for den samlede mængde artikler i ordbogen giver et "programme d'information constant", altså en ensartet struktur, der anvendes på hver artikel.

2.2 Undersøgelse af mikrostrukturen i fagsproglige publikationer

Vi har undersøgt nogle fagsprogsordbøger og leksikografiske specialer ved HHÅ med henblik på at vurdere, om de imødekommer den moderne fagoversætters behov, der kort kan sammenfattes således: ordbøger der muliggør hurtig og effektiv tilgang til relevante oplysninger. Mikrostrukturen i publikationerne varierer meget, hvilket naturligvis bl.a. hænger sammen med deres forskellige funktioner. En nærmere analyse af mikrostrukturerne og deres opbygning viser, at de leksikografiske principper, der – bevidst eller ubevidst – er anvendt, i nogle tilfælde er et tilfældighedsprincip og et princip om at ville variere eller underholde ud fra den betragtning, at det kan brugeren sikkert få megen fornøjelse af. Vi mener, at denne praksis er uhensigtsmæssig for den moderne fagoversætter, der søger præcise informationer og ikke tilfældige oplysninger, der kunne være alment dannende.

Vi vil i det følgende ved konkrete eksempler illustrere denne praksis, idet vi har valgt at fokusere på følgende oplysningskategorier:

* definition/forklaring
* encyklopædiske oplysninger
* eksempler
* kollokationer

På basis af de principper, vi har iagttaget, har vi opstillet en række teser, som danner grundlag for vores argumentation. Nogle af teserne er vi uenige i, andre derimod anbefaler vi. Teser vi **ikke** kan anbefale er markeret med asterisk.

3 Strukturering af oplysningsfelter

Bergenholtz m.fl. (1997:§ 23) nævner, at det kan være hensigtsmæssigt at opdele ordbogsartiklen i felter, og det vil vi naturligvis også anbefale ud fra førnævnte betragtning om hurtig tilgang til oplysningerne.

3.1 Eksempler og kollokationer

En af de opdelinger, vi finder væsentlige, er den, der vedrører kollokationer og eksempler. Vi mener, der er tale om to oplysningskategorier, og vi mener, at de skal placeres i hver sit oplysningsfelt. Det er der ikke en entydig tradition for i leksikografien, hverken i den almensproglige eller i den fagsproglige, og det skal vi komme ind på senere.

Afgrænsningen mellem kollokationer og eksempler er uklar. Dette skyldes bl.a., at der er uenighed blandt forskerne om definitionen af begrebet 'kollokation'. Den mest liberale af de definitioner, der foreligger, er, at en kollokation er "forekomsten af to eller flere ord, der indgår i en

syntaktisk enhed", jf. Bergenholtz & Tarp (1994). Vi vil indledningsvis konvertere denne definition til vores første tese:

***Tese 1:**	**Kollokationer er enhver kombination af ord, der indgår i en syntaktisk enhed**

I sætningen "manden i båden drak en øl" vil således hverken "manden i" eller "båden drak" være kollokationer, men derimod nok "manden i båden" og "drak en øl". Men også hele sætningen må ifølge ovennævnte definition være en kollokation. Vi mener, der er behov for en afgrænsning.

Den manglende skelnen mellem ordforbindelser af forskellig art er typisk for mange ordbøger, såvel monolingvale som bilingvale, der præsenterer lemmata hhv. ækvivalenter i eksempelsætninger og syntagmer/dele af sætninger i flæng.

De følgende 2 eksempler er taget fra en dansk monolingval ordbog:

(3)

disci'plin *-en,* (i betydn. 1) *-er:* – **1.** videnskabsgren, fag *geometri er en matematisk d.* – **2.** opretholdelse af orden og lydighed *der herskede en stram d. i hæren; holde d. i skolen* – [fra latin *disciplina* afl. af *discipulus,* se *discipel*] – **disciplin-** 'nar- i sammensætn. – [afl. af *disciplin 2*] – **disciplinarmiddel** el. **disciplinær-middel** – **disci'pli'nere** *-ede:* holde under disciplin; især i perf.-partc. *mandskabet er i enhver henseende disciplineret* – [afl. af *disciplin 2*] – **disci'pli'nær** adj.: *han blev straffet for en d. forseelse* ɔ: en forseelse mod disciplinen, en ulydighed; *han fik en d. straf* ɔ: straf idømt af en overordnet for brud på disciplinen – [afl. af *disciplin 2*] – **disciplinær-straf**

(4)

rist I *-en, -e: stege kød på r.; risten i en kakkelovn; tør fødderne af på risten foran havedøren* – [fællesnord.: oldnord. *rist;* besl.m. tysk *rost* rist, gitter, og m. eng. *roast* stege, jf. *roastbeef*]
rist II subst.: hvile; kun i forbind. *r. eller ro: hun havde hverken r. eller ro, før hun fik sin vilje* – [fra plattysk *rist;* besl.m. *rast*]

I de to udvalgte artikler ses den nævnte tendens til variation i ordenes kontekst – hele sætninger, infinitte konstruktioner og substantivsyntagmer – uden at pointen med det specifikke valg af kontekster fremgår. Næste eksempel er fra en bilingval fagsprogsordbog – en engelsk-spansk/spansk-engelsk handelsordbog. Også her ser vi kontekstuelle variationer. Substantivsyntagmer, infinitte verbalkonstruktioner og hele sætninger, der tilsyneladende optræder i flæng.

(5)

error *noun* error *m or* equivocación *f;* **he made an error in calculating the total** = cometió un error al hacer la suma; **the secretary must have made a typing error** = la secretaria debe de haber cometido un error de mecanografía; **clerical error** = error de copia *or* error de oficina; **computer error** = error de ordenador; **margin of error** = margen de error; **errors and omissions excepted** = salvo error u omisión; **error rate** = coeficiente de errores; **in error** *or* **by error** = por error; **the letter was sent to the London office in error** = la carta fue enviada a la oficina de Londres por error

Det måste komma ett verb efter fundamentet i svenskan, så verbet kommer ofta före subjektet, men i estniskan brukar subjektet komma före verbet i huvudsatsen, även om satsen börjar med någon annan satsdel. I estniskan kan objektet komma före det finita verbet, medan det skulle se ytterst onaturligt ut i svenskan. Flera olikheter beror på att svenskan har subjektstvång, men inte estniskan – vi har ingen motsvarighet till det svenska formella subjektet, så det förekommer satser utan något klart uttryckt subjekt alls – det underförstås i predikatet. Andra omständigheter som kan orsaka stora skillnader i ordföljden och ge översättningsproblem är placeringen av adverbial, aktiv och passiv form av verbet, samt negationer. Allmänt sett har estniskan friare ordföljd än svenskan, för vi kan visa relationer mellan satsdelar med hjälp av fler böjningsformer.

Eftersom korpusen innehåller både skönlitteratur och facklitteratur har jämförelser tagits fram mellan formell och informell satsstruktur i estniska och svenska.

Utbyggning och framtida användning

Som läget är ligger den första versionen av den svensk-estniska ordboken maskinläsbar i datafiler, men den har inte satts in i någon riktig databas. Frågan prioriteras, särskilt eftersom vi har börjat ett nätverksprojekt med Uppsala universitet där man har hållit på med att redigera en estnisk-svensk ordbok under en tid nu. Vi vill försöka vända på vår ordbok och se om den estnisk-svenska behöver kompletteras.

En annan idé är att koppla ihop ordboksdatabasen med korpusen, så att om man slår upp ett svenskt eller estniskt ord eller eventuellt fras, kan man förutom informationen i ordboksartikeln få fram fraser, satser och stycken ur korpusen där ordet eller frasen finns, köra analysatorer på det man hittar och få fram böjningsformer och valensuppgifter.

För övrigt behöver studerande i lingvistik och översättning en forskningsbas. Beroende av vilka texter som kommer att läggas till i korpusen kan olika språkvariationer undersökas.

Litteratur

Gellerstam, Martin 1994: Produktionsordböcker – vad är det? I: Henning Bergenholtz (red.), *LexicoNordica 1–1994. Tidsskrift om leksikografi i Norden utgitt av Nordisk forening for leksikografi i samarbeid med Nordisk språksekretariat.* Oslo: Nordisk forening for leksikografi, 43–52

Johansson, Stig/Jarle Ebeling 1994: *The English-Norwegian Parallel Corpus: Introduction and Applications.* Paper submitted to The XXVIII International Conference on Cross-Language Studies and Contrtastive Linguistics. 15–17 December 1994, Rydzyna, Poland.

Johansson, Stig/Jarle Ebeling & Knut Hofland 1996: *Coding and aligning the English-Norwegian parallel corpus.* I: Karin Aijmer o.a. (red.): *Languages in Contrast. Papers from a Symposium on Text-based Cross-linguistic Studies.* I: Sven Bäckman & Jan Svartvik (red.): *Lund Studies in English* 88. Lund University Press, 87–112.

Kaalep, Heiki-Jaan 1996: *ESTMORF: A Morphological Analyser for Estonian.* I: Haldur Õim (red.) *Estonian in the Changing World.* Tartu

Sperberg-McQueen, Michael C.M./Lou Burnard 1994: *Guidelines for Electronic Text Encoding and Interchange.* TEI P3. Chicago/Oxford: Association for Computers and the Humanities/Association for Computational Linguistics/ Association for Literary and Linguistic Computing (electronic version)