KEEL JA KIRJANDUS 5/1991

EESTI TEADUSTE AKADEEMIA JA EESTI KIRJANIKE LIIDU AJAKIRI
XXXIV AASTAKÄIK

Lingvistilised korpused keeleuurimises

HALDUR ÕIM

Mulle tundub, et eesti keeleteadlastel on praegu ülim aeg — ka tähenduses 'ülimalt sobiv aeg' — tõsiselt mõelda arvuteil realiseeritud tekstide fondi loomisele, kust iga eesti keele uurija võiks hankida oma näitematerjali (selle asemel, et seda lõpmatuseni sedeldada ja sedeldada). Fondis sisalduvaid tekste saaks uurida vormilisest ja sisulisest, stilistilisest ja kultuurilisest aspektist. Selline fond muudaks arvuti iga lingvisti tõeliselt efektiivseks töövahendiks. Sellest võiks saada eesti keeleteaduse väga tugev edasiviedav ja -lükkav jõud.

Maailmas nimetatakse selliseid fonde üldiselt korpusteks. Ja maailmas on juba rohkesti korpusi, suuri ja väikesi. Neid saab osta ja müüa, kuid meil on sellest vähe kasu, sest me vajame eesti keele korpust.

Alljärgnevas tahaksin tutvustada, mida korpus endast kujutab, mida selle tegemine tähendab, mida korpusega teha saab. Ja niipalju kui oskan, arutada ka seda, mis meie peaksime oma korpus(t)e loomiseks ette võtma.

Mis on lingvistiline korpus?

Korpus on üks 1950-ndate aastate strukturaallingvistika põhimõisteid. Siis mõeldi selle all vastava keele (allkeele, dialekti) kasutust esindavat tekstide kogu, millele kindlaid analüüsiprotseduure rakendades võis lingvist tuletada selle keele grammatika. Zellig Harris esines näiteks oma klassikalises raamatus «Methods in structural linguistics» järgmise tuntuks saanud väitega: «Deskriptiivse lingvistika uurimistöö seisneb mingi dialekti lausungite kirjapanemises. Kirjapandud lausungite kogum moodustab andmete korpuse ja teostatav analüüs kujutab endast korpuse elementide distributsiooni kompaktset kirjeldust.»[1] Z. Harrisele ja temaaegsetele strukturalistidele oli korpus ainus arvessevõetav andmete allikas, millega võidi keelt kirjeldades opereerida. Ühelt poolt selline korpus ja

[1] Z. Harris, Methods in structural linguistics. Chicago, 1951, lk. 12.

teiselt poolt sama rangelt määratletud keelelise (põhiliselt distributiivse) analüüsi protseduurid, mille abil korpusest grammatikafaktid tuletati — see oli klassikalise strukturalismi alus.

Klassikaline strukturalism taandus keeleteaduse metodoloogiana generatiivse grammatika võidulepääsuga ja ühes sellega kaotas ka korpuse mõiste oma põhimõttelise rolli. Generatiivse grammatika järgi ei olnud keeleteaduse ülesandeks mitte keelekasutuse, vaid keelepädevuse kirjeldamine, see tähendab, et objektiks on inimese võime moodustada ja mõista oma emakeeles (lõputut) hulka lauseid, kaasa arvatud sellised, mida keegi enne võib-olla kusagil kasutanud pole. Selge, et niisuguse teoreetilise lähenemise korral ei saanud korpus kui toimunud suhtlusaktide lõplik registreerija mingit põhimõttelist rolli mängida. Lause aktsepteeritavuse üle otsustamisel on lõppkriteeriumiks keele valdaja (*native speaker*) intuitsioon, mitte see, kas lause kusagil korpuses on fikseeritud või ei.

Tegelikult on küll selge, et kirjeldatud vastandus puudutas rohkem teoreetilis-metodoloogilist tasandit kui küsimust, kas keeleuurijale on korpust vaja või ei.

Ei usu, et ka kõige järjekindlam generativist tõsisemat keelekirjeldust üritades ei kogunud eelnevalt andmeid vastavate keelefaktide kohta, vaid püüdis neid meeleheitlikult peast välja mõelda.

Kindel võib siiski olla selles, et suhteline ebasoosing teoreetilisel, üldkeeleteaduslikul tasandil aeglustas märgatavalt 1960.—1970-ndatel aastatel kogu vastavat keelt autentselt katvate tekstikorpuste loomist ja töid nendega.

Kuid algust tehti tõeliselt suurte ja süstemaatiliste tekstikorpuste loomisega siiski just 1960-ndatel aastatel. Need — nagu ka hiljem loodud korpused — aga ei olnud ega ole mingil otsesel viisil seotud strukturalismi ja generatiivse grammatika vastandusega. Need ei olnud mõeldud realiseerima Harrise arusaama korpusest ega ka ümber lükkama generatiivse grammatika ideid. Need korpused on loodud keeleteadlastele abivahendiks ja koostatud kindla süsteemi järgi, nii et nad katavad võimalikult adekvaatselt kogu keelekasutuse vastavas sotsiolektis. Tüüpiline korpus kujutab endast seetõttu seotud teksti väljavõtete kogumit, mis esindab vastavas kultuuris levinud põhilisi žanre. Enamik neist sisaldab kirjutatud tekste, kuid on ka kõnekeelekorpusi, samuti segakorpusi. Ja veel üks oluline moment: kõik tänapäeva korpused on realiseeritud arvutil, sest ainult see teeb võimalikuks efektiivse töötamise. Seetõttu kuulub kogu korpuste loomine ja nende töötlemine — kuid muidugi mitte tulemuste kasutamine — arvutuslingvistikasse. Iseenesest on jõudnud kasutusele tulla termin korpuslingvistika (*corpus linguistics*), mille all mõeldakse arvutuslingvistika seda osa, mis tegeleb korpustega, kusjuures korpuste all ei mõelda enam ainult klassikalisi tekstikorpusi, vaid ka mitmesuguseid eriotstarbelisi suuremaid tekstikogumeid, nagu näiteks seletavad sõnaraamatud vms., mida (mille osi) saab analüüsida kui sidusaid tekste. Korpuslingvistika alal korraldatakse enamasti igal aastal eraldi konverents, korraldajaks ICAME (*International Computer Archive of Modern English*). Tõsi, selle raames tegeldakse vahetult ainult inglise keelega. Seesama ICAME annab välja ka ajakirja «ICAME Journal» (enne 1987. aastat «ICAME News», väljaandmise kohaks on *Norwegian Computing Centre for the Humanities*).

Kuid tutvugem lähemalt mõne korpuse endaga.

Lingvistilisi korpusi

Seni tuntumaid on kolm inglise keele korpust: nn. Browni korpus, Lancasteri-Oslo/Bergeni korpus ja Londoni-Lundi korpus.

Browni korpus (USA Browni ülikooli järgi, kus see välja töötati; algne ametlik nimetus: *A standard corpus of present-day edited American Eng-*

lish) valmis 1964. a.[2] See sisaldab väljavõtteid Ameerika Ühendriikides väljaantud trükitud tekstidest. Kõik väljavõtted on ühesuurused, kujutades endast 2000 sõne pikkusi seotud tekste, ja neid on 15 tekstiliigist (žanrist): ajakirjandusest, ilukirjandusest, teaduslikust ja populaarteaduslikust kirjandusest jne. (vt. allpool), nii et nad eeldust mööda katavad proportsionaalselt kogu «kirjutatud kultuuri». Homogeensuse saavutamiseks on tekstid valitud selliselt, et kõigi väljaandmisaasta on 1961. Väljavõtete arv on erinevates tekstiliikides erinev, kuid kokku on neid 500. Niisiis on korpuse kogumaht 1 miljon sõnet.

Nii tekstiliikide määratlemine kui ka selle kindlaksmääramine, mitu väljavõtet ühest või teisest liigist teha, oli omaette probleem. Seda otsustas eriline ekspertide konsiilium. See, kust konkreetselt üks või teine väljavõte teha, püüti määrata mitmesuguste valiku juhuslikkust tagavate protseduuridega, et välistada subjektiivsust.

1970. a. alustati Inglismaal Lancasteri ülikoolis tekstikorpuse loomist, mis igati järgis Browni korpuse koostamispõhimõtteid, selle vahega, et kõik tekstid esindasid briti inglise keelt. Korpuse loomisel osalesid ka Oslo ülikool ja Bergenis asuv *Norwegian Computing Centre for Humanities*. Siit ka korpuse nimetus: Lancasteri-Oslo/Bergeni korpus, lühendatult LOB.[3] See valmis 1978. a. Nii nagu Browni korpus, sisaldab ka LOB 500 tekstiväljavõtet, igas 2000 sõnet, s. o. kokku 1 miljon. Tekstide ilmumisaasta on samuti 1961. Ka tekstide liigitus on sama, kuid väljavõtete arv liigiti erineb mõningal määral Browni korpuse omast, kajastades LOB-i loojate arusaama vastava tekstiliigi erinevast kaalust inglise kultuuris ameerika omaga võrreldes. Ülevaatlikkuse huvides esitame kummagi korpuse tekstide liigituse ja iga liigi väljavõtete arvu:

	Browni korpus	LOB
1. Ajakirjandustekstid: reportaažid	44	44
2. Ajakirjandustekstid: juhtkirjad	27	27
3. Ajakirjandustekstid: ülevaate- ja probleemartiklid	17	17
4. Religioosne kirjandus	17	17
5. Harrastused, oskused	36	38
6. Populaarkirjandus (*popular lore*)	48	44
7. Biograafiad, esseed	75	77
8. Ametlikud dokumendid	30	30
9. Teaduslik kirjandus	80	80
10. «Üldine» ilukirjandus (*general fiction*)	29	29
11. Detektiiv- ja põnevuskirjandus	24	24
12. Ulmekirjandus	6	6
13. Seikluskirjandus, vesternid	29	29
14. Poeesia, armastuslood (*love stories*)	29	29
15. Huumor	9	9
Kokku:	500	500

Mõlemad korpused sisaldavad, nagu öeldud, ainult kirjalikke (trükis avaldatud) tekste. Kui korpuse üldine eesmärk on olla keelekasutuse uurimise aluseks, siis on see muidugi oluline kitsendus. Kõnekeele korpuse loomine on aga märksa raskem ülesanne. Keerulisem on keelematerjali hankimine, eriti keeruline aga transkribeeritud materjali arvutisse viimine. Siiski on seni loodud ka mitmeid kõnekorpusi. Neist esimene ja kahtlemata tuntuim on nn. Londoni-Lundi korpus.

Selle aluseks on tekstikorpus *Survey of English usage*, mis loodi Lon-

[2] N. W. Francis, H. Kučera, Manual of information to accompany a standard corpus of present-day edited American English, for use with digital computers. Providence, R. I., 1964; vt. ka: H. Kučera, N. W. Francis, Computational analysis of present-day American English. Providence, R. I., 1967.

[3] S. Johansson, G. Leech, H. Goodluck, Manual of information to accompany the Lancaster-Oslo/Bergen corpus of British English, for use with digital computers. Oslo, 1978.

donis *University College*'is 1960-ndail aastail. See oli algselt tavaline, mitte arvutikorpus. Korpus sisaldas 5000-sõnelisi valimikke, mida oli 200. Neist pool kujutas endast originaalis kirjalikke tekste, alustades tavalise ilukirjandusproosaga ja lõpetades raadiouudistega (kokku 100 valimikku), ja teine pool algupärast kõnematerjali: 24 monoloogi (näit. spordikommentaarid) ja 76 dialoogi-vestlust.

1975. a. tehti Lundis algust projektiga *Survey of spoken English*, mille põhieesmärgiks oli viia eelkirjeldatud *Survey of English usage* arvutisse. Nii kujuneski Londoni-Lundi korpus. Osa sellest koos korpuse kirjeldusega on ilmunud ka raamatuna [4].

Arvutikorpuses on 87 valimikku, kokku 435 000 sõnet, valimikud jagunevad järgmiselt:

1. Vahetu vestlus (*face-to-face conversation*)	46
2. Telefonivestlused	10
3. Diskussioonid, debatid, intervjuud	12
4. Avalikud ettevalmistamata kõned, kommentaarid jm.	12
5. Avalikud ettevalmistatud kõned	7
Kokku:	87

Transkriptsiooni on arvutikorpuses originaaliga võrreldes lihtsustatud, eriti prosoodia osas, kuid põhilised näitajad, mis on olulised lause grammatilise ja semantilise organisatsiooni seisukohalt (rõhud, intonatsioon, pausid jms.), on kodeeritud.

Lisaks neile kolmele tuntuimale on maailmas loodud veel mitmeid korpusi, mida tasub nimetada. Piisavalt andmeid on kahjuks ainult inglise keele korpuste kohta.

Indias Kolhapuris Shivaji ülikoolis on loodud India inglise keelt esindav nn. Kolhapuri korpus.[5] See järgib Browni korpuse ja LOB-i põhimõtteid: tekstide liigitus, korpuse kogumaht (1 miljon sõnet) ja väljavõtete arv ning maht on samad. Kasutatakse samuti ainult trükis ilmunud tekste, kuid tekstide ilmumisaasta on 1978. Olulisim erinevus puudutab üksikutesse tekstiliikidesse kuuluvate väljavõtete arvu, põhjuseks taas vastavate tekstiliikide erinev kaal India kultuuris. Näiteks on Kolhapuri korpuses seikluskirjandusest ja vesternitest 18 väljavõtet Browni ja LOB-i korpuse 29 asemel; ilukirjanduse üldliigist (*general fiction*) aga 58 väljavõtet Browni ja LOB-i korpuse 29 asemel.

1985. a. alustati Austraalia inglise keele korpuse loomist Austraalias Macquarie ülikoolis [6]. Selle esimene järk kujutab endast põhimõtteliselt Browni korpuse ja LOB-i analoogi. Tekstid on valitud 1986. a. ilmunud trükistest. Erinevus on taas üksikutele tekstiliikidele omistatavas kaalus.

Teise järguna on aga kavas luua nn. monitorkorpus: korpus, mis ei esindaks ainult üht sotsiolekti («educated English»), vaid ka keelekasutuse muid variante, kaasa arvatud kõnekeele variandid, ja mida saaks vajadust mööda pidevalt täiendada uue materjaliga.

Seega järgib kirjeldatud projekt juba varem Birminghami ülikooli tekstikorpuse loomisel rakendatud põhimõtteid [7]. Ka selles eristatakse nn. «läbilõikekorpust» (*sample corpus*), mis on suhteliselt väike, kuid annab läbilõike kogu keelest (s. t. ühest sotsiolektist; just sellised on LOB ja

---

[4] J. Svartvik, R. Quirk, A corpus of English conversation. Lund, 1980.

[5] S. V. Shastri, C. T. Patikulkarni, G. Shastri, Manual to accompany the Kolhapur corpus of Indian English for use on digital computers. Kolhapur, 1986. Vt. ka: S. V. Shastri, The Kolhapur corpus of Indian English and work done on its basis so far. ICAME Journal 12. Bergen, 1988, lk. 15—26.

[6] P. Peters, Towards a corpus of Australian English. ICAME Journal 11. Bergen, 1987, lk. 27—38.

[7] J. M. Sinclair, Reflections on computer corpora in English language research. Rmt.: S. Johansson (toim.), Computer corpora in English language research. Bergen, 1982, lk. 1—6.

Browni korpus), ja teiselt poolt monitorkorpust, mis on märgatavalt suurem (Birminghami korpuses oli 1985. a. 7,5 miljonit sõnet), on pidevalt täiendatav uue materjaliga ja kajastab keele erinevaid kasutusvariante. Samal ajal võib selline monitorkorpus olla orienteeritud teatavat kindlat tüüpi lingvistilisele uurimistööle. Birminghami korpus on mõeldud eelkõige leksikoloogilisteks ja leksikograafilisteks töödeks (1987. a. ilmus põhiliselt selle korpuse najal tehtud inglise keele seletav sõnaraamat [8]). Hollandis Nijmegeni ülikoolis loodud/loodav sama tüüpi inglise keele korpus aga on orienteeritud keele süntaktiliste struktuuride uurimisele.

Sellega olemegi jõudnud lingvistiliste korpuste teise tüübi juurde. Seni vaadeldud korpused taotlevad anda läbilõiget kogu (tänapäeva) keelest, seejuures nii, et ühe või teise žanri tekstide maht kajastaks vastava tekstiliigi kaalu kultuuris. Teine võimalus aga on teha spetsialiseeritud korpusi, kus korpusesse lülitatavad tekstid valitakse mingite täiendavate kriteeriumide alusel.

Esimesena nimetatagu diakroonilisi korpusi, mis on mõeldud keeleajaloolisteks uuringuteks. Helsingi ülikoolis loodud inglise keele diakrooniline korpus sisaldab tekste tuhandeaastasest ajavahemikust, VIII—XVIII sajandini.[9] Kogumaht on umbes 1,5 miljonit sõnet. Väljavõtete pikkus varieerub, kuid enamasti on see 5000—10 000 sõne vahel. Nagu autorid kirjutavad, on nad püüdnud valida tekstid nii, et korpus oleks representatiivne järgmise nelja faktori osas: 1. teksti ilmumisaeg (kirjutamisaeg); 2. geograafiline dialekt; 3. teksti tüüp; 4. stiilitüüp.

Tekstid on periodiseeritud esmalt inglise keele ajaloo traditsioonilise jaotuse järgi (vaadeldavasse ajavahemikku jäävad *Old, Middle* ja *Early Modern English*), nende perioodide sees aga sajandite järgi. Teksti tüüp on määratletud kolmekümne viie parameetriga, mis varieeruvad erakirjavahetusest Shakespeare'i teosteni ja piibli erinevate väljaanneteni. Kategooriaga «stiilitüüp» on silmas peetud eelkõige kaht stiilieristust: esiteks teksti formaalsuse—vahetuse telge ja teiseks populaarsuse—neutraalsuse—retoorilisuse—ametlikkuse telge.

Diakroonilisi korpusi on ka väiksemate ja kitsamalt piiritletud tekstitüüpide kohta. Clevelandi ülikoolis (USA) on loodud nn. ühe sajandi proosa korpus (*The century of prose corpus*), mis haarab ajavahemikku 1680—1780 jäävat inglise proosakirjandust.[10]

Nijmegeni ülikoolis loodud korpuse eesmärgiks on uurida keelelist variatiivsust, seda eriti süntaktiliste struktuuride osas.[11] Selleks et ühte väljavõttesse saada võimalikult suurt arvu varieeruvaid struktuure, valiti väljavõtte suuruseks vähemalt 20 000 sõnet. Korpus tervikuna sisaldab 1,5 miljonit sõnet.

Võrdlemisi omapärane on Carnegie Melloni ülikooli lastekeele korpus, mis rahvusvahelise projektina üritab olla lastekeelest transkribeeritud tekstide vahendamise alus.[12] Süsteem hõlmab andmestikku ennast ja andmete transkribeerimise programmi. Materjali on ka teistest keeltest peale inglise keele. Andmebaasi kogumahuks on antud 140 miljonit täheüksust.

Võiks mainida mitmesuguseid muid tekstikogumeid, mida kasutatakse kui korpusi eriotstarbeliste uuringute puhul. Näiteks saab seletava sõnaraamatu sõnaseletuste kogumit käsitada korpusena, töödelda seda selli-

[8] J. Sinclair, P. Hanks, G. Fox, R. Muon, D. Stock (toim.), Collins COBUILD English language dictionary. London—Glasgow, 1987.

[9] O. Ihalainen, M. Kytö, M. Rissanen, The Helsinki corpus of English texts: Diachronic and dialectal. Rmt.: W. Mejs, Corpus linguistics and beyond. Amsterdam, 1987, lk. 21—32; M. Kytö, Progress report on the diachronic part of the Helsinki corpus. ICAME Journal 13. Bergen, 1989, lk. 12—15.

[10] T. Milič, A new historical corpus. ICAME Journal 14. Bergen, 1990, lk. 26—39.

[11] N. Oostdijk, A corpus for studying linguistic variation. ICAME Journal 12. Bergen, 1988, lk. 3—14.

[12] B. MacWhinney, C. Snow, The child language data exchange system. ICAME Journal 14. Bergen, 1990, lk. 3—25.

sena ja teha vastavaid uurimusi — põhiliselt semantilisi. Lähemalt käsitleme seda tüüpi töid järgmises alajaotuses.

Töö korpustega

Korpustega tehtavad tööd võib jagada kaheks: esiteks korpuste töötlemine eesmärgiga varustada tekstid (neis sisalduvad sõnavormid, laused, ka pikemad tekstilõigud) relevantse grammatilise ja semantilise informatsiooniga; teiseks suvaline lingvistiline uurimistöö ise, mille puhul korpus esineb andmete allikana (see on töö, mille jaoks korpused tehaksegi).

Korpus, kus sinna viidud tekstid säilitatakse muutumatul kujul, ilma mingi täiendava infota, oleks võrdlemisi ebaefektiivne: vajaliku keelematerjali kättesaamine sellest võib osutuda keeruliseks. Keeleteadlast, kes tahab korpust kasutada mingi probleemi lahendamisel, huvitab enamasti mingi sõna (või sõnade), grammatiliste kategooriate või nende kombinatsioonide ja süntaktiliste konstruktsioonide esinemine mitmesugustes tekstides ja kontekstides. Selleks, et arvuti abil võiks niisuguste tunnuste järgi korpusest tekste otsida ja vastavat analüüsi teha, ongi vaja, et tekstid korpuses oleksid eelnevalt varustatud vajaliku grammatilise ja semantilise infoga. See aga tähendab, et enne kasutamiskõlblikuks saamist tuleb korpuse tekstid allutada grammatilisele (morfoloogilisele, süntaktilisele), semantilisele jne. töötlusele. Inglise keeles tähistatakse vastavaid protseduure sõnaga *tagging*, eesti keeles võiks see siis ehk olla *märgendamine*. Need on protseduurid, mille sisendiks on korpuse «toored» tekstid ja väljundiks tekstid, mis on varustatud vastavate grammatiliste, semantiliste jne. märgenditega.

Praktiliselt moodustab praegustes korpustes märgendamise põhisisu sõnavormide varustamine morfoloogiliste kategooriatega: sõnaliik, arv, kääne, finiitsed ja infiniitsed verbivormid oma kategooriatega. See analüüs on aluseks ka nn. lemmatiseerimisele, protseduurile, kus mingi sõna erinevad grammatilised vormid tekstis identifitseeritakse kui nimelt selle sõna vormid (alles selle järel on võimalik automaatselt jälgida sõnade ja mitte ainult individuaalsete sõnavormide esinemisi tekstides). Sõnu saab tänapäeval morfoloogiliselt märgendada peaaegu täielikult arvutite abil, kasutades automaatse morfoloogilise analüüsi meetodeid. Erinevates keeltes tekivad siin muidugi erinevad probleemid. Inglise keele puhul on üks keerukamaid probleeme sõnaliikide eristamine, iseäranis nimisõnade ja tegusõnade mitmete vormide korral (*turn* või *wish* võivad olla nii nimisõna kui tegusõna vormid). Appi võetakse kontekstina esinevad teised sõnad ja mitmesugused tõenäosuslikud kriteeriumid.

Rikkaliku infleksiooniga eesti keeles ei ole sõnaliigi määramine nii raske, ehkki sõnavormide homonüümiat on rohkesti meilgi.[13] Kuid meil on ilmselt põhiline probleem infleksiooniafiksite tuvastamine ja selle kindlakstegemine, missuguse sõnatüvega tegemist on.

Niisamuti saab teha automaatselt töid, mis tuginevad sõnade morfoloogilistele märgenditele: tekstide indekseerimist sõnade või grammatiliste kategooriate järgi ja vastavate konkordantside koostamist, sõnade, sõnavormide või grammatiliste kategooriate esinemissageduste arvutamist tekstides jne. Tänu lihtsamate märgendamisoperatsioonide, aga ka tekstidega opereerimise üldiste meetodite ulatuslikule automatiseerimisele (tekstide redigeerimine, tekstimassiivide organiseerimine arvuti mälus) on viimastel aastatel loodud süsteeme, mis võimaldavad korpusi ehitada automatiseeritult. Inimestel on vaja valida tekstid ja viia need arvutisse. Kõik ülejäänu teeb arvuti ise, kuni kasutamiskõlblik korpus on valmis. Ka tekstiliikide ja üksikute tekstide identifitseerimise koodid pakub süsteem, kuid korpuse tegija peab muidugi otsustama, missuguse sisu ta neile koodidele

[13] Vt. Ü. Viks, Sõnavormide homonüümia eesti keeles. «Keel ja Kirjandus» 1984, nr. 2, lk. 97—105.

annab. Geoffry Kaye kirjeldab üht sellist süsteemi, mis on mõeldud korpuste tegemiseks personaalarvuteil, kusjuures tekstikorpuse mahuks on arvestatud kuni miljon sõnet [14]. Tõsi, ka siin on konkreetselt silmas peetud inglise keelt ja teistele keeltele neid vahendeid niisama lihtsalt üle kanda ei saa.

Tekstide süntaktiline märgendamine — lausete osaline või täielik varustamine süntaktiliste kategooriatega (näit. nimisõnafraas, verbifraas, infiniitkonstruktsioon, kõrvallause jne.) ja funktsioonidega (näit. subjekt, predikaat, objekt; agent, patsient, kogeja, teema, reema jne.) — on juba märksa keerulisem ülesanne. Seni on seda tööd suurel määral käsitsi tehtud ja enamasti ei ole see täielik: märgendamiseks valitakse ainult teatavad süntaktilised kategooriad ja/või suuremaid korpusi märgendatakse järk-järgult osade kaupa, sõltuvalt sellest, missugused osad süntaktiliste uuringute aspektist huvi pakuvad. Selles suunas töötatakse siiski aktiivselt ja võib kindel olla, et paari-kolme aasta pärast võib rääkida olulistest edusammudest nendegi tööde automatiseerimisel.

Juba praegu pakuvad personaalarvutid häid võimalusi teha uurimusi interaktiivselt, nii et lingvist töötab materjaliga, mis on näha arvuti ekraanil, ja arvuti ise osaleb aktiivselt.

Neist ettevõtmistest, kus süstemaatiliselt on üritatud automatiseerida korpuse süntaktilist märgendamist, on kahtlemata tuntuimad Nijmegeni ülikooli juures tegutseva korpuslingvistika uurimisrühma tööd.

Juba 1970-ndail aastail alustati seal esimesi katseid ulatuslikumate tekstikorpuste automaatseks süntaktiliseks märgendamiseks.[15] Siis töötati välja ka esimesed põhimõtted, mida korpuste süntaktiline märgendamine peaks järgima. Muu hulgas osutati, et kuivõrd korpused on mõeldud kasutamiseks laiale keeleteadlaste ringile, kelle teoreetiline taust on vägagi erinev, siis peab märgendamisel kasutatav süntaktiliste kategooriate süsteem olema võimalikult neutraalne ja laialdaselt aktsepteeritav ning mõistetav, mitte aga järgima mingit kitsast teoreetilist süntaksikontseptsiooni.

Tõsisemalt asuti nende küsimuste kallale 1980-ndate aastate esimesel poolel, kui alustati töid projekti TOSCA (= *Tools for syntactic corpus analysis*) kallal, mis töötas välja vahendite süsteemi korpuse täielikuks süntaktiliseks märgendamiseks (s. o. igale lausele korpuses antakse täielik süntaktiline kirjeldus vastava grammatika termineis) [16]. Sellele järgnes projekt TOSCA II, mille eesmärgiks oli väljatöötatud süsteemi kasutades märgendada ühe miljoni sõneline briti inglise keele korpus (TOSCA I raames märgendati 130 000-sõneline korpusefragment) [17].

Süsteemi töö aluseks on analüüsigrammatika ja leksikon. Leksikoni, millesse kuuluvad analüüsiks vajalikku infot sisaldavad sõnaartiklid, võib pidevalt «kasvatada» korpuse märgendamise käigus. Eeldatakse, et korpus on enne morfoloogiliselt märgendatud. Süntaktiline analüüs toimub interaktiivselt, s. o. lingvistil on võimalus pidevalt sekkuda, valida mitme analüüsivariandi hulgast õige, kõrvaldada valesid analüüsitulemusi jne. Ja teiselt poolt on süsteemil võimalus saada lingvistilt puuduolevat infot (näit. tundmatute sõnade kirjeldusi).

Kasutatav formaalne grammatika, nn. *Extended Affix Grammar* on üks fraasistruktuurigrammatika variante. See tähendab, et lausete komponendid kategoriseeritakse tüüpiliste fraasistruktuurigrammatika kategooriate

---

[14] G. Kaye, A corpus builder and real-time concordance browser for an IBM PC. Rmt.: J. Aarts, W. Mejs (toim.), Theory and practice in corpus linguistics. Amsterdam, 1989, lk. 137—161.

[15] F. Keulen, The dutch computer corpus pilot project. Rmt.: J. Aarts, W. Mejs (toim.), Corpus linguistics II: Recent advances in the use of computer corpora in English language research. Amsterdam, 1986, lk. 127—162.

[16] The Nijmegen research group for corpus linguistics, TOSCA. Nijmegen, 1987.

[17] The Nijmegen research group for corpus linguistics, TOSCA, lk. 44.

järgi ja varustatakse süntaktilist funktsiooni kajastava märgendiga: *subject, direct object, verb, modifier* jne.

Märkimist väärib, et TOSCA I raames välja töötatud süsteemi on katsetatud ka teiste keelte tekstide märgendamiseks: projektis ASCAMSA araabiakeelsete tekstide (ASCAMSA = *Automatic syntactic corpus analysis of modern standard Arabic*) ja projektis ASATE hispaaniakeelsete tekstide puhul (ASATE = *Análisis sintáctico automatizado des textos Españoles*). 1988. a. töötati araabia keele jaoks välja sõnastik, milles oli 10 000 sõnaartiklit ja suurem osa analüüsigrammatikast. Hispaania keele jaoks oli olemas 5000-tüveline sõnastik, analüüsigrammatika kirjutamine oli algstaadiumis. Araabia keele korpus oli 300 000-sõneline, hispaania keele oma poole miljoni sõneline.[18] Tänaseks päevaks on töö mõlema projekti kallal ilmselt hoopis kaugemal.

Nagu siit nähtub, tuleb iga keele jaoks siiski välja töötada oma analüüsigrammatika ja oma sõnastik, mis on ka igati arusaadav. Nende aluseks võib küll olla ühtne ideestik, mis oluliselt hõlbustab konkreetsete probleemide lahendamist.

Süntaksiga seoses on kohane mainida ka seni vist ainust eesti keele alal tehtud tööd, mis mahub tekstikorpuse raamesse. Kaja Tael, olles KKI juures aspirantuuris, kirjutas väitekirja eesti keele sõnajärge mõjutavatest teguritest. Et objektiivsemat pilti saada, koostas ta 3000 tekstilausest koosneva pisikese tekstikorpuse, mis viidi arvutisse ja märgendati.[19] Eeskujuks oli analoogiline uurimus soome keele kohta.[20] Et korpus oli mõeldud väga spetsiifilise probleemi uurimiseks, siis oli ka kategooriate süsteem võimalikult detailne, haarates kõiki morfoloogilisi, süntaktilisi, semantilisi ja pragmaatilisi kategooriaid, millest võis oletada, et nad mõjutavad sõnade järjestamist eestikeelses lauses. Kokku hinnati iga lauset 74 kodeeritava kategooria seisukohalt, alustades sõnaliikidest ja traditsioonilistest süntaktilistest kategooriatest, nagu subjekt ja objekt, ning lõpetades näit. semantiliste rollidega, nagu agent, kogeja, patsient, benefitsient, ja lause üldise semantilise ja pragmaatilise tüübiga. Niisugust kodeerimist sai muidugi teha ainult käsitsi. Kuigi korpuse loomisel peeti silmas üht kindlat sihti, on materjal kasutatav ka teistsugusteks uuringuteks, eelkõige lausesemantika ja -pragmaatika uurimiseks.

Märgendatud korpused võivad olla väga spetsiifilise eesmärgiga. Märgendid ei pruugi sugugi kajastada ainult keeleüksuste morfoloogilisi ja/või süntaktilisi tunnuseid. Põhimõtteliselt saab ju märgendada igasugust informatsiooni, mis lingvistidele huvi võib pakkuda.

On olemas korpusi, kus märgendatavateks on hoopis kõrgema tasandi kategooriad. Sellise näitena võib tuua süsteemi, kus korpuse moodustavad dialoogid, milles märgendatakse erilised dialoogiüksused.[21] Viidatud süsteemis eristatakse hierarhiliselt dialoogiüksuste viit tasandit: 1) dialoogi kui terviku tüüp (infohankimine, konsultatsioon), 2) vooru tüüp (suhtluse avamine/sulgemine, metakommentaar jms.), 3) lausungi tüüp (initsiatsioon, vastus jt.), 4) lausetüüp (näit. küsimus, direktiiv, väide); 5) moodustaja tüüp (osalause, verbifraas, nimisõnafraas jt.). Dialoogide märgendamine toimub interaktiivselt: pärast seda, kui lingvist on valinud kõrgema tasandi üksuse mingi dialoogilõigu tähistamiseks (näit. «küsimus»), pakub süsteem välja selle üksusetüübi võimalikud madalama

---

[18] The Nijmegen research group for corpus linguistics, TOSCA, lk. 43 jj.

[19] K. Tael, Sõnajärjemallid eesti keeles (võrreldurna soome keelega). Preprint KKI-56. Tallinn, 1988.

[20] A. Hakulinen, F. Karlsson, M. Vilkuna, Suomen tekstilauseiden piirteitä: kvantitatiivinen tutkimus. Helsingin yliopiston yleisen kielitieteen laitoksen julkaisuja, nr. 6, 1980.

[21] L. Ahrenberg, A. Jönsson, An interactive system for tagging dialogues. «Literary and Linguistic Computing» 1988, kd. 3, nr. 2, lk. 66—70.

tasandi liigid («üldküsimus», «alternatiivküsimus» jne.), mille hulgast märgendaja peab valima sobiva.

Arusaadav, et sellise süsteemi rakendamine eeldab detailse dialoogimudeli olemasolu. Niisamuti on see muidugi keele kõrgemate tasandite (semantika, pragmaatika) märgendamisega; erinevalt morfoloogiast ja (osalt) süntaksist on nende puhul esimeseks probleemiks märgendatavate kategooriate süsteemi valik.

Korpused lingvistilises uurimistöös

Korpuste puhul on lingvistide jaoks suuremaks ja töömahukamaks osaks korpuse tegemine. Kui korpus on valmis, siis on selle kasutamine keeleteadlase igapäevatöö osa. Vahetult aga lihtsustab korpus vaid töö üht etappi — materjali hankimist — ning tagab süstemaatilisuse ja autentsuse keele (sotsiolekti vm.) selle osa ulatuses, mida korpus haarab. Tuleks aga rõhutada kaht momenti. Esiteks, paljud tööd saavad praktiliselt võimalikuks alles korpuse olemasolu korral. Seetõttu ei maksa lasta end eksitada formuleeringu «lihtsustab materjali hankimist» tähenduse näivast lihtsusest. Iga natukegi mastaapsem keeleuurimistöö on ju selline, et materjali hankimine on selles üks vaevarikkamaid ja ka aeganõudvamaid etappe, millele võib kuluda aastaid. Sobivalt märgendatud korpuse olemasolu korral kulub samaks tööks mõni päev või nädal.

Teiseks, korpust ei looda mitte üheks-kaheks uurimistööks, vaid selle mõte on olla kättesaadav igale keeleuurijale temale tarviliku materjali hankimisel, puudutagu see siis sõnavara, morfoloogiat või süntaksit, tekstiuuringuid.

Toome siin illustratsiooniks mõne näite erinevatelt keeleuurimisaladelt.

Lundi ülikoolis on käivitatud programm uurimaks mitmesuguseid inglise keele nn. suhtlusväljendeid (*conversational phrases*). Allikaks on Londoni-Lundi kõnekeelekorpus.[22]

Mõeldud on põhiliselt modaalse sisuga väljendeid, mille põhifunktsioon on diskursuse organiseerimine, suunamine, näiteks dialoogis järgneva vooru sissejuhatamine. Neid võib olla väga erinevaid tüüpe. Kui otsida eesti keele vasteid, siis näiteks väljendid *kui päris aus olla, kui (sulle) tõtt öelda* jms. juhatavad tüüpiliselt sisse pöördumise, mille sisu kõneleja arvates on kuulajale teataval määral ebameeldiv. Teine tüüp väljendeid võib signaliseerida (vahetus vestluses) kuulaja valmisolekut edasi kuulata, seda, et tal pole arusaamisega probleeme ja et ta ei soovi ise rääkima hakata (nn. *carry-on signals*): *jaa, muidugi, täiesti, tõesti?*.[23]

Korpuse olemasolu võimaldab süstemaatiliselt uurida selliste organiseerivate väljendite esinemist erinevates diskursusetüüpides, erinevates situatsioonides, nende sõltuvust muudest faktoritest, nende omavahelisi korrelatsioone. Korpuse kasutamise peamine efekt seisneb siin materjali operatiivses kättesaamises.

Nijmegeni ülikoolis on sealset süntaktiliselt märgendatud korpust kasutades tehtud algust projektiga, mille eesmärgiks on uurida enam levinud konstruktsioonide distributsiooni tänapäeva inglise keeles.[24]

Uuritakse näit. seda, missuguste struktuuridena realiseeruvad subjekt, objekt või muud lause funktsionaalsed komponendid, mis tingimustel esineb inversioon, missugused on nimisõnafraaside tüüpilised struktuurid, millest sõltub nende erinevate variantide esinemine jne.; seda kõike kõrvu-

[22] K. Aijmer, Work in progress within the project Conversational phrases in English. ICAME Journal 14. Bergen, 1990, lk. 44—48.

[23] A.-B. Stenström, Carry-on signals in English conversations. Rmt.: Corpus linguistics and beyond. Amsterdam, 1986, lk. 87—119.

[24] P. de Haan, Structure frequency counts of modern English: Progress report. ICAME Journal 13. Bergen, 1989, lk. 53—55.

tavalt neis erinevates tekstitüüpides, mida korpus sisaldab. Seni on ilmunud nimisõnafraasi väga põhjalik käsitlus.[25]

See on juba täiesti ilmselt seda tüüpi töö, mida võib küll teoreetiliselt kujutleda käsitsi tehtavana, aga mille praktiline teostamine täies süstemaatilisuses ja põhjalikkuses on mõeldamatu ilma korpuseta, mida arvuti abil analüüsitakse. (Põhimõtteliselt samasse tüüpi kuulub ka eespool mainitud eesti keele sõnajärje uurimine.)

Teist tüüpi näide on Göteborgi ülikoolis käsil olev sõnade semantilise analüüsi alale kuuluv ulatuslik projekt.[26] Töö vahetu eesmärk on suure, tähenduskirjetega varustatud sõnadest koosneva sõnamassi teisendamine nn. formaalseks leksikoniks, kus sõnad oma tähenduskirjelduste kaudu oleksid kindlates seostes. Sõnade tähenduste kirjeldused moodustavad siin korpuse (selles on kirjeldatud üle 150 tuhande sõnavormi). Projekti esimene ülesanne on esitada tähenduste kirjeldused, mis ju tüüpiliselt on loomuliku keele fraasid, formaalsete süntaktiliste struktuuridena (konkreetselt sõltuvuspuudena). See vastab enam-vähem korpuse süntaktilisele märgendamisele. Kuid et tegu on sõnade tähenduste kirjeldustega, on tulemuseks semantika seisukohalt vägagi suurt huvi pakkuv semantiline võrk. (Näit. ühe sõna kirjelduses esinev teine sõna on ka ise kusagil kirjeldatavaks; erinevate sõnade kirjeldustes võib sama sõna esineda erinevates rollides jne.) Selle võrgu töötlemine ja uurimine, semantiliste klasside, hierarhiate või ahelate vms. väljaselgitamine on projekti kaugem ja põhiline eesmärk.

Lõpuks üks näide tekstiuuringute alalt.[27] Douglas Biber Lõuna-California ülikoolist on tegelnud tekstide tüpoloogiaga, lähtudes mitte nende sisulisest või funktsionaalsest tüübist (žanrid vms.), nagu seda seni on enamasti tehtud, vaid teatavate iseloomulike keeleliste (leksikaalsete, süntaktiliste) tunnuste koosesinemisest tekstides, s. o. puhtal kujul objektiivsel keelelisel alusel. Analüüsitud on 481 kirjalikku ja kõnekeeleteksti (valitud LOB-ist ja Londoni-Lundi korpusest) 67 erineva keelelise tunnuse koosesinemise seisukohalt: teatavad leksikaalsed sõnaklassid, pro-vormid, prepositsioonifraasid, passiivi erinevad tüübid, kõrvallausete tüübid, eituse väljendusvormid jne. Faktoraalanalüüsi abil on nende vormide koosesinemise analüüsi põhjal identifitseeritud viis tekstuaalset dimensiooni, mida on võimalik ka sisuliselt interpreteerida. Dimensioonid on skaalad, kuhu erinevad tekstid kindlal viisil paigutuvad ja mis neid tekste suuremal või väiksemal määral iseloomustavad. Biber kirjeldab neid dimensioone järgnevalt: 1) isiklik/informatiivne huvitatus; 2) narratiivne/mittenarratiivne väljendus; 3) eksplitsiitne/situatiivne viitamine; 4) veenmise väline väljendus; 5) abstraktne/mitteabstraktne stiil. Edasi analüüsitakse klasteranalüüsi abil, kuidas korpustes esindatud tekstitüübid — reportaažid, juhtkirjad, populaarteaduslik kirjandus, seikluskirjandus jne. kirjalike tekstide puhul, vahetu vestlus, telefonivestlus jne. suulise kõne korral (vt. eespool toodud korpuste kirjeldusi) — nende dimensioonide järgi rühmituvad. Niimoodi saadaksegi huvipakkuv tekstide tüpoloogia: tekstid jagunevad kaheksasse tüüpi, mille vahel jagunevad korpuste algupärased tüübid-žanrid, kusjuures näiteks selge vahe suuliste ja kirjalike tekstide vahel kaob. Üks ja sama žanr võib jaguneda — küll erinevates proportsioonides — mitme tüübi vahel. Näiteks tekstid, mis algupäraselt olid žanris «biograafiad», kuuluvad osalt 3. tüüpi «teaduslik esitus» (*scientific exposition*), osalt 6. tüüpi «üldine narratiivne esitus» (*general narrative exposition*), jne.

---

[25] P. de Haan, Postmodifying clauses in the English noun phrase. A corpus based study. Amsterdam—Atlanta, 1989.

[26] J. Järborg, Towards a formalized lexicon of Swedish. Rmt.: M. Gellerstram (toim.), Studies in computer aided lexicology. Stockholm, 1988, lk. 140—158.

[27] D. Biber, A typology of English texts. Linguistics 27. The Hague, 1988, lk. 3—43.

See uus klassifikatsioon esindab Biberi järgi «ehtsat» lingvistilist klassifikatsiooni, sest lähteks olid tekstide keelelised tunnused.

Eesti keele tekstikorpus?

Mainisin artikli alguses, et meil on ülim aeg, aga ka sobivam aeg kui kunagi varem asuda looma eestikeelsete tekstide korpust. Pidasin silmas eelkõige kaht asjaolu. Esiteks on korpuste loomine ja eriti nende kasutamine maailmas massiliseks muutunud; inglise keele uurijale on näiteks enesestmõistetav, et materjali oma töö jaoks saab ta mõnest korpusest. Tänu tehtud ja tehtavate tööde massilisusele on võimalus kasutada olemasolevaid rikkalikke ja piisavalt standardiseeritud kogemusi.

Teiseks oleme varustatud arvutustehnikaga juba küllalt hästi — pean silmas nii Tartu Ülikooli kui KKI-d — ja seda tehnikat on suhteliselt lihtne juurde hankida, kui käivitame nii mastaapse ja eesti keeleteaduse ning mõnes mõttes kogu eesti kultuuri seisukohalt olulise töö.

Oleks lihtsalt mugavus ja inertsus, kui me praegu asja ette ei võtaks.

Töö teine aspekt on hoopis keerulisem: kuidas ja missuguste jõududega see teostada? Eelnevast peaks selge olema, et korpuse loomine ei ole ühekahe aasta töö ja seda ei tee ära paari inimese igapäevase nokitsemisega. Korpuse tegemine nõuab hoolikat planeerimist. Tööd alustades peab olema selge, mida tahame tulemusena näha.

Minu arvates on meil eelkõige vaja «tänapäeva eesti kirjakeele korpust», mille loomine enam-vähem järgiks Browni korpuse, LOB-i jts. ideid. See peaks küll olema üksnes eesti keele tekstikorpuse tuum. Paralleelselt vajaksime näit. kõnekeelekorpust, murdetekstide korpust, vana kirjakeele tekstide korpust jt. Viimased jäävad siiski eelkõige nende õlule, kes vastavate uurimustega tegelevad: juba sellepärast, et need on spetsiifilised ja ei pea olema eriti mahukad. Tänapäeva kirjakeele korpuse loomine on aga märksa olulisem, see eeldab olemasolevate jõudude ühendamist (ülikoolist, KKI-st) ja olemasolevate materjalide ärakasutamist. Eesti keele tekstide arvutikorpuse loomine on praegu lülitatud Tartu Ülikooli eesti keele laborí teadustööde kavva, ja seni on see vist ka ainuke allüksus, mille plaanis niisugune ülesanne figureerib. Ent korpuse lõpliku sisu ja struktuuri fikseerimine nõuab ilmselt üldisema lingvistide konsiiliumi koostööd ja otsust.

Missuguseid olemasolevaid materjale saaksime kasutada? Kahjuks selliseid materjale praktiliselt pole. Näiteks KKI suur näitekartoteek — ehkki iseenesest kahtlemata väga hinnaline näidete allikas — ei sobi juba sellepärast, et sisaldab valdavalt üksiknäiteid, korpuse üks algpõhimõtteid on aga, nagu võisime veenduda, et see moodustub sidusatest tekstilõikudest. Iseenesest on tehtud ka piisavalt eesti keele alaseid arvutuslingvistilisi töid, mis opereerivad sidusate tekstidega. Eelkõige tuleb nimetada J. Tuldava juhendamisel TÜ-s tehtud keelestatistilisi uurimusi, nende hulgas eriti ilukirjanduslike tekstide autorikõne sagedussõnastiku koostamist. Kuid ka siinne lähtematerjal on kogutud hoopis teisi eesmärke silmas pidades (ehkki iseenesest võib seegi väga hästi moodustada eesti keele tekstikorpuse ühe kõrvalharu). Sama võib öelda KKI arvutuslingvistika sektoris leiduvate tekstilaadsete fondide kohta (näit. ENEKE-se mõisteartiklid).

Niisiis tuleb praktiliselt alustada päris algusest. Seda võib pidada loomulikukski, sest tekstikorpus peab moodustama süsteemse terviku.

Folkloristikas on kõik korras

(„Eesti folkloristika täna ja homme")

RUTH MIROV

Ennemuiste ka elati,
kui maada küüsil künnetie.

Oleme üle elanud selle aja, kui eesti arhitektidele kinnitati, et sisult sotsialistlik ja vormilt rahvuslik arhitektuur väljendub kõige paremini vene klassitsismi kaudu. Sel ajajärgul olid kõik inimsoo avastused ja leiutised tehtud geniaalsete vene iseõppijate poolt, kellel ainult tsaarivalitsuse masendavate tingimuste tõttu jäi au ja kuulsus saamata. Oleme üle elanud selle aja, mil viini saiast sai moskva sai ning berliini pannkoogist pontšik. Osa neist reaalidest on saanud ka rahvaluuleks, näiteks kas või seda liiki naljandid nagu «Ivan, ma näen sind läbi!» (röntgeni leiutamine). Küllap me saame seeditud ka need *snackbar*'idest müüdavad *hot dog*'id ja *hamburger*'id. Poleks eriline au selles ameerikalikus sulatusahjus massikultuuriks sulada.

Palju on räägitud sellest, kui kitsas ja suletud oli tavalise inimese elu eelmistel aastasadadel ja kui suured olid seisuste vahed. Ei olnud kiireid liiklus- ja levivahendeid. Kirjaoskajaid oli vähe, rääkimata haritud inimestest. Ja ometi — kui avatud oli elu eelmistel sajanditel! Nagu imekombel-nõiaväel levisid teated ja teadmised rahvalt rahvale, seisuselt seisusele. Puhtalt meie folkloori seisukohastki — kui palju on rahvusvahelist ainest muinasjuttude, naljandite, vanasõnade, mängude, laulude, loitsude ja kõige muu hulgas. Kõik on eesti keelde ümber pandud ja muuski mõttes ümberrahvustatud. Rahvas on võimas ja võimekas. Nagu muistend sõja- ja katkuaegadest jutustab, oligi kord ainult kaks inimest järele jäänud. Hea, et sattusid olema eri soost. Nii me kestame.

Hoolimata sellest, et meist on palju «aegu» ja sõjamasinaid üle veerenud, ei ole meie aeg siiski väga palju teistsuguseks saanud, kui oli Hurda aeg. Jakob Hurdal oli suur mure, et haritud eestlased saksastuvad. Nagu nüüdki on parema elu peale minejaid. See on mõistetav, kuid põhimõtteliselt siiski hukkamõistetav. Aga jätkus tookord Eesti asja edendajaid ja jätkub ka nüüd. Ühestainsast Hurdastki jätkus väga paljuks.

Ja üheainsa aastaga suutis Jakob Hurt eesti lapsekingades folkloristika viia maailma tähelepanu keskpunkti. 1888. aastal algatas Hurt rahvaluule suurkogumise ning 1889. aastal ülistati Euroopa kultuuri südames Pariisis folkloristide kongressil Hurda rahvaluulekogu ja kogumismetodoloogiat.

Humanitaaria trükkitoimetamiseks pole vist Eestimaal kunagi eriti soodsaid olusid olnud. Hurda ajal olid nad väga ebasoodsad nagu nüüdki. «Setukeste laulud» I—III (mis samuti saavutas rahvusvahelise tähelepanu ning on tänaseni arvestatav ka oma editsiooniprintsiipide poolest) nägi päevavalgust Helsingis tänu Soome Kirjanduse Seltsile ja eriti Kaarle Krohnile. Kust nüüd abi saada?

Rahvusteadusliku kirjanduse väljaandmisega oli Nõukogude Eestis kogu aeg raskusi, küll niisuguseid, küll naasuguseid. Praegusel üleminekuajal on olukord aga lausa košmaarne. Räägitakse, et pole raha, pole paberit, ja — äri enne muud. Ma ei hakka pidama ilukõnet, et rahvusteadustel peaks olema eelisõigused, kui tahetakse taasehitada rahvusriiki ning säilitada oma rahvust elujõulisena ja kultuurivõimelisena. See