Peter Widell og Ulf Dalvad Berthelsen (udg.):
11. Møde om Udforskningen af Dansk Sprog
Århus 2006

INTERCORP – PARALLELT DANSK-TJEKKISK KORPUS

Af Kateřina Haušildová (Aarhus Universitet)

1. Indledning

Korpuslingvistikken, som er sprogvidenskabens spædbarn, har oplevet en svimlende udvikling i de seneste år. Siden de første usikre skridt i 60'erne, hvor tjekken Henry Kučera og amerikaneren Nelson Francis byggede et computerbaseret korpus over moderne amerikansk engelsk[1], *Brown Corpus* (efter Brown University i USA), har korpuslingvistikken taget et stort spring mod avancerede og multifacetterede elektroniske tekstkorpora med en bred vifte af anvendelsesmuligheder.

2. Begrebet *korpus*

En hurtig udvikling af korpuslingvistikken har selvsagt medført etablering af et specifikt terminologisk apparatus med *korpus* som det centrale begreb. I forhold til en generel forståelse af begrebet korpus som en hvilken som helst samling af tekster, hvad enten den er elektronisk eller trykt, definerer korpuslingvistikken *korpus* som en „afgrænset mængde af tekster som er indsamlet efter klart definerede kriterier, beskrevet efter en given standard og som er tilgængelige i elektronisk form." (Kirchmeier-Andersen 2002:12)

3. Opdeling af tekstkorpora

Tekstkorpora kan variere både i størrelse og indhold afhængigt af, hvilke formål korpusset skal tjene og hvilke undersøgelser man ønsker at foretage.

Man kan således skelne mellem *specielle tekstkorpora*, der opfylder visse forudbestemte og meget specifikke kriterier, fx. korpus over en bestemt dialekt eller Shakespeares værker, og *almene tekstkorpora*, der på den anden side skal dække hele sprogområdet, herunder diverse genrer og tekstyper, for at gøre det muligt at beskrive sproget i dets helhed.

Det aldersmæssige kriterium gør sig gældende ved opbygning af synkrone og diakrone korpora. Mens de *synkrone tekstkorpora* viser nutidens sprog og således tjener som en vigtig

[1] Kucera, H., and Francis, W.N. Computational Analysis of Present-Day American English. Brown University Press, Providence, R.I.. 1967

informationskilde om forskellige igangværende sproglige fænomener og processer, afspejler de *diakrone korpora* sprogets tidligere stadier, og deres omfang er som regel betydeligt mindre end de synkrone korporas på grund af en møjsommelig digitalisering.

De fleste tekstkorpora er *skriftsprogskorpora* bestående af elektroniske tekstuddrag fra bøger, aviser, blade og andre skriftlige tekster. *Talesprogskorpora* indeholder derimod transkriberet talesprog, og netop den besværlige transkription har indtil videre gjort sig skyldig i mindre udbredelse af talesprogskorpora til trods for talesprogets ubestridte primat som kommunikationsmiddel. Udvikling af nye programmer til automatisk transkription af talesproget (et sådant program for tjekkisk bliver aktuelt testet ved Det Tekniske Universitet i Liberec, Tjekkiet) vil imidlertid uden tvivl styrke talesprogets position inden for korpuslingvistikken.

Et vigtigt potentiale for oversættervirksomhed, fremmedsprogspædagogik, udvikling af værktøjer til maskinstøttet oversættelse og kontrastiv lingvistisk forskning ligger i anvendelse af parallelle tekstkorpora. Disse korpora indeholder indholdsmæssigt identiske tekster forfattede på forskellige sprog (parallelle tekster, dvs. tekster og deres oversættelser til andre sprog).

4. Standarder for beskrivelse af tekster

Det fremgår af ovenstående, at kendskab til kriterierne for sammensætning af korpora og mulighed for stor præcisering af søgekriterierne under arbejdet med korpusset er afgørende for, hvorvidt en undersøgelses resultat kan betragtes som repræsentativt, pålideligt og validt. Af denne grund er råteksten forsynet med yderligere information, den ekstra-lingvistiske og den lingvistiske opmærkning. Den *ekstra-lingvistiske opmærkning*, også kaldt "header", indeholder blandt andet oplysninger om tekstens omfang, oprindelse, indhold, medium, sprog osv. Den *lingvistiske opmærkning* eller "tags" rummer information, "som knytter sig til de enkelte enheder i teksten" (Kirchmeier-Andersen 2002:15), dvs. oplysninger om ordklasser, syntagmegrænser, semantiske informationer osv.

5. Det Tjekkiske Nationale Korpus

På trods af Henry Kučeras store indsats på området slår den tjekkiske korpuslingvistik ikke igennem som et selvstændigt forskningsområde før i 1994, hvor der oprettes Institut for Det Tjekkiske Nationale Korpus ved Karlsuniversitetets Filosofiske Fakultet i Prag. Siden oprettelsen har instituttet været beskæftiget med opbygning og udvikling af diverse korpora, herunder det almene korpus SYN2000 med 100 mio. ord, der tilsammen udgør Det Tjekkiske

Nationale Korpus, og med andre korpusrelaterede aktiviteter som undervisning og dyrkning af faget *korpuslingvistik*.

6. Projekt *Intercorp*

Et af delprojekterne afviklet under Det Tjekkiske Nationale Korpus hedder *Intercorp*. Projektet, der blev bevilliget af det tjekkiske Ministerium for Skolevæsenet, Ungdom og Sport som MSM 0021620823, blev sat i gang i januar 2005 med to overordnede ambitiøse målsætninger: opbygning af parallelle synkrone tekstkorpora med tekstmateriale fra 25 sprog[2], heriblandt dansk, med tjekkisk i centrum (som atomkerne omgivet af elektroner) og efterfølgende korpusbaseret sammenlignende lingvistisk forskning[3] med fokus på kvantitativ undersøgelse af sproglige fænomener i parallelle tekster og identifikation af centrum og periferi. Projektet afvikles i to faser, en konstruktionsfase, hvor der fokuseres på indsamling og digitalisering af skønlitterære, samfundsvidenskabelige og naturvidenskabelige tekster fra perioden efter 1950 og deres fremmedsprogede udgaver, og en forskningsfase. Forståeligt nok overlapper begge faser hinanden, idet opbygning af korpora er en langtrukket og i princippet uendelig proces. Ifølge planen skal korpusset indeholde mindst 2,5 mio. ord inden 2010, og i 2011 afsluttes projektet med en konference om parallelle tekstkorpora, deres metodik og anvendelsesmuligheder. Uafhængigt af projektets enkelte faser udvikles og testes løbende søgeværktøjet Bonito 2, der skal foreligge i løbet af 2007.

7. Et parallelt korpus bliver til

Opbygning af parallelle tekstkorpora sker under ledelse af eksperter fra filologiske fag med tilknytning til Det Filosofiske Fakultet. De enkelte koordinatorer har overordnet ansvar for anskaffelse, digitalisering og alignering af tekster, og de skal sørge for en prompte opdatering af den fælles tekstdatabase med ekstra-lingvistiske informationer om tekster og med den aktuelle status for tekstbearbejdningen. Koordinatorerne har også mulighed for at hyre studenter eller kolleger til projektarbejdet, primært til digitalisering og alignering af tekster, der aflønnes pr. kB i henhold til præcist definerede kriterier.

[2] Følgende sprog skal være repræsenteret i korpusset: engelsk, tysk, fransk, russisk, slovakisk, spansk, italiensk, arabisk, bulgarsk, dansk, finsk, litauisk, ungarsk, makedonsk, norsk, polsk, portugisisk, slovensk, svensk og evt. japansk og kinesisk

[3] Som *tertium comparationis* anvendes tjekkisk som netværkets kerne.

7.1 Anskaffelse af tekster

Det er projektledelsens bestræbelse at anskaffe elektroniske tekster frem for tekster i papirform, fordi en manuel digitalisering af teksternes papirudgaver er en tidskrævende og møjsommelig proces, der sker på bekostning af korpussets samlede størrelse. Der er blevet identificeret en række elektroniske kilder med bl.a. juridiske tekster (EUR-Lex og Korpus Acquis Communautaire) samt elektroniske biblioteker som eksempelvis The Oxford Text Archive, mulighederne er imidlertid begrænsede. Årsager dertil er af forskellig art, og jeg vil kun nævne de vigtigste:

Et alment synkront korpus skal indeholde vidt forskellige teksttyper og genrer, men dette kriterium er praktisk umuligt at imødekomme, især når det gælder små sprog som dansk og tjekkisk. EU's tekstdatabaser ligner ganske vist et overdådigt tag-selv-bord, men det er begrænset, hvor mange juridiske tekster der kan indgå i et alment korpus. Et andet problem er oversættelsernes sproglige og faglige kvalitet. Langt de fleste internetsider foreligger på engelsk og tysk, de største virksomheder og organisationer kan endda pryde sig med en hjemmeside på samtlige EU-sprog, men teksternes kvalitet er tit tvivlsom. Mange oversættelser fra tjekkisk til dansk bliver eksempelvis udført i Tjekkiet af tjekkiske oversættere med mangelfuldt kendskab til dansk fagterminologi, men selv de danske oversættere har tit problemer med tekniske oversættelser fra tjekkisk til dansk, dels på grund af manglende ordbøger (både tekniske og almensproglige) og dels på grund af manglende erfaring inden for de forskellige fagområder.

Den mest pålidelige kilde til tekster og deres oversættelser er skønlitteraturen. Oversættelsernes ophavsmænd og –kvinder er som regel kendt og kan i bedste tilfælde direkte stille deres oversættelser til rådighed i elektronisk form. Desuden er en del vigtige eller opsigtsvækkende værker blevet oversat til de fleste europæiske og nogle ikke-europæiske sprog og kommer således i betragtning som fælles kerne i et potentielt parallelt korpus. Mange tekster forfattet på engelsk, fransk og tysk kan opdrives elektronisk på internettet (The Oxford Text Archive er et uvurderligt hjælpemiddel på dette punkt), men små sprog som dansk er sjældent repræsenteret, især når det gælder oversættelser fra andre sprog. *Intercorp* har derfor henvendt sig til en række forlag med henblik på at få tilsendt tekster i elektronisk form. Samarbejdet foregår som regel på grundlag af en samarbejdsaftale, hvori det fastslås, under hvilke vilkår teksterne frigives til brug i forbindelse med projektet *Intercorp*.

7.2 Indscanning af tekster

Hvis de relevante tekster ikke foreligger i elektronisk form (som det er tilfældet for størstedelen af ældre oversættelser fra tjekkisk til dansk og omvendt), skal teksterne scannes ind og konverteres til en tekstfil. Dette sker normalt ved hjælp af en almindelig scanner og programmet FineReader, der kan udføre OCR[4] og eksportere filen til WORD. FineReader har mange praktiske funktioner, bl.a. fejldetektion, men netop denne funktion kan være meget upålidelig ved tekster af ældre dato trykt på papir af ringe kvalitet med gammeldags skrifttyper eller ved tekster, der viser tydelige tegn på brug som pletter, fingeraftryk osv. Disse defekter bliver typisk genkendt som skrift, og det er derfor vigtigt at rense teksten for alle uønskede elementer og fejl forud for konverteringen.

7.3 Editering forud for alignering[5]

Forud for alignering skal den elektroniske tekst i Word-format konverteres til en tekstfil med passende kodning for at sikre korrekt visning og forsynes med opmærkning af afsnit, sætninger og skrifttyper ved hjælp af HTML-tags for at opnå kompatibilitet med XML-standarden. Ændringerne udføres automatisk gennem to macro-løsninger.

7.4 Alignering i ParaConc

Til sidst foretages alignering af begge parallelle tekster i programmet ParaConc. Dette sker gennem manuel splitning, sammenfletning og sletning af enkelte segmenter eller afsnit og indsættelse af tomme segmenter eller afsnit i den fremmedsprogede udgave af teksten efter den tjekkiske, kanoniske udgave. Den tjekkiske tekst indgår i Det Tjekkiske Nationale Korpus' tekstdatabase og står således til rådighed for alle projektdeltagere. Af denne grund er den tjekkiske tekst låst for editering i ParaConc. Når antallet af afsnit er identisk og de ækvivalente afsnit står side om side, udfører programmet automatisk alignering på segmentniveau. Koordinatoren har herefter mulighed for at tjekke teksten for uoverensstemmelser, inden den færdige tekst eksporteres fra ParaConc og afleveres.

7.5. Tekstdatabase

Eftersom de enkelte aktører i projektet arbejder mere eller mindre selvstændigt og uafhængigt af de andre hold, er det væsentligt at holde hinanden opdateret om anskaffelse af nye tekster,

[4] Optical character recognition

[5] Parallel opstilling af de ækvivalente tekstenheder (afsnit, segmenter) og efterfølgende oprettelse af link mellem disse

der kan have interesse for andre hold og om status for arbejdet. Kommunikation mellem holdene foregår hovedsageligt gennem en database, der indeholder aktuelle oplysninger om alle tekster, der søges anskaffet, der er anskaffet i papirform, der er blevet digitaliseret og der er afleveret i aligneret form til servicecentret ved Instituttet for Det Tjekkiske Nationale Korpus. Derudover udskifter projektdeltagere deres erfaringer på et forum, der er tilgængeligt fra *Intercorp*s hjemmeside https://trnka.ff.cuni.cz/ucnk/intercorp.

8. Anvendelse af parallelle tekstkorpora

De parallelle tekstkorpora kan med fordel anvendes af både oversættere, undervisere og sprogforskere og som et vigtigt redskab til systematisk bearbejdning af flersproglig terminologi og ordbøger samt til udvikling af værktøjer til maskinstøttet oversættelse. Selv om den endelige version af søgeværktøjet Bonito 2 ikke foreligger endnu, lader forsøgsundersøgelser i ParaConc ane, at de parallelle korpora potentielt har en stor betydning for den sammenlignende sprogvidenskab.

En hyppighedsundersøgelse i den danske og den tjekkiske version af Jan Otcenaseks efterkrigsroman "Romeo, Julie og Mørket" viser for eksempel, at det hyppigste ord i den tjekkiske tekst er det refleksive personlige pronomens akkusativform *se* efterfulgt af konjunktionen *a* (*og*) og præpositionen *na* (på, i), hvorimod det personlige pronomen *han* er hyppigst i den danske udgave af teksten, efterfulgt af *og* på andenpladsen og præpositionen *i* på tredjepladsen.

En søgning på kopulaverbet *budu* (1. pers. sg. indikativ futurum af verbet *být* - at være), der indgår i det perifrastiske futurum af tjekkiske imperfektive verber, og på ækvivalente danske former i samme tekst afslører, at den tjekkiske imperfektive fremtidskonstruktion ikke har en kanoniseret ækvivalent i dansk, men oversættes med både *skulle, ville, vil, blive* (+ infinitiv) og almindelig præsens. En parallel søgning på *skal* og *budu* derimod giver ingen resultater, dvs. ingen fremtidsform med *budu* bliver oversat med det danske modalverbum *skal* og infinitiv, hvilket er overraskende i betragtning af, at modalverbet *skal* tit beskrives som fremtidsmarkør (fx Mikkelsen 1911).

9. Perspektiver

Projekt *Intercorp* har endnu langt til målet, men korpussets vækst og de foreløbige resultater tyder på, at arbejdet vil bære frugt. Det næste skridt bliver at forsyne korpusset med morfosyntaktisk opmærkning med henblik på bedre udnyttelse af korpussets potentiale, men dette mål ligger indtil videre uden for projektets ramme.

Litteratur

Čermák, František, Schmiedtová, Věra (2004) Český národní korpus – základní charakteristika a širší souvislosti. *Národní knihovna. Knihovnická revue,* årg. 15, nr. 3:152-168.

Černý, Jiří (1996) *Dějiny lingvistiky*. Olomouc: Votobia.

Kirchmeier-Andersen, Sabine (2002) Dansk korpusbaseret forskning. *NyS*, nr. 30:11-26

Mikkelsen, Kristian (1911) *Dansk Ordföjningslære*. København: Hans Reitzels Forlag.

Det Tjekkiske Nationale Korpus'hjemmeside http://ucnk.ff.cuni.cz/