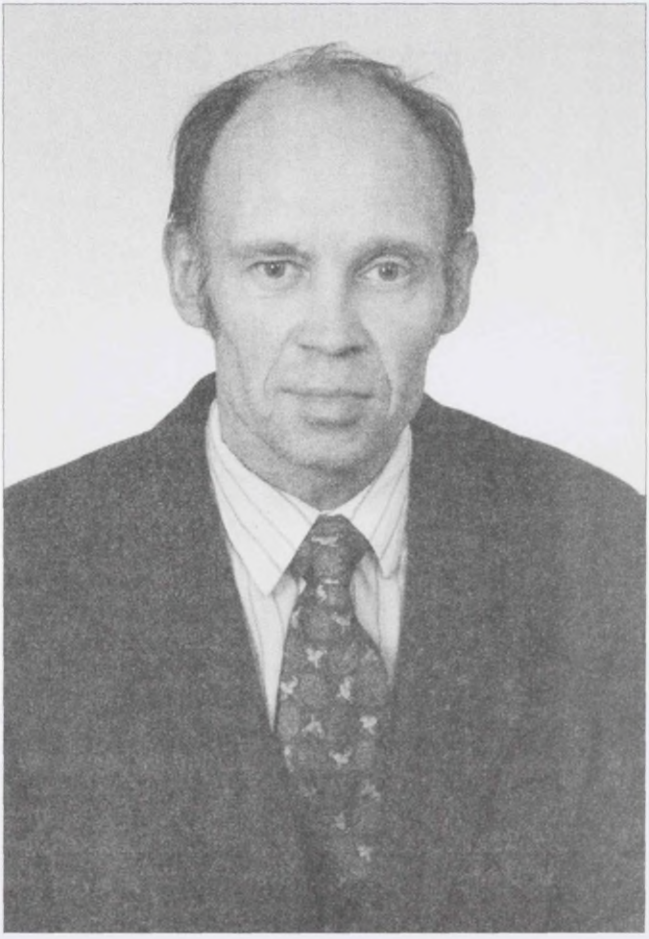
Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 3
Publications of the Department of General Linguistics 3
University of Tartu

Tähendusepüüdja
Catcher of the Meaning

Tartu 2002

Tähendusepüüdja

Pühendusteos
professor Haldur Õimu
60. sünnipäevaks



Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 3
Publications of the Department of General Linguistics 3
University of Tartu

Tähendusepüüdja
Catcher of the Meaning

Pühendusteos professor Haldur Õimu
60. sünnipäevaks 22. jaanuaril 2002

Festschrift for Professor Haldur Õim
on the occasion of his 60th birthday

Toimetajad / Edited by
Renate Pajusalu & Tiit Hennoste

Tartu 2002

Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 3
Publications of the Department of General Linguistics 3
University of Tartu

Tähendusepüüdja / Catcher of the Meaning

Pühendusteos professor Haldur Õimu 60. sünnipäevaks
Festschrift for Professor Haldur Õim
on the occasion of his 60th birthday

Toimetajad / Edited by Renate Pajusalu & Tiit Hennoste

Kujundaja / Layout Roosmarii Kurvits

ISSN 1406–619X
ISBN 9985-4-0230-8
Tartu Ülikooli Kirjastus
Tiigi 78, Tartu 50410
Tellimus nr. 1

Saateks

See raamat koosneb artiklitest, mille on professor Haldur Õimu 60. sünnipäevaks 22. jaanuaril 2002 kirjutanud tema sõbrad, kolleegid ja õpilased. Tartu ülikooli üldkeeleteaduse professori Haldur Õimu elust ja tööst on täpsem ülevaade antud esimeses artiklis. Kogumiku artiklite laialdane temaatika, mis eelkõige (aga mitte ainult) seostub tähendusega nii arvutilingvistikas kui keeleteaduse traditsioonilisemates valdkondades, näitab kaudselt ka juubilari kahte olulisemat omadust: mitmekesiseid teaduslikke huvisid ja avatust uutele teooriatele.

Toimetajad

To the reader

The volume consists of papers by friends and colleagues of Professor Haldur Õim to celebrate his 60th birthday on January 22, 2002. The first paper (in Estonian) deals more thoroughly with Haldur Õim's life and work – *inter alia* – professor of general linguistics at the University of Tartu. The wide spectrum of linguistic studies (both computational and non-computational), as covered by this volume, reflects Haldur Õim's manifold interests and openness to new theories.

Editors

Sisukord / Contents

Tabula gratulatoria

Ajakiri Akadeemia
Ajakiri Keel ja Kirjandus
Jens Allwood
Angelina Tshaikovskaja
István Bátori
Niels Ole Bernsen
M. V. Boldasov
Eugene H. Casad
Svetlana Dikareva
Laila Dybkjær
Eesti keele instituudi leksikograafiasektor
Eesti keele instituut
Eesti Teaduste Akadeemia
Tiiu ja Mati Erelt
Turid Farbregd
Olga Gerassimenko
Külli Habicht
Auli Hakulinen
Helsingin yliopiston suomalais-ugrilainen laitos
Tiit Hennoste
Annika Hussar
Airi Jansons
Heiki-Jaan Kaalep
Neeme Kahusk
Kaarel Kaljurand
Marja Kallasmaa
Fred Karlsson
Reet Kasik
Krista Kerge
Valve-Liivi Kingisepp
Birute Klaas
Mare Koit
Arvo Krikmann
Maret Kullasaar
Kaja Kährik
Matti Leiwo ja Jyväskylän yliopiston suomen kielen laitos
Liina Lindström

Maia Madisso
Einar Meister
Helle Metslang
Meelis Mihkla
Kadri Muischnek
Sinikka ja Jussi Niemi
Ellen Niit
Heili Orav
Oulun yliopiston suomen ja saamen kielen ja logopedian laitos
Hille Pajupuu
Renate ja Karl Pajusalu
Külvi Pruuli
Peeter Päll
Tiit Roosmaa
Kristiina ja Jaan Ross
Andriela Rääbis
Mare Sepp ja Jüri Valge
S.A.Sharoff
E.G. Sokolova
Krista Strandson
Seppo Suhonen
Helena Sulkala
Margit ja Urmas Sutrop
V.A. Zhigalov
Kaja Tael
Tallinna pedagoogikaülikooli eesti filoloogia osakond
Tallinna pedagoogikaülikooli idamaade keskus
Tallinna pedagoogikaülikooli idamaade õppetool
Jüri Talvet ja Tartu ülikooli kirjanduse ja rahvaluule osakond
Tamperen yliopiston kieli- ja käännöstieteen laitos
Tartu ülikooli eesti keele õppetool
Tartu ülikooli eesti keele (võõrkeelena) õppetool
Tartu ülikooli eesti kirjanduse ning teatriteaduse ja kirjandusteooria õppetoolid
Tartu ülikooli filosoofiateaduskond
Tartu ülikooli keelekeskus
Tartu ülikooli keeletehnoloogia õppetool
Tartu ülikooli kognitiivse lingvistika seminar
Tartu ülikooli vene keele õppetool

Urve Talvik
Silvi Tenjes
Piret Toomet
Ilona Tragel
Turun yliopiston suomalaisen ja yleisen kielitieteen laitos
Ene Vainik
Silvi Vare
Ann Veismann
Enn Veldi
Riina Vellerind
Anna Verschik
Kadri Vider
Jüri Viikberg
Ülle Viks
Krista Vogelberg
Martin Volk
Evely Vutt
Erling Wande
Kalevi Wiik
Eberhard Winkler
Jaan Õispuu

Teadusele pühendunud

Tutvusin Haldur Õimuga 1965. aastal, kui Huno Rätsepa eestvõttel alustas Tartu Ülikoolis tollase eesti keele kateedri juures tegevust seminar, kus hakati käsitlema uudseid meetodeid keeleteaduses, eeskätt Chomsky generatiivseid grammatikaid. Hiljem sai selle seminarirühma nimeks generatiivse grammatika grupp. Rühmas osales nii lingvistika- kui ka matemaatikaõppejõude ja -üliõpilasi. Haldur Õim õppis siis individuaalprogrammi alusel strukturaalset ja matemaatilist lingvistikat. Mäletan seminaridest tema sagedasi, hästi põhjendatud sõnavõtte ja vaikset häält.

Pikast tutvusest hoolimata ei ole ma kuigivõrd pädev asjakohaselt hindama Haldur Õimu kogu mahukat tegevust ja tema teaduslikku panust. Alljärgnev on eeskätt minu subjektiivne nägemus sellest mitmetahulisest õpetamis- ja uurimistööst.

Haldur Õim sündis 22. jaanuaril 1942 Valgamaal Helme vallas. Käis Madi algkoolis, Tõrva keskkoolis ja astus 1960 Tartu Riiklikku Ülikooli õppima eesti filoloogiat. Ta lõpetas ülikooli 1965. aastal ja jätkas samas õpinguid aspirantuuris 1966–1969. Kandidaadiväitekiri "Isiku mõistega seotud sõnarühmade struktuur eesti keeles" valmis tähtajaks (mis ei olnud tavaline tollal ega ole seda ka praegu) ning selle eduka kaitsmise järel omistati H. Õimule 1970. aastal filoloogiakandidaadi teaduslik kraad.

Pärast aspirantuuri lõpetamist asus Haldur Õim vanemteadurina tööle TRÜ kriminoloogia laboratooriumis, kus oli keeleeksperdiks eesti- ja venekeelsetest juriidilistest dokumentidest info otsimise süsteemi JURIPS väljatöötamisel ja arvutil juurutamisel.

1977. aastal alustas ta tööd eesti keele kateedris, algul vanemõpetajana, 1983. aastast aga dotsendina.

1983. aastal kaitses Haldur Õim filoloogiadoktori väitekirja semantikast ja keele mõistmise teooriast ning 1985. aastal omistati talle professori kutse.

Praegu on ta Tartu Ülikooli eesti ja soome-ugri keeleteaduse osakonnas üldkeeleteaduse professor ja 1994. aastast ka Eesti Teaduste Akadeemia liige.

Lisaks on Haldur Õim töötanud pikemalt ka välismaal: 1975–1976 Zürichi ülikoolis järeldoktorina, 1981–1984 Helsingi ülikooli

eesti keele lektorina. 1991–1992 oli ta külalisprofessor Saksamaal Koblenz-Landau ülikoolis.

Professor Õim on avaldanud ligi 200 teaduspublikatsiooni. Esimese artikli avaldas ta juba üliõpilasena, 1964. aastal, ja selle pealkiri sobib sümboliks: *Keeleteadus ja matemaatika*. Kui püüda leida neid teemasid, mis lähevad läbi Haldur Õimu teaduslike tööde, siis saame kolm selget rida: semantika, pragmaatika ja arvutid.

Haldur Õim alustas semantikuna. Tema varased artiklid kõnelevad sõnade semantilisest struktuurist ja tema esimene suur töö on monograafia (artiklina avaldatud kandidaadiväitekiri) *Isiku mõistega seotud sõnarühmade semantiline struktuur eesti keeles* (1971). Tema esimene rahvusvahelist tähelepanu pälvinud artikkel *On the semantic treatment of predicative expressions* ilmus 1973. aastal kogumikus *Generative Grammar in Europe*. Ka edaspidi ei kao tähendused ja semantika tema töödest kuhugi.

Aga juba 1970. aastate alguses ilmub tema tööde pealkirjadesse ka pragmaatika mõiste, nt artiklites *Keel, keeleteadus ja pragmaatika* (1973) ning *Towards a theory of linguistic pragmatics* (1977).

Samal ajal lisandub kolmas põhiteema – inimese ja arvuti suhtlus – ning koos sellega küsimused, kuidas mõista teksti, kuidas on üles ehitatud dialoog, suhtlus, kuidas on suhtlemiseks vajalikud teadmised organiseeritud inimese ajus.

See tee viib Õimu 1970. aastate teiseks pooleks teoreetilise keeleteaduse uude paradigmasse, kus kõneldakse freimidest, teksti mõistmisest jms. Ilmuvad artiklid nagu *Language, meaning and human knowledge* (1981) ajakirjas *Nordic Journal of Linguistics*, mis ennustab uut paradigmat keeleteaduses ja kõneleb vajadusest üle minna kognitiivsele keeleuurimisele. Ilmub koos Madis Saluveeruga kirjutatud artikkel *Frames in linguistic descriptions* (1985) ajakirjas *Quaderni di Semantica*, milles vaadeldakse uute keelekirjeldusüksustega (freimid, stsenaariumid) seotud probleeme lingvistilisest seisukohast ja viidatakse freimiloogika vajadusele ning piirangutele tema rakendamisel. Selle teema viljelemise üheks tipphetkeks on tolle aja nõudmiste taustal väga noorelt ja sealjuures üliedukalt kaitstud doktoriväitekiri *Semantika ja keele mõistmise teooria. Eesti keele direktiivse suhtluse leksikoni ja tekstide analüüs* (1983).

1970. aastate lõpus kujuneb välja Haldur Õimu juhitav uurimisrühm, kellega koos on kirjutatud hulk artikleid arvuti ja inimese suhtlemisest: inglise filoloog Madis Saluveer ning matemaatikud-

informaatikud Sergei Litvak, Tiit Roosmaa ja allakirjutanu. Selle grupi tööna valmis 1980. aastatel eksperimentaalne eestikeelseid tekste mõistev arvutiprogramm, milles realiseeriti tol ajal uudsed tähenduse esitamise skeemid – freimid. Sellest grupist on tänaseks välja kasvanud üldkeeleteaduse õppetooli juures tegutsev arvutilingvistika uurimisrühm.

1980. aastate lõpupoole läheb Õimu uurimine üha enam dialoogikeskseks ning märksõnadeks saavad suhtlusstrateegiad, kognitiivne semantika ja naiivsed teooriad.

Oluline suund, mida ta on arendanud kuni viimase ajani, on naiivsete teooriate mõiste ja sellel põhineva keelekontseptsiooni arendamine. Sellel teemal on ilmunud mitmeid artikleid. Artiklis *Naïve theories: a conception of cognitive semantics*, mis ilmus ajakirjas *Cognitive Linguistics* (1995), esitatakse seisukoht, et inimesed loovad ja kasutavad oma igapäevasel kogemusel rajanevaid mudeleid, nn naiivseid teooriaid, mis juhivad nende käitumist, ning et sellised inimeste mentaalset maailmapilti väljendavad teooriad on loomulike keelte põhialuseks.

Teiseks põhiteemaks on olnud suhtlusstrateegiad, mis juhivad inimese ja arvuti loomulikus keeles suhtlemist. Seda suunda kajastavad mitmed ettekanded rahvusvahelistel konverentsidel, näiteks (koos allakirjutanuga) *An approach to the organization of natural reasoning* (AIMSA 1990) ja *Developing a model of dialogue* (LREC 2000), mis käsitlevad inimese käitumist reguleerivatest printsiipidest lähtuvate suhtlusstrateegiate modelleerimist arvutil ning nende rakendamist inimese suhtlemisel arvutiga.

Haldur Õimu puhul on silmapaistev see, et temas on ühendatud harva esinevad võimed, ühelt poolt originaalne mõtlemine ja oskus uut avastada, ning teiselt poolt oskus keerulisi teaduslikke teooriaid a teistele selgitada. Viimasele poolele jäävad muu hulgas kaks ülevaateraamatut: *Semantika* (1974) ning *Inimene, keel ja arvuti ehk kompuuterlingvistika* (1983). Siia kuuluvad ka omal ajal diskussioone alustanud ja hiljem mitmete üliõpilaspõlvede poolt loetud sissejuhatav ülevaade *Teoreetilise keeleteaduse vanast ja uuest paradigmast* (1981) ning samasugune sissevaade *Pragmaatika ja keelelise suhtlemise teooria* (1986). Samasse sarja kuulub ka rahvusvahelisele üldsusele kirjutatud ülevaade *Language understanding and problem solving: on the relation between computational linguistics*

and artificial intelligence arvutilingvistika käsiraamatus *Computational Linguistics. An International Handbook* (1989).

Just Haldur Õimult on saanud Tartu ülikoolis pildi sellest, mis on keeleteooria ja teoreetiline lingvistika, mitu põlvkonda üliõpilasi. Sellest on ilmunud ka mõned artiklid, nagu *Keeleuurimine ja keeleteooria läbi aegade* (2000), *Eesti keeleteadusliku mõtte areng XX sajandil* (2000). Jääb veel oodata ülevaatlikku raamatut keeleteooriast.

1990. aastatel on Haldur Õim olnud ka suurte arvutilingvistiliste projektide organiseerija, mida ta on läbi viinud koos oma nooremate kolleegidega. Just tänu prof Õimu isiklikele kontaktidele ja tema rahvusvahelisele tuntusele on TÜ ja Eesti arvutilingvistid saanud osaleda mitmetes Euroopa Liidu jm keeletehnoloogiaprojektides. Värskematest väärib mainimist näiteks projekt EuroWordNet, mille raames alustati eesti keele semantilise andmebaasi väljatöötamist. Kui 1997. aastal käivitus Eesti Informaatikakeskuse poolt koordineeritud Eesti keeletehnoloogia sihtprogramm, siis sai Haldur Õimust üks selle eestvedajaid.

Professor Õim on olnud aktiivne võitleja eesti keele püsimajäämise eest infotehnoloogilises ühiskonnas. Tema juhtimisel ja osavõtul on valminud Eesti keeletehnoloogia arenduskava, palju kordi on ta sel teemal sõna võtnud ajakirjanduses ja mitmesugustel seminaridel. Tema esinemised on alati tasakaalukad, sisaldavad selget sõnumit ja veenvaid põhjendusi.

Professor Õimu eestvõttel on Tartu Ülikoolis töötatud välja arvutilingvistika eriala õppekava ja käivitatud arvutilingvistide süstemaatiline ettevalmistamine eesti ja soome-ugri keeleteaduse osakonnas alates 1997. aastast. Ta ise õpetab mitut kursust arvutilingvistika üliõpilastele, tõmmates neid ühtlasi kaasa uurimistöösse. Eestvedajana ja innustajana on ta suunanud arvutilingvistika uurimisrühma, koondanud sinna noori uurijaid, kavandanud ja toetanud nende arengut. Minu arvates on see parim, mida saab teha Eesti teaduse järjepidevuse heaks.

Juubilarile kestvat vaimuvärskust soovides,
kolleegide nimel
Mare Koit

Language & time

Jens Allwood
Göteborg University

Introduction

Language and time are related in many ways. This paper presents some reflections on two of the most important ways in which language and time are related. In the first case, the point is that language is a phenomenon which itself is found in time and is used and developed in time. Below I will briefly point to four aspects of the time relatedness of language. In the second case, language is a means for structuring, representing, and conceptualizing time. Language is one of our chief instruments (probably the most important one) for understanding time. The second part of the paper, thus, discusses a kind of methodological prerequisite of the first part, making use of language in a kind of "metalinguistic" manner to relate to the first part.

1. Language in time – use and evolution

Human languages are not constant phenomena but are continually undergoing dynamic processes of use and evolution. Let me briefly introduce four types of dynamics associated with human language. With terms derived from Latin, the four types can be designated *phylogeny*, *macrogeny*, *ontogeny*, and *microgeny*.

Phylogeny

Phylogenetic dynamics concern those millennia of development which can be associated with the origin of human speech. Some researchers see a connection between the origin of humans as a species *(Homo sapiens sapiens)* and the origin of human language. Human spoken language was a factor which probably had survival value for the human species. Through the development of language humans gained access to a gradually improved capacity for collective information processing. The surrounding world could be observed, categorized, and discussed collectively. Collective plans and undertakings became possible. The possibilities for complex

cooperation improved radically. Through language in effect a collective memory was created. The improved possibilities for cooperation provided by language, thus, gave humans access to an important survival factor. The importance of cooperation for survival can perhaps be seen even more clearly in other species, for example, in ants, who despite being a relatively simple organism, with a simpler means of cooperation than that of humans, have survived on earth very effectively for a long time.

In a similar way, through language humans gained the ability to survive by means of cooperation. Cooperation through spoken communication is also connected with the development of another fundamental trait of humans, namely their social nature. Humans do not evolve in isolation, but in interaction with others, and this interaction is most often spoken.

Macrogeny

By macrogeny, I mean the evolution of the language of a particular group. This is a process which, at least for groups which are not too small in number, usually takes centuries. We can, for example, compare the English of 800 years ago with the English of today. The process of group change is for the most part greatly dependent on different types of influence from the surrounding world, which most often carry with them linguistic influence as well. It is also dependent on the economic, technological, and political dynamics within a group. The development of written language has played an especially important role, primarily in that it makes the macrogenetic changes slower. The usage of earlier generations can be preserved through writing and can exert an influence on later generations. This influence has probably been strengthened by printing and by the advent of generally accessible dictionaries and grammars. Secondarily, writing, especially after the rise of the mass media, has also made the more rapid spread of linguistic innovations possible. However, these innovations are also preserved by writing, instead of disappearing, as they might have done in a spoken language culture without writing.

Macrogenetic linguistic evolution can pertain to groups other than national groups. It can apply to the "dialects" of regional groups or the language within social institutions such as government entities and corporations (the language within public health services or at a

car company). It can also pertain to language within a certain activity or genre such as, for example, the development and change of the language of instruction or the language in a cookbook.

Ontogeny

A third type of dynamics concerns an individual's development of his linguistic abilities from infancy to old age.

We develop our language throughout life. There is always something to learn. We learn the most, the easiest, and the quickest during the very early ages (0–5 years). The high pace persists up into one's teens, after which, for most people, it decreases somewhat but, provided one is not stricken with illness, the ability to learn and develop language never wholly disappears.

Microgeny

A considerably swifter linguistic dynamic than the three we have discussed up to now is found in ordinary conversation. How do we develop internal impulses into externally accessible messages through gestures, speech, or writing? The course of events is so fast that most often we do not have a clear picture of whether thoughts precede words or if rather thoughts and speech are articulated simultaneously and in unison. The dynamic also comprises how others consciously and unconsciously are influenced by and react to what we say, and how we work jointly and structure both the relations between us and the content about which we communicate.

2. Language as a means for structuring, representing, and conceptualizing time

After having, thus, contemplated the fact that language is a temporal phenomenon with processes that extend over millennia (phylogeny), centuries (macrogeny), decades and years (ontogeny) and micro-seconds (microgeny), we will now move on to the question of how we with the aid of language structure, represent and conceptualize time. This question can be formulated more precisely in terms of two main questions: (i) which categories of linguistic expression (linguistic means) are used to structure time in different languages, and (ii) which categories of temporal content are found in different languages?

Linguistic means of expression

Some of the relevant linguistic means are the following:

(i) Affix: for example, tense endings as in *I talk* (now–present) and *I talked* (past time – preterite)

(ii) Vowel change: for example, *sit, sat*

(iii) Reduplication: for example, *he ran and ran and ran* to indicate that a process is extended over time.

(iv) Simple and compound words: for example, (adverbs like *now, then, tomorrow, yesterday, yesteryear,* nouns like *second, minute, hour, week, month, year, afternoon, fortnight,* adjectives like *long, short-lived, etc.*

(v) Intonation: for example, vowel lengthening to show a long time duration.

(vi) Body movement: for example, quick movements to show that something happens quickly.

(vii) Implied correspondence between that which is expressed and the actual course of time. Consider for example the sentences *Olle ate a sandwich and went to bed* and *The man jumped up, ran, and stopped.* In both these types of examples we assume that the described series of events are also in time sequence. That which is mentioned first happened first.

(viii) Information that is not expressed linguistically at all but that can be inferred from the speech situation, for example, the point in time when something is said.

Of all linguistic expressions for time, perhaps one can say that tense markers have been discussed the most. Tense markers are found on verbs and indicate how the process the verb stands for should be temporally anchored. Adverbs of time probably come in second place. These can be more freely combined than tense with all parts of speech, for example, *food yesterday, drink today.*

From the examples we see that in English, one can express time relations in many different ways. This also pertains to other languages. There are often many ways to express and formulate time in a given language. There is also variation regarding the ways which are used most in a certain language. Sometimes a certain way to express time is lacking. In Chinese, for example, time is not expressed as in English and many other languages, by linguistically

changing the forms of verbs, so that the process that is being described can be anchored in time in various ways. Instead, adverbial expressions such as *now, then, tomorrow*, etc., are used to a higher degree than in English, together with letting situationally given time information be implied by what is said, for example through agreement between linguistic description and the sequence of events (as in the English examples above) or through the situation necessitating a certain temporal anchoring.

One might wonder, then, whether the difference in modes of expression reflects something more than a random variation. Many linguists have maintained that this is the case. The above-described list (i)–(viii) is believed to reflect degrees of integration in a linguistic system. Whatever is expressed through grammatical means (for example, inflectional and derivational morphemes, or prosodic or syntactic patterns) is more integrated in a language than what is found in the vocabulary, which in turn is more integrated than what is implied contextually or situationally. If one limits one's perspective to inflectional and derivational morphemes (and disregards syntax and prosody), this implies for the examples discussed above that Chinese has a less integrated need for temporal anchoring in its linguistic system than English.

Linguistic structure of temporal content

A good starting point in discussing linguistically structured temporal content is the relationship between time and change.

Without change, it is uncertain whether we would have any concept of time, and perhaps one can say that if time did not exist, no change could occur. Time and change appear to be analytically conceptually bound together. Changes show themselves most distinctly in processes, occurrences, and courses of events. It is therefore natural that verbs (those words in the language which indicate processes and courses of events) have an especially close relationship to temporal anchoring and temporal duration.

Let us now consider two fundamental ways by which a speaker (or experiencing subject) can perceive change.

(i) I (the experiencing and possibly speaking subject) change in relation to the environment.

(ii) Space changes in relation to me.

The first way is associated with the constant stream of new contents in attention and consciousness, while the other is associated with changes which are independent of human beings, such as the continuous change of night to day and the change of the seasons. These are phenomena which, independent of our own activity, force an experience of change on us.

These two points of departure for the experience of change are connected to three different methods used in human languages for temporal anchoring, where the third method consists in a combination of the first two.

(i) The point of departure is the "now of the ego" in the speech situation – a perpetually ongoing "flow of experience/attention" or "now-flow". This leads to so called deictic (pointing or indicating) temporal anchoring.

(ii) The point of departure is an external way to measure time – the sun, a clock, etc. This leads to so called calendric temporal anchoring.

(iii) The point of departure consists in a combination of the "now of the ego" and an external way to measure time.

The first point of departure is the one we find behind all so-called "deictic" time expressions, for example, tense affixes and adverbs of time. The point of departure is the speaker's (or experiencer's) "now" and all of the deictic words which receive their anchoring relative to this "now". The preterite *I went yesterday* indicates a point in time before "now", and the pluperfect *I had already gone yesterday* indicates a point in time which lies before this prior point in time. The present *I go* indicates that the event spoken about includes the "now" and the future *I will go* indicates a point in time that lies after the "now".

With this system, time is structured as a sort of line where the point of departure, the "now" moves from "prior or earlier" to "after or later." Consequently, each time indication becomes relative to the speaking subject. Two classic accounts of the system, from two different perspectives, are Bühler (1934) and Reichenbach (1947).

The other way to temporally anchor ourselves is to use some external way to measure time for the anchoring; let us call it "calendric anchoring". Instead of saying *now* or *again*, we say the fourteenth of October 2001, or the thirtheenth of October 2001. The

point of departure here, historically, is likely to be astronomical – diurnal rhythms and annual rhythms, which were later strengthened by the invention of different artifactual ways to measure time. From this perspective, differences arise concerning the expression of time in different languages, because people in different cultures have had different ways to measure time and have had different ideas about which holidays they wished to observe. Before modern clocks were constructed, there was not much reason to have words for seconds or even minutes and hours. For similar reasons, there has not been any great reason for non-Christian cultures to have words for Christmas Eve or Good Friday. Since even the seasons are experienced with different degrees of distinctness in different parts of the world, there are not always words for *summer*, *fall*, *winter*, and *spring* but rather, for example, for *rainy season*, *dry season*, or *monsoon season*.

The third way to bring about temporal anchoring occurs through combining the calendric way of anchoring with the deictic, which can be observed in expressions such as:

the day before yesterday	yesterday	today	tomorrow	the day after tomorrow
	last week	this week	next week	
	last year (previous year)	this year	next year	

The point of departure is a time interval including "the now-flow", and then one uses external temporal anchoring is used in addition – diurnal change, the ability to count (a week), or seasonal change to secondarily anchor the event being referred to.

Conceptual categories related to subjective temporal anchoring

Through deictic temporal anchoring, the point of departure for temporal anchoring becomes a the "now-experience" or "flow of attention" of a self. These make it possible and natural to combine time with other elements in the "now-experience". Such elements are provided, for example, by the experiences of certainty, uncertainty, intention, will, and feelings of obligation, which can be associated with "now-flow." The link is so natural that it has influenced our ways of linguistically structuring time. Time expressions (e.g., tense

expressions) are therefore in many languages difficult to distinguish, for example, from what often is called "aspect" and "modality."

By "aspect" (sometimes the term *aktionsart* is used) we mean factors which have to do with the temporal structure of a course of events. Do we regard an event as a process or as the end state which the event has resulted in? Do we see an event from the outside as some completed whole, or do we see the event from the inside as something on-going? Are we speaking of an isolated event or of repeated, routine events and actions? With some reflection, one can perhaps see why tense (time) and aspect often fit together. An event that is completed is often an event in the past, while an event that is uncompleted often points to a future event. One can therefore use linguistic endings for tense and aspect (or choice of verb type – *aktionsart*) in order to, depending on the circumstances, express something about the inner structure of a process in relation to time. From a linguistic semantic point of view, concepts like the ones below have been used in order to capture how different linguistic forms relate to processes. (See further, Dahl 1985; Comrie 1976; Lyons 1977.)

1) change (dynamics) – state (statics), e.g., the difference between *Peter is running* and *Peter is happy.*
2) duration , e.g., *Peter runs for an hour.*
3) beginning, e.g., *Peter begins to run.*
4) continuation, e.g., *Peter is running now.*
5) completion, e.g., *Peter has finished running.*
6) punctuality, iterativity, e.g., *Peter gave a start – Peter runs often.*

Temporal anchoring is also closely bound up with what is usually called "modality." By "modality," we mean phenomena that have to do with that which is necessary or possible, certain or uncertain, obligatory or permitted, desired or undesired, intentional or unintentional.

Once more there is a connection between phenomena of this type and temporal anchoring. If I want to go to the movies, then with a certain likelihood I am going to (will) do so. To say something about "what one wants to do" can in this way be a way to say something about "what one is going to do." In English, this process has gone so far that *I will* most often has nothing at all to do with desire, even though it originally did. In the same way, something that

is an obligation, *I shall*, can also cross over to express that something will occur in the future.

Further, there is a relationship between epistemic modality (certainty–uncertainty) and time. That which happens now tends to be what one is most certain of, while that which has happened or will happen has a less certain status. In certain North American Indian languages, there is precisely this coupling of epistemic status and temporal anchoring. In Swedish and English there is also a coupling of this kind, though it moves rather in the opposite direction, going from temporal anchoring to the state of knowledge. If someone says *Olle kommer att vara hemma nu* (*Olle is going to be at home now*), *kommer att* (going to) expresses the future, but since the word *nu (now)* also is found in the utterance, we reinterpret *kommer att* to mean approximately "if we check," which gives the entire utterance a hypothetical epistemic status. In a similar way, in Swedish one can say *om jag var dum skulle jag gå* (*if I was (were) stupid, I would go*) and use the preterite of the verb *vara* not to express the past time but rather to express counterfactuality in the present. The past's non-presence, in combination with the word *om (if)*, causes the utterance to be interpreted as a possibility. English is different here since it has the subjunctive *were*.

As has been suggested above, then, there are interesting connections in many languages (inluding English and Swedish) between how one expresses the categories of content underlying what linguists call tense, modality, and aspect. The future has a connection with the uncertain, with what one intends, wants, or ought (must, obligation) to do, as well as with that which is on-going and not completed. The now or the present has a connection with the certain and the on-going. The past has a connection with the uncertain and the completed but can also be used to express that which is unreal or only a supposition in the present.

Other ways to express time linguistically

Time is expressed linguistically not merely through different ways to temporally anchor a course of events, but also through those qualities, processes, and relations which we tend habitually, in the language, to ascribe to time.

The following are some qualities that are relatively often ascribed to time in English.

scarce, short time
good, long, eternal time
strange, difficult, changing times

In all cases, these qualities seem to have to do with some course of events the speaker is engaged in. This course of events has a duration, and words such as *short* and *long* specify its extent in relation to some actor in the course of events, e.g., the speaker her/himself. Words such as *strange*, *difficult*, or *changing* specify instead a set of evaluations which the speaker makes about different events or states in a certain period of time. If we look at processes where time is attended to, one says for example that

time flies, passes, lasts
time runs out, is up
the time is at hand, is approaching

Only in the first type of example does the process seem to be an attempt to speak about time-in-itself and its relationship to the deictic "now-experience". In the latter two examples, on the other hand, the most natural interpretation is that it has to do with an actor's perspective on a certain course of events with a finite extension of time. A less probable interpretation of the second type of example of processes would be that time as a conceptual or physical quantity will no longer exist.

Finally, if we look at the linguistic structuring of relationships to time, this happens to a large extent with the help of prepositions, for example:

for, in, during, after	an hour
on	the stroke of twelve
at	about twelve
through	the centuries

These prepositions have been claimed by many authors (see, e.g. Lyons 1977) to be primarily spatial. They have then, through some sort of metaphorical process, been extended to time, where the differences that are found between their meanings in a spatial context receive a partly different interpretation. In relation to time, the differences become roughly "aspectual" (see above), i.e., they deal with whether an action is, for example, completed (in), on-going (for) or quite simply localized within a certain period of time (during).

He wrote the letter in an hour. (the action concluded after an hour)
He wrote for an hour. (on-going, continuous action)
He wrote only a line during one hour. (relevant period for the action is one hour)

The linguistic structuring of relationships to time is often also the point of departure for speculations about whether time is essentially discontinuous (a series of points) or continuous, structured in a flow of intervals. The discontinuous perspective is supported by a calendaric temporal anchoring and can linguistically be expressed through a "bare time expression" such as, for example, *12 o'clock* or *the year 1200*, possibly with the addition of adverbs such as *precisely*, *exactly*, or *just*. In much the same way, the continuous perspective is supported naturally by a deictic anchoring (in an on-going "now-flow"), for example, *He ran to the city for an hour/in an hour*, and by the relationships most of the prepositions contribute to the structure of time, such as whether it consists of intervals with duration rather than of points without duration.

References

Bühler, K. 1934. Sprachttheorie. Jena: Fischer.
Comrie, B. 1976. Aspect. An Introduction to the Study of Verbal Aspect and Related Problems. Cambridge: Cambridge University Press.
Dahl, Ö. 1985. Tense and Aspect Systems. Oxford: Basil Blackwell.
Lyons, J. 1977. Semantics, Vol. 2. Cambridge: Cambridge University Press.
Reichenbach, H. 1947. Elements of Symbolic Logic. New York: The Free Press, & London: Collier, MacMillan.

Modellierung der Etymologien – von der Suche nach Ur-Sinn zu etymologischen Datenbasen

István Bátori
Universität Koblenz-Landau

Zielsetzung etymologischer Wörterbücher und die Modelle

Die nachfolgenden Überlegungen entstanden im Rahmen eines Arbeitsvorhabens für die Erstellung einer etymologischen Datenbasis. Einigkeit herrscht lediglich darüber, dass die Zielsetzung der Klärung (Enthüllung) der Herkunft der Wörter ist, aber wie diese Zielsetzung instrumentalisiert wird, bleibt selbst in modernen etymologischen Werken unausgesprochen.

Der Überblick über die Entwicklung von naiven Wortdeutungen bis zu der wissenschaftlicher Etymologie erhebt nicht den Anspruch auf Vollständigkeit. Die vorgestellten diversen Ansätze, Vorgehen, Methoden und Techniken sind kein Selbstzweck, sondern sie signalisieren die Parametrisierungsdimensionen für ein modernes db-orientiertes etymologisches Speicherungs- und Retrievalsystem.

1. Die antike Etymologie: Etymologie als Worterklärung

Die **Suche der ursprünglichen Bedeutung** ist die älteste, bereits bei den alten Griechen betriebene) Urform des Etymologisierens: Die Menschen waren offensichtlich schon immer an der richtigen (zutreffenden) Bedeutung der Wörter interessiert, die sie mit der ursprünglichen (alten) Bedeutung gleichgesetzt haben, die in dem grauen Alltag verdorben worden, oder verloren gegangen ist. Der antike Etymologe sollte den ursprünglichen Sinn des Wortes herausfinden. Er hatte keine methodischen Restriktionen, allerdings auch keine besondere Hilfsmittel. Er dürfte alles auswerten. In der Praxis wurden allerdings nur schriftlichen Quellen verwertet.

Das antike Modell ist nicht an Lexikon gebunden: Die ersten Etymologien, die als die Grundsteinlegung der Disziplin angesehen werden, sind in den Werken von Heraklit, Homer oder Sokrates zu finden. So wird z.B. in Kratylos über die Richtigkeit des Namens von *Hermogenes* "vom Stamme des Hermes" diskutiert (Pisani 1975: 11–35).

Festzuhalten sind die folgenden Modelleigenschaften:

1. Das Modell ist **informell und ohne Restriktionen:** Die Etymologie ist Improvisation in der Form von unformatiertem, einfachem fortlaufendem Text.
2. Gesucht wird die ursprüngliche unverfälschte Bedeutung und die ursprüngliche unverdorbene autung.
3. **Konsultiert wird die untersuchte Sprache selbst.** Obwohl das Heranziehen von anderen Sprachen nicht untersagt wird, werden in der antiken Praxis lediglich das Lateinische und Griechische verwertet.

Das naive Etymologiemodell lebt bis zu heute fort und prägt den Etymologiebegriff. Die antike Etymologie, wie naive Sprachmodelle im Allgemeinen, ist semantikzentriert.

2. Etymologie im Mittelalter: Abstammung

Im christlichen Mittelalter stand die **Bibel** im Mittelpunkt der Gelehrsamkeit und sie prägte auch das sprachliche Weltbild.

Erstens war die Bibel ein sakraler Text und zweitens war die Bibel eine alte Überlieferung. Dank der Bibel war das **Hebräische** als besonders alt, älter als alle anderen Sprachen angesehen und nach dem scholastischen Weltbild stammten alle Sprachen von dem Hebräischen ab. Demnach war die eigentliche Zielsetzung der Etymologie nicht die Klärung der Urbedeutung der Wörter, sondern deren Ableitung aus dem Hebräischen. Das scholastische Etymologiemodell beruhte auf der Annahme, dass alle Sprachen aus dem Hebräischen hervorgehen.

Die Scholastik erhob keine besondere erkenntnistheoretische Ansprüche. Die Wissenschaft (und darunter die Aufstellung der Etymologien,) sollte lediglich Instanzen der göttlichen Weltordnung demonstrieren. Da die hebräische Abstammung aller(?) Sprachen sachlich verfehlt war und sie mit den scholastischen Methoden gänzlich verworfen worden ist, wurde übersehen, dass sie trotz aller methodischen Mängel und Unzulänglichkeiten das Etymologiemodell mit dem Abstammungskonzept bereichert hat. Abstammung ist eine **vertikale Relation zwischen Sprachpaaren** (zwischen Hebräisch und den einzelnen Vernacularen).

Demnach stützt sich die etymologische Erklärung auf die Existenz der **Wörter in einer Ursprungsprache, aus welchen die gegenwärtigen Wörter abgeleitet werden.**

3. Etymologie als Verwandtschaftbeweis

In der Neuzeit vermehrte sich das Sprachwissen und es entstanden neue Voraussetzungen und Anforderungen:

- Die Humanisten entdeckten nicht nur die klassische Antike, sondern zeigten auch ein erhöhtes Interesse für das eigene Vernacular.
- Die hebräische Abstammung der Sprachen implizierte auch ihre Verwandtschaft und erwies sich unhaltbar und unpraktikabel.
- Dahingegen entstand ein Ordnungsbedarf für die wachsende Anzahl von Sprachen, die im Laufe der Zeit bekannt wurden.
- Auf der technischen Seite entstand im Laufe der Zeit die **Kollationierungstechnik.**

Instrumental waren für das neue Etymologiemodell die Enzyklopädisten, Katharina die Große und ihre Expeditionen, (Fischer, Bacmeister usw., vgl. (Winkler 1999; Gulya 2000), die das kritische Rohmaterial geliefert haben. Die geistige Synthese der erfolgte durch die Werke von Sajnovics und Gyarmathi.

János Sajnovics Sajnovics stützte sich auf die zeitgenössischen "linguistischen" Kollationierungstechniken und schrieb 1770 seinem "Demonstratio Idioma Ungarorum et Lapponum idem esse", in dem er die Sprachverwandtschaft des Ungarischen mit dem Lappischen nachzuweisen anstrebte. Er postuliert keine gemeinsamen Ursprache, sondern kollationiert lediglich ungarische und lappische Wörter. Sámuel (Gyarmathi 1799) präsentierte 30 Jahre später eine breitere, sieben Sprachen umfassende Kollationierung (genannt "Affinitas"), indem er durch Zeigen der lexikalischen Ähnlichkeit die Verwandtschaft der finnisch-ugrischen Sprachen beweisen wollte.

Das Etymologie-Modell von Sajnovics und Gyarmathi sind sehr ähnlich. Gyarmathi ging über Sajnovics hinaus indem er auch die grammatikalische Struktur für die Sprachverwandtschaft verwertete. Die beiden, wie generell die Enzyklopädisten, nehmen keinen Bezug auf eine Ursprungsspache sondern listen lediglich die verwandten Wörter, die die ebenengleiche, **horizontale Verwandtschaft** unmittelbar enthüllen sollen. Sprachverwandtschaft wird als Zugehörigkeit

zu einer gemeinsamen Menge aufgefasst. Demnach ist die Zielsetzung der Etymologien die **Darlegung der Verwandtschaft unter der Sprachen.**

Die Kollationierung implizierte natürlich eine ursprüngliche Gemeinsamkeit, die aber Sajnovics und Gyarmathi nicht näher spezifiziert wird.

Etymologie als bloße Kollationierung ohne restriktive Bezugnahme auf Sprachgeschichte und Lautwandel lebt auch weiter und trägt seltsame Früchte. Besonders verbreitet sind diversen Verwandtschaftsnachweise im Zusammenhang mit dem Ungarischen, vgl. (Rédei 1998).

4. Wissenschaftliche Etymologie und die Vergleichende Sprachwissenschaft

Die moderne wissenschaftliche Etymologie ist ein Adoptivkind der Vergleichenden Sprachwissenschaft.

Die Vergleichende Sprachwissenschaft hob die Etymologie von der Ebene der anekdotenhaften Unverbindlichkeiten auf die einer **wissenschaftlichen Teildisziplin** und schuf ein neues, wissenschaftliches Etymologie-Modell. Wissenschaftlich bedeutet systematisches, nachprüfbares Vorgehen, und kontinuierliche Integrierung der vorangehenden Forschungsergebnisse.

Auch in einem anderen Sinn besaß das neue, komparatistische Etymologiemodell einen anderen Status als das der vorangehenden: Das wissenschaftliche Modell war zuerst ein **Rahmen zu füllen,** ("Paradigma" im Sinne von Kuhn), es war ein Programm, woran die Forscher noch arbeiten müssten.

Aus der Perspektive der Komparatistik war nicht mehr die Erhellung der Urbedeutung der Wörter, oder die Demonstration der Zusammengehörigkeit der Sprachfamilien in der Etymologie das wesentliche, sondern die genaue Topologie der Sprachverwandtschaft und darunter insbesondere ihre Reflexion im Wortschatz. Die eigentliche Aufgabe lautete:

1. Die Erfassung der Änderungen, d.h, die vertikal (von älteren zur neueren) Sprachstadien wirkenden sprachspezifischen Lautwandeln und
2. Die Systematisierung der bestehenden, horizontal konzipierten Verwandtschafts-beziehungen.

Methodisch folgenschwer war die Bezugnahme auf die historischen Lautveränderungen, wonach die verglichenen Wörter nicht mehr direkt übereinstimmen, sondern vielmehr regelmäßige Entsprechungen aufweisen müssten.

Die Anwendung der **Lautgesetze** führte dazu, daß Abstammung und Verwandtschaft der Wörter nicht durch die oberflächige Ähnlichkeit, sondern durch die Systematizität der Lautentssprechungen entschieden wurde. Pointiert gesagt: Entscheidend ist bei den Lautentssprechungen nicht die materielle Übereinstimmung, sondern die Systematizität der Abweichungen.

Verglichen und rekonstruiert werden allerdings nicht bloße Lautsequenzen, sondern **Wurzel, Stämme, Wörter** d.h. Struktureinheiten der Sprache. Hierdurch sind zwei Prozesse eingeleitet worden, die die Substanz der Etymologie bis zu der Gegenwart stark mitprägen: 1. die **Linguistisierung**, und 2. **Professonalisierung**. Durch Linguistisierung wird die systematische Nutzung vom linguistischen Wissen gemeint, wie Postulierung des Sprachwandels (und darunter des Lautwandels) und deren systematischen Verfolgung und die Auswertung der Mophemstruktur. Alle linguistische Erkenntnisdomäne (wie Onomatipoetica, Lehrbeziehungen, Bedeutungsrelationen usw.) konnten und müssen für die Etymologien ausgewertet werden. Die Etymologen sind Linguisten geworden und die Etymologie zu einem Anwendungsfeld der Komparatistik. Die unverbindliche Kollationierung der Wörter wurde durch die wissenschaftliche Feuerprobe der Lautwandelregeln ersetzt, die eine linguistische Vorbildung bedarf, über welche nicht alle verfügen. Hierdurch entzog sich die Etymologie der populären Kontrolle, denn erst im Kenntnisse der Lautgesetze kann eine Etymologie postuliert bzw. zu gewürdigt werden. Lingustisierung ist ein Teil der Professionalisierung.

Die Lautgesetze werden zwischen eine Ursprache und einer Tochtersprache formuliert, also in einem Schema:

$$E: \text{Wort}_{\text{Ursprache:x}} \Longrightarrow \text{Wort}_{a}\ ,\ \text{Wort}_{b,}\ \ldots\ \text{Wort}_{i,}$$

zwischen Ursprache und Sprache_{a} , Ursprache und $\text{Sprache}_{b,}$ ... Ursprache und $\text{Sprache}_{i,}$ Die Etymologie gilt als bestätigt, wenn alle Laute in den Wortpaaren den Lautgesetzen von Ursprache zu TochterSprache gehorchen.

4.1. Lautwandel und Lautgesetze

Bopp und Grimm waren keine Etymologen, sondern vergleichende Sprachwissenschaftler. Aber ihr Programm öffnete für die Etymologie neue Perspektiven und ihre Bezugnahme auf die Sprachwandel und Sprachstruktur war für die Etymologen überzeugend und attraktiv, Das davon abgeleitete neue Modell blieb jedoch für die Etymologien praktisch leer, da zuerst die Regeln der Lautwandel und das Netz der Verwandtschaftsbeziehungen aufgestellt werden müssten. Vielfach müssten auch noch konzeptuelle Einzelheiten geklärt werden. Es dauerte bis zu der zweite Hälfte des 19. Jahrhundert bis die ersten repräsentativen wissen-schaftlichen etymologischen Wörterbücher erschienen.

Federführend waren dabei die Junggrammatiker, für die die Lautgesetze im Mittelpunkt standen: Sie wollten exakt arbeiten, weniger spekulativ aber methodisch sauber: nicht rekonstruieren, sondern lediglich vergleichen. Sie wollten die Lautgesetze strikt („ausnahmslos") anwenden und in der Verwirklichung dieses Programms steckt ihre eigentliche Leistung.

Bopp postulierte die Erste Germanische Lautverschiebung, wonach die Ver-schlusslaute vom Indogermanischen (repräsentiert durch Sanskrit) zu Germanischen (repräsentiert durch Gotisch) systematisch verschoben worden sind:

Sanskrit:	p	b	bh	th	d	dh	k	g	gh
Gotisch:	f	p	b	þ	t	d	?	k	g

wie sie in den lexikalischen Entsprechungen in Sanskrit bzw. Germanisch belegt sind:

Sankrit:	pad	bhar	asadat	dhaman-	
Germanisch:	fōtus	bairan	sitan	ga-dēþ	usw.

Dabei blieb es unerklärt, warum *Bruder* und *Vater* anders lauten, obwohl im Sanskrit beite mit -t- belegt sind (bhrātar bzw. pitā). Der Jungrammatiker Verner ermittelte, daß diese abweichende Entsprechung durch die Betonung verursacht ist und belegte mit Dutzenden von Beispielen: bhratar aber pita: -t- bleibt vor der Betonung im Germanischen erhalten.

Die beschreibungstechnische Präzision der Junggrammatiker setzte neue, hohe Standards, die nicht mehr zu Seite geschoben werden können, terminierte jedoch nicht in einem neuen Etymologie-

modell. Obwohl sie die romantischen Rekonstruktionen leidenschaftlich bekämpften, haben sie nicht das Konzept der Rekonstruktion beanstandet, sondern die mangelnde empirische Begründung.

Die Präzision in der Beschreibung der Lautgesetze ging möglicherweise auf die Kostendes Gesamtmodells: die Junggrammatiker haben nur Belege und Lautentsprechungen gesehen und Sie stellten auch die semantische Seite der Vergleiche zurück.

4.2. Die historische Rekonstruktion

Die Ähnlichkeit der scholastischen Abstammungsmodell mit den modernen komparatistischen Etymologiemodell macht auch den grundlegenden Unterschied deutlich: Die Scholastiker gingen von der (bereits gegebenen = hebräischen) Ursprache aus und wollten beweisen, daß die Nationalsprache davon abstammt:

Ursprache ==> Nationalsprache.

Die Komparatisten dahingegen müssten zuerst die Ursprache erschließen, rekonstruieren bevor sie zu der Ableitung fortschreiten könnten:

Ursprache <== Nationalsprache.

Rekonstruktion war eine Hypothese und warf zwei Probleme auf:

1. Man musste die Rekonstruktion **auf der theoretischen Ebene** näher bestimmen: Konzeptuell, methodisch und statusmäßig. Die rekonstruierte indoeuropäische Ursprache war nicht identisch mit Sanskrit oder Urgermanisch mit Gotisch. Die Rekonstruktion war eine Hypothese, sie hatte also einen anderen Status als die Sprache der (schriftlichen) Quellen, sie war aber methodisch unerlässlich.
2. Man musste die Rekonstruktionsebenen **konkret entwerfen,** die gleichzeitig auch die Verwandtschaftsbeziehungen zwischen den Sprachen und Sprachfamilien wieder- spiegelte und die Lautrekonstruktionen physikalisch aufstellen.

Die Einsicht, daß in der Rekonstruktion, die methodischen Überlegungen wichtiger sind als die Produktion von *-Formen und *-Texten, setzte sich im Laufe der Zeit durch. Das Konzept der etymologischen Rekonstruktion wird weiter präzisiert:

1. Der Hauptunterschied zwischen einer Belegsprache und einer rekonstruierten Sprache (Ursprache, kantakieli) besteht darin,

daß eine belegte Sprache sowohl als Parole, als auch als Langue (im Sinne von Ferdinand de Saussure) existiert, während eine Rekonstruktionssprache lediglich eine Langue-Ebene besitzt. Aus dieser Gegebenheit ergeben sich weitere Folgen für die Etymologien, die bereits verschiedentlich beobachtet worden sind:

2. Die Rekonstruktionen sind nicht nur hypothetisch sondern auch unterbestimmt, denn vielfach können wir die Formen mangels zeitgenössischen Belege nur unvollkommen (Vordervokal, Hintervokal, Spirant u.ä.) bestimmen, und die Unterbestimmtheit der Rekonstruktion muss berücksichtigt werden (vgl. Anttila 1972: 343).
3. Rekonstruiert wird das System. Dieses System ist nicht nur die Summe aller Folgen der Lautgesetze, sondern gleichzeitig ist sie auch eine menschliche Sprache, d.h. die Sprachuniversalien müssen beachtet werden (vgl. Prosdocimi 1978: 90).
4. Die Rekonstruktion muß systematisch durchgeführt werden: Verglichen dürfen nur koexistente Formen, und nicht solche, die nicht nur hypothetisch sind sondern auch verschiedene Zeitstufen angehören: auch die Zwischenstufen müssen rekonstruiert werden, wenn man Rekonstruktion ernst nimmt (Honti 1976: 136).

4.3. Die innere Rekonstruktion und die Etymologie

Wörter verändern sich nicht nur von Sprache zu Sprache, sondern auch innerhalb. Innere Rekonstruktion basiert auf der Diskrepanz zwischen morphologischen und phonologischen Gesetzmäßigkeiten in einer Sprache, mit deren Hilfe die Geschichte der Sprache quasi zurückverlängert werden kann, daher ist die Methode insbesondere für Sprachen ohne ältere schriftliche Überlieferungen wichtig.Im Zusammenhang mit dem Sprachvergleich werden hier die älteren (ältesten) Formen der Wörter, die für die Eymologien benötigt werden, bestimmt. Wörter lauteten früher nachweislich anders als heute.

Für die Etymologien (Vergleich und Rekonstruktion) sind also nicht die heutige Formen interessant, sondern die jeweils ältest erreichbaren Formen. Andererseit werden Wörter in unterschiedlichen Formen realisiert: in den Flexionsparadigmen können regelmäßig Laute wegfallen oder scheinbar unmotiviert

hinzutreten. Der Wechsel in dem finnischen Nominalparadigma Nom *vesi* – Gen} veden „Wasser" ist unverständlich wenn man nicht weiß, daß hier *-si}* auf frühere *-ti}* zurückgeht. Morphologische Eigenarten können für die Etymologien interessante und wichtige Schlüssel liefern: Syrj. *mus* – *musk-* „Leber" weist in den obliquen Formen eine morphologische Unregelmässigkeit auf, die sich allerdings voll erklärlich ist wenn man Permisch *musk-* dem ostseefinnischen Stamm *maksa* gegenübersteht.

4.4. Die Emanzipation der schriftlosen Sprachen und Dialekten und phonetische Transkription

Für die Pioniergeneration bedeutete die Etymologie die Beschäftigung mit der Vergangenheit: mit alten Schriften, aus welchen hervorging, wie die ursprüngliche (wahre) Bedeutung war bzw. die korrekte Form lautete. Die Gleichstellung der Volkssprachen mit den traditionsreichen (sakralen) Sprachen – einen einfachen Kurdendialekt mit Sanskrit oder Lulelappisch und Schwedisch auf der Ebene der Lautwandel- war nicht nur populistisch, sondern eröffnete neue Perspektiven und vermehrte vor allem das komparatistisch verwertbare Belegmaterial. Nicht nur die relativ spät verschrifteten Nationalsprachen (wie Tschechisch, Ungarisch) wurden in die etymologische Betrachtung einbezogen, sondern auch die schriftlosen Sprachen die erst jetzt aufgezeichnet worden sind (wie die amerikanischen Indianersprachen), also Sprachen die u.U. nur als Dialekte existierten ohne dialektübergreifende schriftliche Normierung.

Die wissenschaftliche Zuwendung zu den Dialekten führte als Side Effect zur der Problematik der authentischen Lautwiedergabe. Die nähere Betrachtung der nationalen Schriftsysteme machte bereits deutlich, daß für die wissenschaftliche Etymologie maßgebend die Lautwerte sind, und nicht die Buchstaben. Für die etymologische Auswertung der neuen, bisher schriftlosen Sprachen (Dialekte) könnte die Benutzung einer einheitlichen, normierten Schrift Vorteile bringen. Allerdings erwiesen sich die Normierungsversuche als wenig erfolgreich. Die Lautschrift findet generell Verbreitung und Eingang in die Etymologieforschung.

Die etymologische Forschung war in der Indogermanistik schriftsprachlich zentriert. Dabei entstand eine Ungleichheit zwischen gut und weniger gut dokumentierten Sprachen. Es werden

stets die gut dokumentierten Sprachen herangezogen und bekommen ein unverdientes Übergewicht. Die phonetische Betrachtung ließ die schriftlosen, traditionsarmen Sprachen auf dieselbe Ebene mit der Schriftsprachen.

Die Verwertung der dialektalen Belegungsdichte liefert interessante etymologische Zusatzinformation, die in das komparatistische Etymologiemodell nur schwer hineinpassen: Die Belege aus den Dialekten wiesen die Verbreitung des Wortes in der Diachronie nach und so förderte die Zuwendung zur Dialektologie einen neuen verallgemeinerten Etymologiebegriff: und die Aufwertung der mündliche .Belege.

4.5. Die Überprüfung der Wortbedeutungen

In den Etymologien müssen auch Bedeutungsrelationen gleichbereichtigt mit dem Lautvergleich beachtet und verwertet werden.

Obwohl die These, daß die etymologisch verwandten Wörter auch bedeutungsmäßig ähnlich sein müssen, niemals umstritten war, konzentrierte sich die etymologische Forschung in der junggrammatisch geprägten Epoche auf die Lautentsprechungen. Man sah keine Möglichkeit die Bedeutungsrelationen zu erfassen. inzukamdiepraktische Erschwernis, daß die Bedeutungsexplikationen typischerweise in verschiedenen Sprachen (Lateinisch, Englisch, Russisch, usw.) angegeben waren. Bereits die Übersetzung in eine einzige, normierende Bedeutungsrepräsentationssprache wurde kritisch betrachtet.

Die Forderung wurde aufgestellt, daß die für die Sprachverwandtschaft relevanten Etymologien einigen wenigen Inhaltsklassen angehören müssen: Zahlen, Körperteile, Verwandtschaftsbezeichnungen, elementare Naturphänomene u.ä... Eine systematische Einbeziehung der Bedeutungsrelationen in die Etymologien ist immer noch eine Herausforderung (vgl. jedoch Mikola 1976).

4.6. Analogie, Onomatopoetica, idiomatische Wendungen, Entlehnungen-Etymologie als Quiz

Die Überwindung der junggrammatischen Fixierung auf die historische Lautlehre brachte zweifelsohne eine thematische Bereicherung

der Etymologieforschung und lieferte in Einzelfällen feinere, präzisere Etymologien.

Das additive Vorgehen bei der Einführung neuer in sich interessanter und legitimer etymologisch relevante Gesichtpunkte ergab jedoch kein neues Modell. Das komparatistische Modell blieb informell: Die Prinzipien der Etymologie wurden sehr einsichtvoll verbal geschildert, aber es gab keinen Versuch, die geschilderten Prinzipien in einem Modell kohärent und verbindlich festzulegen (vgl. Joki 1976; Kiss 1976; Malkiel 1993; Hajdú 1978; Kulonen 1996).

Die vielen Erweiterungen machten es schwer und nur für Professionellen möglich etymologische Lösungen zu entwickeln. Etymologien wirkten wie Rätsel oder Quiz.

5. Professionalisierung der Etymologieforschung

Die Professionalisierung der Etymologieforschung begann bereits mit der wissenschaftlichen Wende. Durch die Scientifizierung wurde bereits den Etymologie- Amatören die Lizenz entzogen.

Ein äußeres Zeichen der Professionalisierung war die Einrichtung umfangreicher Etymologiearchiven für die Organisation und Pflege der Bestände, deren kontinuierliche Finanzierung gesichert werden muß.

5.1. Lexikographie und Etymologie

Etymologie ist von ihrer Herkunft her auf die Einzelwörter gerichtet und ihre natürliche Erscheinungsform ist das Wörterbuch. Lexikographie verhalf die Etymologie zu höherer Professionalisierung, durch die Vermittlung der allgemeinen lexikographischen Techniken, und darunter insbesondere auch die Techniken der elektronischen Datenverarbeitung:

1. Strengere, systematische, einheitliche Formatierung der Einträge Font management,
2. Quellen und kritisches Apparat.

Modellierung in der Etymologie nahm erst im lexikographischen Rahmen faßbare Formen an. Die modernen etymologischen Wörterbücher der finnischugrischen, insbesondere (Rédei 1988; SKES 1978) und (Itkonen, Kulonen 1995) bzw. uralischen Sprachen stehen einander konzeptuell sehr nahe: Sie operieren alle mit den Lautgesetzen und Rekonstruktion. Der Unterschied besteht lediglich

darin, daß (Rédei 1988) die rekonstruierten Formen auch ermittelt (und als Sortierschlüssel einsetzt), während die finnischen Werke die rekonstruierten Formen nicht explizit präsentieren.

5.2. Das etymologische Wörterbuch als Datenbasis

Ebenso wie die lexikographische Professionalisierung kein inhaltlich neues Etymologiemodell gebracht hat, bildet eine etymologische Datenbasis zuerst das typographisch geprägte Etymologiemodell nach. Die Überlegenheit des DB-Modells besteht in der Präzision, Explizitheit der Beschreibung und in den eröffneten Perspektiven:

1. Das Modell ist nicht additiv, das die einzelnen später zugefügten Komponenten (Entlehnungen, Onomatopoetica, Analogie usw.) beziehungslos nebeneinander stehen läßt, sondern die Möglichkeit bietet, in einem abstrakten, mehrdimensionalen Datenraum alle relevanten Abhängigkeiten der etymologischen Belege zu bestimmen und ihre virtuellen Werte zu ermitteln.
2. Die Überlegenheit des DB-Modells besteht in der Überprüfbarkeit etymologischen Annahmen. Die Wissenschaftlichkeit der komparatischen Modell war durch die persönliche Integrität (Sprachkenntnisse und Erfahrung) des Wissenschaftlers gebürgt. Das DB-Modell erzwingt eine konsequente Modellierung, bietet die Möglichkeit einer Selbsverifizierung und steht dadurch auf einer höheren Stufe der Objektivität.

Literatur

Anttila, Raimo 1972. Introduction to Historical and Comparative Linguistics. New York: The MacMillan.

Arens, Hans 1969. Sprachwissenschaft – Der gang ihrer Entwicklung von der Antike bis zur Gegenwart. Freiburg/Munchen: Athenäum Fischer Taschebuch Verlag.

Bátori, István; Csúcs, Sándor 2000. Uralische Etymologische Databasis – Progress Report. Tartu: Universität Tartu. http://www.uni-koblenz.de/uedb/uedb_aktuell/fgr9.ps.

Bátori, István. demn. Az etimológiai szótár adatbázisként. Budapest: Linguistics Institute, Hungarian Academy of Sciences. http://www.uni-koblenz.de/~batori/uedb/bum2.ps.

Gyarmathi, Sámuel 1799. Affinitas lingvae hvngaricae cvm lingvis fennicae originis grammatice demonstrata ... Joann Christian

Dieterich, Göttingen. Nachdruck in ungarischer Übersetzung, Reihe = Regulyana 3, Tinta Könyvkiadó, 1999.

Hajdú, Péter 1978. Rekonstruktion in der Uralistik. – Proceedings of the Twelfth International Congress of Linguists. Hrsg. Wolfgang U Dressler, Wolfgang Meid. Innsbruck. 99–101.

Honti, László 1976. Az alapnyelvi rekonstrukciók kérdéseiről. – Az etimológia elmélete és módszere, Volume 89 von Nyelvtudományi Értekezések. Budapest, Akadémiai Kiadó. Papers of the International Conference on Etymology, 22–24. August 1974. Ed. by Lorand Benkő, Éva K.Sal. 131–137.

Itkonen, Erkki; Kulonen, Ulla-Maija (toim.) 1992, 1995. Suomen sanojen alkuperä. – Etymologinen sanakirja 1–2. Suomalaisen Kirjallisuuden Seura / Kotimaisten Kielten Tutkimuskeskus.

Joki, Aulis J. 1976. Prinzipien und Praxis der Redaktion des Finnischen Etymologischen Wörterbuches. – Az etimológia elmélete és módszere, Volume 89 von Nyelvtudományi Értekezések. Budapest. Hungarian Academy of Sciences. Papers of the International Conference on Etymology, 22–24. August 1974. Ed. by Lorand Benkő, Éva K.Sal. 156–160.

Kiss, Lajos 1976. As etimológia kutatások újabb fejlődése külföldőn. – Az etimológia elmélete és módszere, Volume 89 von Nyelvtudományi Értekezések. Budapest, 1976. Hungarian Academy of Sciences. Papers of the International Conference on Etymology, 22–24. August 1974. Ed. by Lorand Benkő, Éva K. Sal. 27–50.

Krahe, Hans 1962. Indogermanische Sprachwissenschaft. Sammlung Göschen Band 59. Berlin: Walter de Gruyter.

Kulonen, Ulla-Maija 1996. Sanojen alkuperä ja sen selittäminen. Etymologiasta leksikografiaa. Helsinki: Suomalaisen Kirjallisuuden Seura.

Malkiel, Yakov 1993. Etymology. Cambridge: Cambridge University Press.

Mikola, Tibor 1976. Hangtan és jelentéstan az etimológiában. – Az etimológia elmélete és módszere, Volume 89 von Nyelvtudományi Értekezések. Budapest. Hungarian Academy of Sciences. Papers of the International Conference on Etymology, 22–24. August 1974. 209–211.

Pisani, Vittore 1975. Die Etymologie. Internationale Bibliothek für Allgemeine Linguistik. München: Wilhelm Fink Verlag.

Prosdocimi, Aldo L. 1978. Diachrony and reconstruction. – Proceedings of the Twelfth International Congress of Linguists, Innsbruck. Ed. by Wolfgang U. Dressler, Wolfgang Meid. 84–101.

Rédei, Károly (Hrsg.) 1988, 1991. Uralisches Etymologisches Wörterbuch I–III. Unter Mitarbeit von Marianne Bakró-Nagy, Sándor

Csúcs, István Erdélyi†, László Honti, Éva Korenczy†, Éva K. Sal und Edit Vértes. Budapest und Wiesbaden: Akadémiai Kiadó und Otto Harrassowitz.

Rédei, Károly (Hrsg.) 1998. Östörténetünk kérdései – a dilettáns nyelvhasonlításról. Volume 10 von Budapesti Finnugor Füzetek. ELTE, Finnugor Tanszék, Budapest.

Sajnovics, János 1770. Demonstratio. Idioma ungarorum et lapponum idem essse. Collegia Academici Societatis Jesu, Koppenhagen/Nagyszombat, Nachdruck in ungarischer Übersetzung, Reihe = Regulyana 2, Budapest, ELTE, 1994.

Toivonen, Y. H. (toim). 1978. Suomen kielen etymologinen sanakirja I–VI. Helsinki: Suomalais-Ugrilainen Seura.

To a possible GIS based multimodal question answering system

M. V. Boldasov, S. A. Sharoff, E. G. Sokolova, V. A. Zhigalov
Russian Research Institute for Artificial Intelligence, Moscow

The paper presents the design of a question-answering system aimed at providing information for a tourist traveling in a city. The analysis component is based on the semantic-oriented analysis of A. S. Nariniani in the framework of InBASE system performed in Russian Research Institute for Artificial Intelligence (RRIAI). The generation component is based on the multilingual English–German–Russian grammars developed in the framework of John Bateman's KPML, which, in its turn, is based on Halliday's Systemic Functional Linguistics. The aim of the paper is to discuss some correlations between extralinguistic knowledge of spatial relations and realization of the extralinguistic knowledge in terms of language-motivated structures, integrating lexical-semantic and grammatical means as well as text planning.

Introduction

What we have in mind is an intelligent interactive software system that helps people to access information and orient in spatial configuration of objects in some area. The objects can be buildings, streets and parks in a city, buildings in a palace area, farmstead or something else. In this paper we present the design of a question-answering system aimed at providing information for a tourist traveling in a city.

Multimodality has become an intrinsic feature of the information presentation in the Internet. All kinds of media are included: text, diagrams, images, video and audio clips, different ways of representation of data sets. A question about the correlation between different forms of information arises. The expressibility of information in different forms depends on the nature of the information. For example, weather parameters (temperature, wind direction and so on) are firstly expressed in numeric tables, and after that they appear in a verbal form as weather forecasts for the public. By contrast, formal specifications for user manuals are primarily expressed in a verbal form.

Recently some systems have appeared which present information expressed in a verbal form as diagrams (Mittal *et al* 1995; Bateman *et al* 2000). Information presented as a geographical map in general cannot be presented in a verbal form (Зацман *et al* 2001), but it can be presented in the particular case when there is a concrete communicative goal – such as, for example, explaining the path from one point on a map to another. One of the possible areas to use the two forms of information representation is a tourist question-answering system.

The core component of the discussed design is geographical information system (GIS). Roughly GIS is a digitized map with layered objects bases attached. Geoinformatic mapping has extensively risen because it can integrate a great amount of cartographic and other spatial-related information on various features and phenomena including economical, cultural or historical areas (Зацман *et al* 2001). In the tourist information system objects are houses (one of the house features is the house number), streets, city traffic including bus stops and metro stations, museums and so on. GIS KB can also see thought of as a set of multiscale 3-dimensional objects – "virtual geoinformational objects" – a combination of space attributes and informative aspects of objects and events. "Virtual geoinformational objects" of GIS are usually classified or typed. This allows the information presented in GIS to be split into layers for the convenience of its processing and user representation. So GIS is a digitized map with layered information attached.

An example of a possible information dialog (user's queries are marked "*U*>", system's reactions as "*S*>".):

U>: Are there Greek restaurants nearby?

The GIS retrieves the requested object and delivers the information on NL:

S>: There is just one Greek restaurant in the vicinity. Its name is "Griechisches Tavern".

The aim of the paper is to discuss some correlations between extralinguistic knowledge of spatial relations and realization of the extralinguistic knowledge in terms of language-motivated structures, integrating lexical semantic and grammatical means as well as text planning. We also consider some systems modeling elements of the designed prototype. A functionally similar system delivering information on Geography is START, developed by Boris Katz at MIT Arti-

ficial Intelligence Laboratory in the USA (http://www.ai.mit.edu/projects/infolab/ailab.html). This technique employs natural language sentences and phrases – annotations – as descriptions of content that are associated with information segments at various granularities. An information segment is retrieved when its annotation matches an input question. This method allows START to handle all kinds of media, including text, diagrams, images, video and audio clips, different ways of representation of data sets and others. The system also calculates paths to get to some geographical point from another point, for example:

U>: "How can I get from Washington to New York?"

The system builds a path to the goal – a table listing all changes of route for a car from the initial to the target point.

1: Start out going East on E ST NW towards ELLIPSE RD NW.	0.2 miles (0.3 km)
2: E ST NW becomes PENNSYLVANIA AVE NW.	0.1 miles (0.1 km)
3: Turn RIGHT onto 14TH ST NW.	0.2 miles (0.4 km)
.....	
39: Turn RIGHT onto CANYON OF HEROES/BROADWAY.	0.0 miles (0.0 km)

Here movement of a point through a scheme is modeled. Pedestrian, bus, metro and other city transport can similarly be involved though including human perception of space and surrounding objects. As an answer the system generates NL text associated with multimodal information presentation elements.

The first component of the tourist system is analysis. Conceptually we base it on the InBASE system performed in RRIAI. To achieve the NL access to data bases (DB) the system uses semantically-oriented approach of A.S.Nariniani and some other modern technologies (Жигалов *et al* 2001). The system is exposed in the Internet and goes through its pre-commercial period; for example, you can see English queries NL interface to a DB (the system works now with Data Base of a small company employees) where you can see internal OQL and final SQL query representations (http://artint.ru/nl/kadry-eng.asp). InBASE technology has twofold benefits. Firstly it does not use parsing, so it can understand free lan-

guage that is specific to the domain. NL query processing is organized as bottom-up gathering of semantic elements into register oriented semantic representation. The second benefit is multilinguality, i.e. the same "semantically-pragmatic" grammar is used for different languages (currently – English and Russian) and different domains (currently: staff DB, DBs of Internet shops for microcomputers, mobile phones, choosing system for auto and some others). The core notion of the interface to a DB is the domain model based on diagram of classes. Using the domain model the NL query is firstly interpreted into representation on Q-language – a subset of OQL (Object Query Language) and then to SQL representation. Expansion of the approach for the tourist system will need some exploration since the register of question answering system demonstrates more diversity of query types than the register of NL access to DBs.

The tourist system is not a dialog system, but information question answering system. In artificial intelligence "a dialog system" is usually equal to "advisory or argumentative type of dialog system", for example (Koit, Õim: 2000). We do not consider this type of systems here. But the "information" dialog is indispensable for the system. It is a type of the START system reaction to user query when it cannot react properly. A more elaborated dialog of the information type is presented in the earliest artificial intelligence system – Winograd's robot (Winograd 1972) where the system informs the user that it has not understand some word, lacks relevant information or need some additional details, to for example, define a concept.

The outlet component of the discussed design is text generation. Text generation is a technology aimed at the production of high-quality natural language text from computer-internal representations of information. The applications of this technology typically involve production of descriptions on demand in areas where texts are changed, for example, weather reports or user manuals. A subclass of text generation is multilingual generation (MLG), in which the same content is expressed in several languages in parallel. The approach MLG contrasts with the more conventional approach of Machine Translation in that the starting point is not a source text in one language, but a non-linguistic specification of the content to be expressed.

A prototype of the multilingual generation system which serves as the basis for the proposed research has been developed within

AGILE (Kruijff *et al* 2000) – a project for generation of user manuals for CAD-CAM systems in three languages of Eastern Europe (Bulgarian, Czech and Russian). The system allows technical writers to specify the content of the instructions for carrying out individual tasks in the CAD-CAM domain, in which a user can perform various operations, like drawing lines, specifying their properties, opening and saving files, etc. This content is then automatically expressed in each of the three languages in parallel. The formal tools for representing the content are based on the Domain Model (DM), a domain-specific extension of the Upper Model (Bateman *et al* 1995) – a multilingual hierarchy of general concepts that appear in a language. For example, UM includes such concepts as DISPOSITIVE-MATERIAL-PROCESS or EXISTENCE, while DM includes such concepts as OPEN-FILE or APPEAR. In addition, DM includes concepts that specify the configuration of an instruction – that it includes GOAL, a set of METHODs for achieving it, a sequence of STEPs implementing a METHOD, and the set of SIDE-EFFECTs that may appear as the result of executing a step. For example, if the task of the user is to draw an arc, then the CAD system provides a method for doing this. The first step consists in invoking a special tool for doing this, a side-effect is the presence of the tool on the screen when the arc is being drawn. The example of generated text is:

To draw an arc
First start the ARC command using one of these methods:
Windows: From the Arc flyout on the Draw toolbar, choose 3 Points.
DOS and UNIX: From the Draw menu choose Arc. Then choose 3 Points.
Now specify three points of the arc.
1. Specify the start point (of the arc). First enter endp. Then select a line. The arc snaps to the endpoint of the line.
2. Specify the second point of the arc. First enter poi. Then select a point. The arc snaps to the point.
3. Specify the endpoint of the arc.

2. User query processing

User queries to a city information system reflect the following situation: the user has a map or tourist scheme of the city, he/she can point out his/her location and the system can show the points user would like to get to or to have information about. The more natural type of interaction between the user and the system is a dialog. The dialog

initiator is the user. He/she formulates a query to the system, the system delivers information or formulates its own query if the user's query is ambiguous or has not sufficient information for answer:

U> How can I get to Red Square?
S> You'll walk, use a bus or metro, or go by car?

GIS is usually multilevel. Oriented to tourist information GIS can have three levels: streets and buildings (walk), urban transport (bus or metro), car-related marking – one-way traffic, pedestrian street, road junctions and so on.

In the register we can consider some types of queries:

1. point the path to;
2. point the location of some object;
3. find all the specified objects;
4. describe the features or attributes of an object;
5. etc.

NL query is analyzed and referred to one of the types. To answer a query, the relevant information need to be extracted from the KB and expressed it on NL. We need to associate significant fragments picked out from the NL query with DM concepts. The type and semantics of the NL fragments in the KB are crucial – for example, *Red Square* and *Sadovoje Ring* belong to the same object type – "ground User can move through", *not far from me* and *some near place* have the same semantics – "in a short distance", *eat* and *look* present the same semantic type "object specific User action". Below we present some queries with some NL fragments in Russian presented in brackets by their DM concepts or types:

- **"point the way to":**
  Как ... добратъся до <object>?
  (How ...get to <object>?)
- **"point location of some object":**
  Где находится < object >? Где <object >?
  (Where is < object >?)
- **"find all the specified objects":**
  Какие <object > естъ в <specified map area>?
  (What <object > are in <specified map area>?)
  Естъ ли в < specified map area> <object >?
  (Are there <object > in < specified map area>?)

Где я могу < object specific User action >?
В каком <object > я могу < object specific User action >?
Где <specified map area > можно <object specific User action >?
(Where can I < object specific User action >?)

- **"describe the features or attributes of an object":**
  Какие <attribute> ... ? <object >?
  (Which <attribute> are in <object >?)

Below we present an example of a possible query analysis to give an idea about the method of analysis in InBASE system. Let it be the following query:

I'm at Petrovka, 38. Where is the nearest Chinese restaurant?

Here the goal of the user is to get Address of Restaurant (an Object), with minimal distance from Current Place (a context Object) at address 'Petrovka 38'; the Restaurant Object must have 'Chinese' feature. In the picture we present the scheme of analysis for the query:

	I'm at	**Petrovka**	**38**	**where is**	**Nearest**	**Chinese**	**restaurant**
Class	Obj	Val	Val	Attr	Min	Val	Obj
Orientation	(Cur Place)	(Street. Name)	(Building. Number)	(Place. Address)	(Distance)	(Restaurant. Type)	(Restaurant)
1		Val(Address)		Restaurant. Address			
2	Pred (CurPlace.Address = 'Petrovka, 38')			Get (Restaurant. Address)	Min (distance (Cur Place, restaurant))	Pred (Restaurant.Type = 'chinese')	

The stages of analysis are:

- The values Street.Name and Building.Numner are unified into Address-hypervalue;
- Values turn into the predicates (Pred) and into giving attributes (Get);
- Object oriented to the current place (I'm at) disambiguates the following values (Petrovka 38);
- Abstract objects (Place) turn into more concrete (Restaurant).

The result SQL-like query representation is:

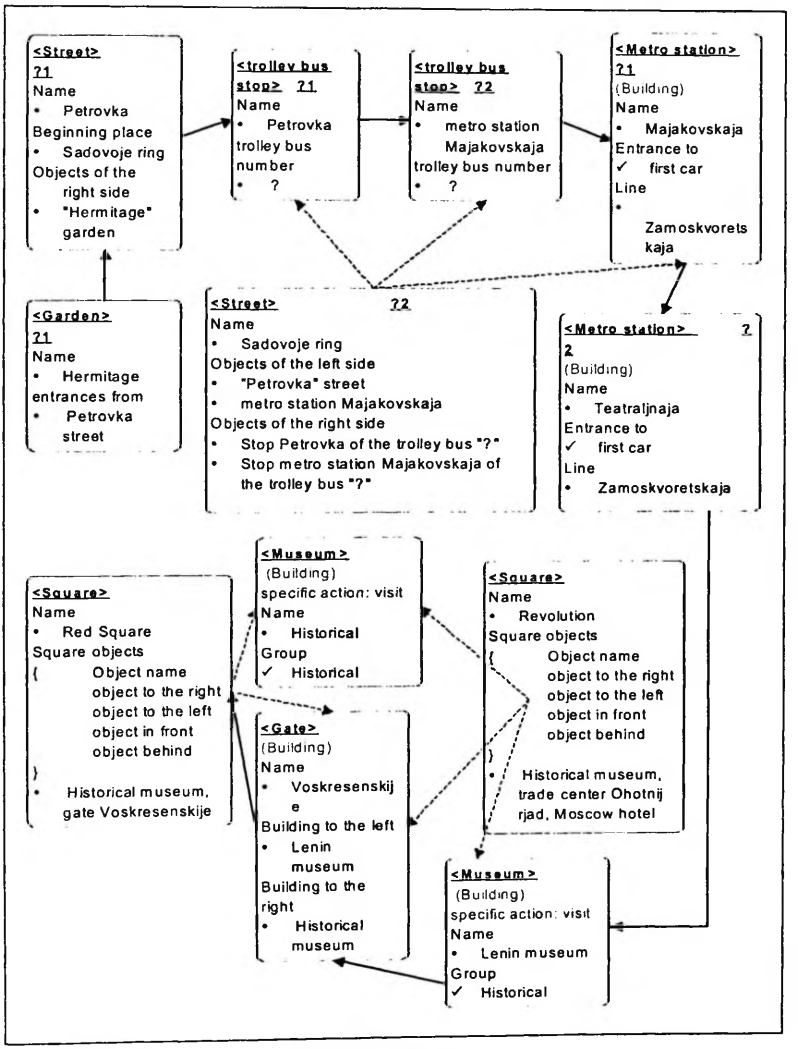


Figure 1. Possible structure of the GIS KB

```
Select Restaurant.Address, Restaurant.Type
From Restaurant, CurPlace
Where Restaurant.Type='Chinese'
      And CurPlace.Address.StreetName='Petrovka'
      And CurPlace.Address.BuildingNumber=38
      And distance(CurPlace, Restaurant)=min
```

3. From GIS to KB

GIS consists of the digitized map on the one hand and objects with attributes of their geographic position on the other hand and it can seek or process the information on their connection to the digitized map. The information processing is in most cases the task of looking up a space-restricted set of objects or of crossing some sets of objects. Moreover, GIS can represent the result of information retrieving on the map in visual form. In most cases such representation is very convenient. However, the combined visual-verbal information representation is not always what the user wants. Not everybody considers graphical information adequate. Furthermore, there are the tasks where textual documents generation is obligatory. The task of displaying nonverbal GIS information in verbal form appears.

We investigate GIS as the source of information, or KB, for the external written text generation system. Relying on the general GIS information representation concept (object concept), we suggest the conception of KB construction as objects of GIS information layers (distinct buildings, streets, squares etc.). Every particular object has its inner structure, the set of attributes. For example, street has its beginning and ending points, the ordered houses along its right and left sides and other objects. Object has also a type according to its inner structure (shown in broken brackets in Figure 4). Among the text examples we have investigated, we picked out the following types of objects: street, square, stop, station, building, gate, museum (i.e, the building; museum inherits the attributes of a building; inheritance is shown in round brackets in Figure 4) etc.

The KB fragment presented in Figure 1 describes the city objects from the perspective of their human-tourist stereoimpression. It is relevant, for example, for the following text:

U> I am in Hermitage garden, which is on Petrovka street. How can I get to Red Square? (Where is Red Square)
S> Are you on foot, driving car or you will use traction?
U> I will use traction.
S> To get from Hermitage garden to Red Square you come out to Sadovoje Ring, cross the street, take the bus ? and get to Majakovskaya metro station. Then take metro train (the first car) to Teatraljnaja station (two stops without changing). Get off. You are on Revolution Square behind the Moskva Hotel. Traverse Revolution

Square in the direction of Lenin museum. Pass it moving right and go through the arc of Voskresenskije Gate. You are on Red Square.

Objects in the KB have attributes and can be connected to a digitized map. They include proper attributes (properties) and pointers to specific predicates designating common user actions with the objects, for example, *to visit* (museum), *to eat in* (restaurant, café) etc. Especially important links between objects are set as arrows in Figure 1. We distinguish the following types of relations between objects: owner (shown with dotted arrow), attribute-inclusion (presented as an object attribute inside the object), and spatial relation (shown with solid arrow). The spatial relation is set by particular GIS module that processes GIS information to find out the best path from a start to the target point. The relation reflects the order in which the key-objects appear in the path.

In the case considered above the generation module takes the ordered object references – the path presented as a "line" in KB and build the text plan using appropriate object attributes (strategic planning). Figure 1 does not give all possible properties of the text plan objects. We have to deal not only with direct object features, but also with the so called oblique features. Here we mean that an object can be reported through the description of surrounding buildings, for example, "*near the red house on your right* "; or "*behind the skyscraper is a small two-storeyed house*". GIS can supply us with the information.

It is very important to know exactly what the information user wants to get from the system. He can forget to reveal part of the information, holding it as evident. For example, if is very important to know what facilities the user has for moving through the city: the same point will be differently available in a car, by traction or on foot. This determines what object layers should be used by GIS for the path construction. In this situation the dialog is necessary. Depending on the user decision, the path will be chosen differently and its description will be in different styles. Therefore we also need to store extra information about the request type (or the user type) in our KB. The text planner will take this information afterwards.

In GIS the information is stored in databases. Unlike in the user-oriented KB, shown in Figure 1, we also consider the original class model for representing city objects. In Figure 2 we pinpoint diagram of classes in the InBASE system style, the inheritance model. It illus-

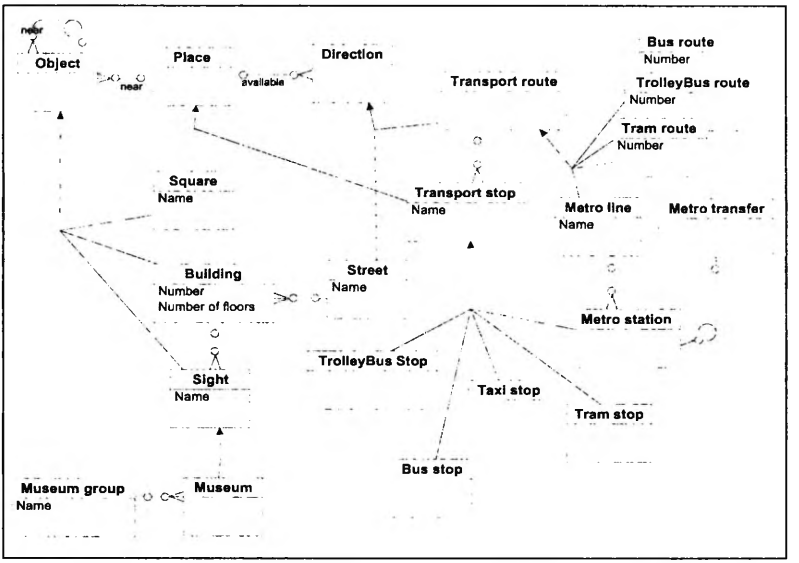


Figure 2. Possible GIS class scheme

trates object-oriented approach to the GIS KB creation. There are 3 base classes: Place, Object and Direction. Other classes are inherited from these. For instance, Street and Bus-Root are directions, Building is an Object, Square and Bus-Stop are places.

Relations between classes (and between objects in the KB) form dependences between classes and objects in the GIS domain. So, there may be some objects near a place, buildings belong to streets, transport route contains stops and so on.

This aspect does not contain such things as "on the right side", or "objects in front", because in the common OO-approach it is difficult to create scenarios and other complicate dynamic structures presenting human perception.

4. Text structure representation

Two types of generation systems can be marked out.

1) The systems which have domain specific information representation as initial source for a text; this type is based on the information sources used in practice, such as for example, the cited numeric tables for weather forecasts.

2) The systems which have communication-specific information representation as a source, for example, in AGILE system the text plan for instruction is created by a technical writer using interface facilities.

The generation process having domain-specific source typically involves three processes or stages: a) selecting the content for a future text and arranging it in a text structure representation following the communicative goal, for example, INFORM, (strategic planning); in the example in Section 3 the content of future text is selected as a "line" of objects presenting the path; b) arranging the content in a linear order in terms of some semantic representations of sentence sized spans (this stage is often called "text planning" or "tactical text planning"); c) expressing the content as grammatically and stylistically correct sentences in a target language (tactical generation). Systems belonging to the second type usually have only two last stages having text structure representation as the source representation. The tourist system we consider belongs to the former type, i.e. has three stages. Below we roughly consider the text structure representation for the example from Section 3 that could be the result of the strategic planning process.

For the text structure representation we use Rhetorical Structure Theory (RST) (Mann 1987) since it is the most widely used and minimally depends on text type. RST is a descriptive theory of text structure. At the heart of the theory is the notion of rhetorical relation that holds between adjacent segments (spans) of text. Usually it is a binary relation between the spans identified as the nucleus and the satellite. So texts are analyzed in terms of immediate constituents linked by rhetorical relations. A tree structure is used to show how each text is recursively subdivided into constituent spans down to the granularity of simple propositions – the leaves of the tree. RST was first used for generation in 1988 by E. Hovy to build a "text structurer". In his model definitions of rhetorical relations were interpreted as plan-operators in a goal oriented planning paradigm by applying constraints on the nucleus, satellite and their relation. Hovy's text structurer was used for explaining generation in an advanced system of advisory dialogs build by J. Moore and S. Paris in 1989. RST is used for text structure representation in many applications, see discussion of some particularities in (Korelsky, Kittredge 1996).

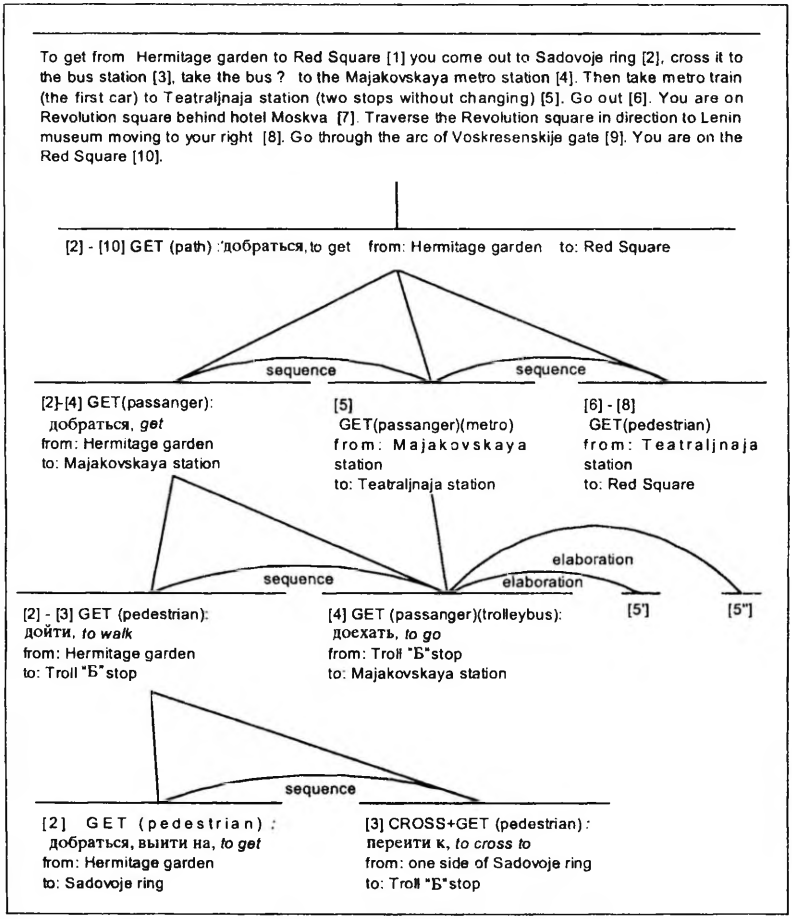


Figure 3. RST-like structure.

In Bateman *et al* (2000) an "intermediate" RST representation was proposed allowing KB concepts to be its nodes. The representation is named "RST-like structures". In this case the planning process is simplified to a mapping between elements of KB and relational structures. The process of text planning follows some domain-specific strategies, for example, if city transport is needed, metro is the first means used and then other kinds of transport are used to get to (or from) the metro station.

In Figure 3 RST is used as formalism to present domain motivated spans and relations, which become textual internal relations

```
(EXAMPLE :NAME Text4-N2
 :GLOSS ( :ENGLISH "In center is-located road.")
 :GENERATEDFORM "В середине находится дорога."
 :LOGICALFORM
 (S / EXISTENCE :LEX NAKHODITJSJA
    :CIRCUMSTANTIAL-THEME-Q CONTEXT
    :DOMAIN (L / DM::ROAD)
    :SPATIAL-LOCATING (K / DM::PICTURE-CENTER)))
```

Figure 4.

between spans of text. In particular, SEQUENCE relation reflects the situation of temporal sequence of events. Granularity of the representation is bounded by the granularity of objects in the KB, not by lexical terms. Thus any branch on any level of the representation can be cut off, predicates can be lexicalized as pointed out and text can be generated.

5. Tactical generation

The task of tactical generation is to output grammatically and stylistically correct sentences from text plans produced in strategic generation. However, this does not mean that the strategic planner is completely independent from any concerns for the grammatical form of utterances. There are many constraints on the grammatical structure of sentences, because they form a text in the specific rhetorical mode of information presentation. Thus, sentence plans are the locus of negotiation between requirements of the strategic planner and the tactical generator. We consider two examples of variation in strategies for information presentation. The first one is a description of the path of motion to reach a goal (following the examples above). The second one concerns presentation of pictures, in which the goal is to make the reader understand what is drawn in the pictures; this follows the research of (Кобозева 1997; Армеева 2001).

Tactical generation process in KPML starts with an input expression represented in the Sentence Plan Language, SPL, (Kasper 1989). An SPL example is given in Figure 4. The backbone of SPLs consists of concepts and relations from the Upper Model (UM) and Domain Model (DM). These are responsible for specification of ideational (i.e., propositional) meaning. Besides the ideational information, SPLs

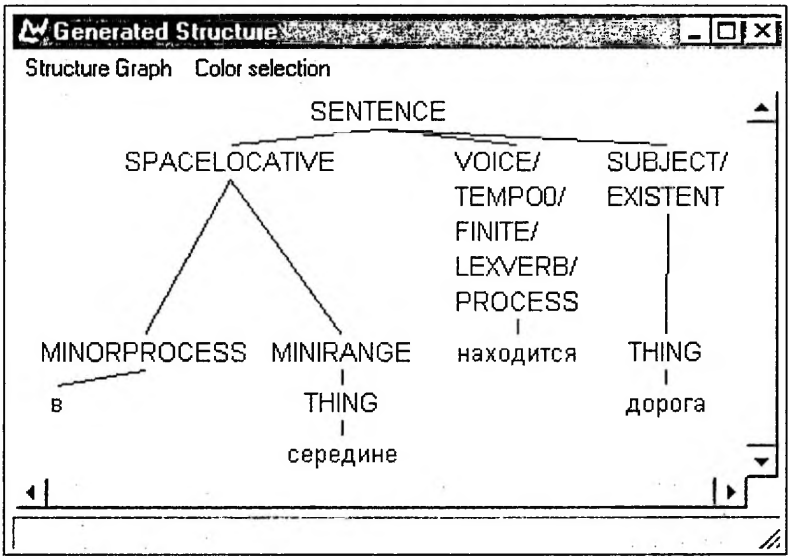


Figure 5.

also contain interpersonal information, which pertains to the exchange between the speaker and the hearer (e.g., `:speechact command`), and textual information, which pertains to the flow of text as a coherent whole (e.g., `:circumstantial-theme-q context`). In terms of the knowledge representation language, SPLs are designed as LISP forms that specify instances of UM/DM concepts (e.g., `existence, dm::road`) and include relations (`:spatial-locating`), inquiries (e.g., `:reference-type-q big`) and macros, which consist of a set of inquiries (e.g., `:tense future`, which invokes a list of semantic specifications). An SPL can also specify additional information, e.g. the target and generated forms and its gloss in another language.

Given an input such as Figure 4, the traversal of the grammatical system network is started. At each choice point (each system), a *chooser* is invoked that checks the input representation for the semantic grounds to make a choice. A chooser is a decision procedure that is associated with a system (Matthiessen, Bateman 1991). Its task is to mediate between semantic and grammatical information. A chooser is organized as a tree, the nodes of which are *inquiries*, which are the actual interpreters of semantic knowledge for the grammar. A response can be either computed by the inquiry as the result of

some reasoning on UM/DM or directly provided by the strategic planner. The process of choosing is applied throughout the traversal of the system network, invoking the chooser of each system, and realization statements successively build up syntagmatic structure at all ranks. Information about the lexical choice can be either directly provided in the SPL, or left implicit, so that the traversal of the (lexico)grammatical network constraints the set of choices for expressing a given DM concept. For example, `DM::CHOICE` in the domain of software manuals can be rendered in Russian as *выбирать, выбрать, выбор, задание, пункт меню*. The grammatical constraints on its realization may include the opposition of its realization at the level of clauses or nominal groups. In the former case, the grammar should choose between the perfective and imperfective aspects, thus constraining the choice to a single lexical item. The tree of constituents produced for SPL in Figure 4 is shown in Figure 5.

What are requirements on input specifications, as imposed by the tactical generator? First of all, the specifications depend on different types of sentences. For example, since the goal of the tourist information system is to describe a path leading to a location, one of the most natural types of texts is represented by a sequence of clauses:

The path expressions:

a) Выйдите на Садовое кольцо, перейдите дорогу, сядьте на троллейбус Б и выйдите на остановке "метро Маяковская". (imperative, perfective)
b) Выходите на Садовое кольцо, переходите дорогу, садитесь на троллейбус Б и выходите на остановке "метро Маяковская". (indicative, imperfective)
c) Выйдете на Садовое кольцо, перейдете дорогу, сядете на троллейбус Б и выйдете на остановке "метро Маяковская". (indicative, perfective)
d) Надо выйти на Садовое кольцо, перейти дорогу, сесть на троллейбус Б и выйти на остановке "метро Маяковская". (indicative, perfective, modal)

The imperative sentences that command the reader to perform a sequence of actions are typically expressed in Russian using verbs in the perfective aspect, because it highlights the result of the action. The imperfective aspect highlights the middle phase, while the reader is performing the action, so the imperative imperfective is possible only in two situations, when the repetition or iteration of a process is high-

lighted, or when a warning is issued in the case of negation. In the first case, the imperfective aspect expresses either one of the two meanings: habituality or multiple instantiation of the same process, e.g. *Нажимайте клавишу r каждый раз...* Thus, the grammar can assume that the default choice for imperative clauses is the perfective aspect, unless it is known that the action should be repeated (then, the strategic generator inserts `:repeated-q repeated` into the SPL).

In the case of indicative clauses, both perfective and imperfective variants are grammatically possible, though the strategic generator should ensure that it produces respective SPLs with the same aspect consistently. The semantics of the imperfective indicative clauses in 0 implies imaginary motion of the reader along a trajectory on the map. Thus, when the strategic planner inserts `:nonstatic-process-aspect-q activity-highlighted`, imperfective clauses are produced, while `:nonstatic-process-aspect-q result-highlighted` leads to perfective clauses. One more possibility is to use modal clauses with infinitives (the result is always highlighted in this case). The strategic generator controls this by inserting a macro: `:modal-property-ascription necessity`, which is expanded into a set of inquiries (as with the imperative clauses, the default choice of the aspect in this case is perfective).

When the strategic planner considers one style of 0 more suitable, it produces SPLs, which include all the elements that ensure a proper output, because the sentence-based tactical generator has no knowledge about the stylistic considerations employed by the strategic planner. On the other hand, the linguistic knowledge of the grammar encoded by the tactical generator should be sufficiently general in order to serve a much wider range of possible applications than simply a tourist information system or generation of CAD-CAM manuals. Hence the input is encoded not in terms of properties of a problem domain, say, `:style list-of-user-actions` for (b), and not in terms of grammatical features, say, `:aspect imperfective`, but in *semantic* terms relevant to the *grammatical* properties of respective outputs. Figure 6 shows an SPL for the third clause of all variants of 0.

Another issue that should be provided at the input of the tactical generator concerns semanticization, i.e. suiting the semantic specification of the SPL level to language-dependent means for expressing

```
(RST / RST-SEQUENCE
...
  :DOMAIN (V / DM::TAKE-LINE
(a) :speechact command
(b) :interactant-q empty :nonstatic-process-aspect-q
    activity-highlighted
(c) :interactant-q empty :nonstatic-process-aspect-q
    result-highlighted
(d) :modal-property-ascription necessity
       :DESTINATION (S / TWO-DIMENSIONAL-OBJECT :LEX
       TROLLEJBUS :NAME "B"))
```

Figure 6.

specific concepts. For example, the SPL in Figure 6 uses the Russian semanticization for the action of taking a bus:

садитесь на троллейбус Б.
sit-imper onto trolleybus B

This is formally represented as a motion process, the destination of which is the carrier. In contrast, the English and German expressions typically use different semanticization: a dispositive material process (*take*), which takes the carrier as its *actee*. Even though the DM concept (`take-line`) is the same, English and Russian SPLs use different relation slots (`:actee` vs. `:destination`).

The same verb of motion can be used in Russian in expressions of another type. *Садиться* can be used not only in the sense of using a carrier, but also for entering into a vehicle. The example 0 can be compared to:

садитесь на троллейбус.
sit-imper into trolleybus

in which the carrier is considered as a three-dimensional container, which allows entering into it. On the other hand, the carrier in 0 is not considered as a container, but as a means of transport, which is realized semiotically as a two-dimensional object, thus *na* (onto) is used in 0 as for other surfaces. Also, in 0 different types of carriers are possible, including persons, like *izvozchik* (cabman), *chastnik* (taxi driver), which cannot be used in the semanticization implied in 0.

One interesting case is the difference in the lexical choice for expressing the location:

The locative expressions for a person:

e) Вы находитесь около гостиницы Москва.

f) Вы окажетесь около гостиницы Москва.

Even though both expressions claim that a person is at a location, (a) implies that this is the current location of a person, i.e. the present tense, while (b) can occur in a description of the result of motion of a person that should follow instructions, i.e. the future tense. The deficiency of *находиться* (like all static processes, it lacks the perfective aspect) requires the choice of another lexical item. Once again, we have one DM concept and several possibilities for its lexicogrammatical realization.

In their experiments, Kobozeva and Armejeva studied the range of narrative strategies for describing the content of pictures to someone who has not seen them. Their corpus consists of recordings of 25 subjects, who produced spontaneous descriptions of 6 pictures each. One picture in their study is especially interesting for our purposes, since it shows a road that goes along several buildings. The speech acts are arranged in a sequence of speech acts that introduce objects and their properties. One possible strategy is to describe the picture along a trajectory of fictive motion of the observer around the scene. In such cases, the thematic development of the description is typically based on the locative theme: a clause starts with the location of an object on the picture (*Справа от дороги, За этой постройкой, На самом дальнем плане*); it is followed by a predicate stating the existence of the object (*находиться расположен, виднеется, стоит,* or various copula forms, including zero). The lexical choice of the verb is sometimes random, i.e. copula forms or such verbs as *находиться расположен* can be freely exchanged, but sometimes the choice is related to properties of the object: *виднеется* (is visible) is applicable only to objects that are on the periphery of the picture, *стоит* (stands) correlates with certain object types, like houses, trees, but not, say, lakes or roads. On the contrary, positions of roads can be semanticized not only by means of existential predicates, but also using verbs of motion: *идет, уходит, ведет.* At the end of the clause, goes the object itself (*небольшое здание, ряд деревянных построек*). An instance of this pattern is presented in Figure 4: the SPL represents a statement of existence, a one-place rela-

tion, the domain of which (in the mathematical sense of this word) is *дорога* (a road) and the lexicalization uses the verb *находиться* (to be located). The method of thematic development (location-to-object) is reflected by the inquiry-response pair `:circumstantial-theme-q context`.

Description of an object is not a straightforward mapping from a predicate to a lexical item. First of all, the speaker (or the strategic generator, in the case of a generation system) should choose the right level of abstraction for presenting the type of an object. Depending on the speaker's viewpoint, the same object can be presented as *здание, постройка, сооружение, церковь, собор, колокольня*. In addition, the speaker is free to choose attributes, on which s/he wants to focus the hearer's attention. Let's consider the case of the description of houses: they can be presented with or without specification of their size, shape, material, approximate date of construction, architectural style, and so on. Further on, when, for example, the size of a house is specified, the speaker has the choice of using a generic specification, e.g. *большой*. When one dimension is highlighted, then it can be specified explicitly, e.g. *высокий, длинный*. Finally, the specification of a generic size leaves a range of options, e.g. for describing small-size objects: *маленький, небольшой, крошечный, миниатюрный, мелкий, скромных размеров*, etc.

The two most frequent neutral options, *маленький* and *небольшой*, look almost synonymous, but their uses differ with respect to the speaker's attitude. It is much less likely that the speaker uses *небольшой* for referring to an object which is described with some specific emotions. The attitude expressed with *маленький* is typically positive *две маленьких сараюшки* (two small barns), unless other qualities or the noun itself imply the negative attitude, e.g. *маленькая заплёванная каморка* (a small dirty closet). The use of *небольшой* instead of *маленький* in both examples should contrast with the emotional attitude towards the object. On the contrary, *небольшой* is more felicitous in neutral contexts. Examples of SPL specifications corresponding to various size specifications are given in Figure 7. Take a note that in the output tree, *деревянный дом* is lexified explicitly, while the size specification is given indirectly and is lexified by the tactical generator.

At the beginning we mentioned that SPLs contain three types of information: ideational, interpersonal and textual. Now we can see

```
:GENERATEDFORM              "небольшой дом"
:LOGICALFORM (A1 / OBJECT :LEX DOM
     :SIZE-PROPERTY-ASCRIPTION  (S / QUALITY :REFERENCE-
     TYPE-Q SMALL))

:GENERATEDFORM              "маленький  дом"
:LOGICALFORM (A1 / OBJECT :LEX DOM
     :SIZE-PROPERTY-ASCRIPTION  (S / QUALITY :REFERENCE-
     TYPE-Q SMALL
                :INTERPERSONAL-TYPE-Q SYMPATHY-INTERPER-
                SONAL))

:GENERATEDFORM              "высокий  дом"
:LOGICALFORM (A1 / OBJECT :LEX DOM
     :SIZE-PROPERTY-ASCRIPTION  (S / QUALITY :REFERENCE-
     TYPE-Q BIG
                :SPATIAL-TYPE-Q DIRECTIONAL :SIZE-TYPE-Q
                VERTICAL-SIZE))

:GENERATEDFORM              "большой деревьянный дом"
:LOGICALFORM (A1 / OBJECT :LEX DOM
     :SIZE-PROPERTY-ASCRIPTION  (S / QUALITY :REFERENCE-
     TYPE-Q BIG)
     :MATERIAL-PROPERTY-ASCRIPTION  (S1 / QUALITY :LEX
     DEREVJANNYJ))
```

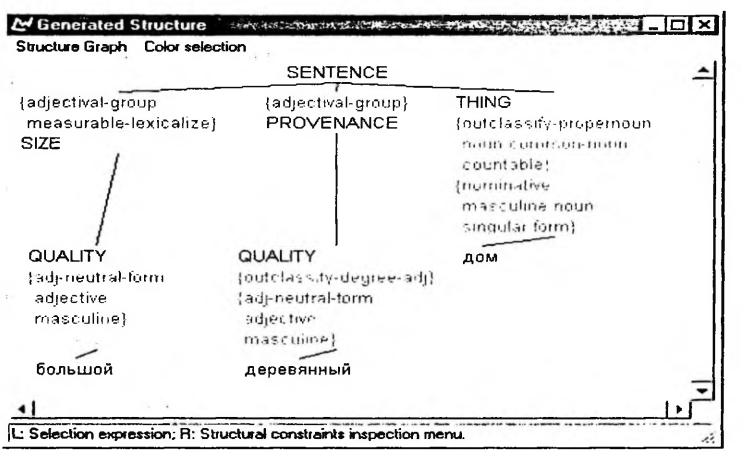


Figure 7.

how they are produced by the strategic generator. Ideational specifications come from predicates representing the scene, even though the mapping is not straightforward: the latter are language-independent statements, while the former are their language-dependent semanticizations. Specifications of interpersonal information come from stylistic constraints on the interaction between the system and the user, e.g. the choice between indicative or imperative clauses, the level of politeness, etc. Textual information ensures coherent discourse, so its source is the text planning mechanism of the strategic planner. This involves reasoning about what objects are introduced and when, how they are contextualized in the flow of discourse, and so on.

6. Conclusion and further research

In the paper we have presented our first approach to design a tourist multimodal information retrieval system. We have shown that for every component of the system there exist some prototype: START for the system design and multimodality, GIS for information storage, InBASE for analysis and partly for reasoning, AGILE and KPML for text planning and generation. Integration of the approaches would be favorable to clarify the fundamental processes of NL communication. Further researches concern explorations of features that could be inserted in any of the design constituents to reach the goals of communication – cognitive, communicative, spatial, visual, subject matter organization, language semantics, lexical semantics.

Literature

Армеева, А. Р. 2001. Когнитивная категория выделенности и ее языковые корреляты. Диссертация на соискание ученой степени кандидата филологических наук. М: МГУ им. М. В. Ломоносова.

Bateman, J.; Magnini, B.; Fabris, G. 1995. The generalized upper model knowledge base: Organization and use. – Proceedings of the Conference on Knowledge Representation and Sharing http://www.darmstadt.gmd.de/publish/komet/gen-um/newUM.html

Bateman, J. A. 1997. Enabling technology for multilingual natural language generation: the KPML development environment. – Journal of Natural Language Engineering 3:1, 15–55. http://purl.org/net/kpml

Bateman, J. A.; Kamps, Th.; Kleinz J.; Reicheberger, K. 2000. The Dart(bio) System: Constructive Text, Diagram and Layout Generation for Information Presentation. (forthcoming in Computational Linguistics)

Болдасов, М. В.; Соколова, Е. Г. 2001. Моделирование генерации описательного текста. Аксаково. – Труды Международного семинара Диалог'2001 по компьютерной лингвистике и ее приложениям. Нариньяни А.С. (ред.).

Halliday, M. A. K. 1978. Language as Social Semiotic. The Social Interpretation of Language and Meaning. London: Edward Arnold.

Halliday, M. A. K. 1985. Introduction to Functional Grammar. London: Edward Arnold.

Kasper, R. T. 1989. A flexible interface for linking applications to PENMAN's sentence generator. – Proceedings of the DARPA Workshop on Speech and Natural Language 1989. Available from USC/Information Sciences Institute, Marina del Rey, CA.

Кобозева, И. М. 1997. Как мы описываем пространство, которое видим: композиционные стратегии. М. – Труды международного семинара по компьютерной лингвистике и ее приложениям "Диалог'97". 132–136.

Koit, M.; Õim, H. 2000. A model of dialog and its application:building a communication traner. – Proceedings of the 7th National Conference on Artificial Intelligence (КИИ'2000) vol. 1. Moscow. 345–353.

Korelsky, T.; Kittredge, R. 1996. On the stratification of rhetorical structures. – The Moscow Linguistic Journal,Vol. 2, 212–226.

Kruijff, G. J.; Teich, E.; Bateman, J.; Kruijff-Korbayová, I.; Skoumalová, H.; Sharoff, S.; Sokolova, L.; Hartley, T.; Staykova, K.; Hana, J. 2000. A multilingual system for text generation in three slavic languages. – Proceedings of the 18th Conference on Computational Linguistics (COLING 2000), Saarbrücken. 474–480. http://Kwetal.ms.mff.cuni.cz/~korbay/Public/agile-coling00.ps

Matthiessen, C.; Bateman, J. 1992 Text Generation and Systemic Functional Linguistics: Experiences from English and Japanese. London: Pinter Publishers.

Mittal, V. O.; Roth, S.; Moore, J.D.; Mattis, J.; Carenini, G. 1995 Generating explanatory captions for information graphics. – Proceedings IJCAI, Montreal, August 20–25. 1276–1283.

Соколова, Е. Г.; Шаров С. А. 1998 К многоязыковой генерации руководств пользователя: начальный этап проекта AGILE. Казань: ООО"Хэтэр". – Труды Международного семинара Диалог'98 по компьютерной лингвистике и ее приложениям. Нариньяни А. С. (ред.). 848–895.

Зацман, И. М.; Куренков, С. А.; Лютый, А. А. 2001. Семиосфера электронного образа земли: основные структурные составляющие и принципы моделирования геотекстов. Аксаково. – Труды Международного семинара Диалог'2001 по компьютерной лингвистике и ее приложениям. Нариньяни А. С. (ред.) т. 2. 145–150.

Жигалов, В. А.; Соколова, Е. Г. 2001. InBASE: технология построения ЕЯ интерфейсов к базам данных. Аксаково. – Труды Международного семинара Диалог'2001 по компьютерной лингвистике и ее приложениям. Нариньяни А. С. (ред.). 123–135.

Winograd, T. 1972. Understanding Natural Language. N.Y., Edinburgh.

UPwards bound in Cora: Cora metaphors for UP

Eugene H. Casad
Summer Institute of Linguistics

1. Introduction

Typical of human categorization in general, the notion UP in Cora emerges from a comprehensive analysis of the data as a complex category with a multiplicity of meanings organized around prototypes grounded in a variety of domains. These are related to one another in a myriad of ways and to differing degrees along a continuum of similarity to one or more prototypes (cf. Langacker 1987: 49, 53, 69; Lakoff 1987: 76, 86, 90; Casad, Langacker 1985).

Cora, a Southern Uto-Aztecan language of Northwest Mexico, has a dazzling variety of locative and directional morphemes that to a large degree collectively shape its grammar and lexicon. These morphemes and lexical items include definite articles, demonstrative pronouns and adverbs, locative particles, procomplements, topographic adverbs, locative prefixes and postpositions (Casad 1977, 1982, 1984). My long time fascination with this system is reflected in a number of works beginning with Casad (1977) and includes the comprehensive study contained in my 1982 dissertation *Cora Locationals and Structured Imagery*. This present paper is based on further analysis of data contained in this latter work, as were several previous ones.

This study is organized as follows: Section 2 gives an overview of all the different kinds of grammatical items in Cora that can be glossed as one variety of UP or another. Section 3 begins with a discussion of the locative vs. directional distinction and continues by documenting some of the versions of UP related to paths in the domain of 3-D space, as well as some instances of abstract motion and proposes the conceptual metaphors that motivate them. Section 4 takes us into the domain of the human body, focussing on one subset of UP locations as identified by the presence in the sentence of the locative particle **y-én** "here-on:top". Section 5 summarizes the findings of this study.

2. Classes of morphemes meaning UP

In this section I present an overview of each kind of locative and directional morpheme that reflects some aspect of UP, describing each usage and providing initial evidence for the locative meanings cited.[1] The particular classes of grammatical items that I cite here include locative particles, first, in their topographical usages and then in a grammaticalized one in which the particle serves as a sequential sentence connector. Additional classes include topographic adverbs and locative verbal prefixes.

2.1. Locative Particles

The first example I cite illustrates the prototypical topographic usage of the particle **máh**, which I gloss as "right off there out in the slope."

(1) m-á-h tú wá-ta-t^{y}au
there-outside-slope we COMPL-REFL-find
'We ran across each other right out there at the side of the hill.'

The initial **m-** in this particle indicates MEDIAL distance from the speaker. It is deictic in its own right and carries the implication that the speaker has the particular location in sight and may well be signalling it overtly to his listener. Notice here, that the verb **wá-ta-t^{y}au** "we found one another" makes no specific reference to any peculiar location. It clearly implies that the interaction took place within a spatial setting without spelling out anything about that setting. The locative prefix **máh** handles that function, i.e. the meeting took place in the middle of the slope of the hill within eyeshot of the speaker as he relates the incident to his interlocutor.

Example (2) is introduced by the sentence initial locative particle **'áh**, which in its topographical usages can be glossed as "off out there in the slope." Here in its grammaticalized text-based usage, we can gloss it as "and then". This form is the DISTAL counterpart to the MEDIAL **máh** of sentence (1).

---

[1] The crucial evidence for such meanings comes, of course, from the explanations offered to me by native speakers of Jesús María Cora. I am very grateful to each of them.

(2) 'á-h pú nu'u í í-rá
outside-slope SUBJ QUOT SEQ NARR-face

'a-h-cuná
outside-slope- spring up

'Then, so they say, she jumped right up to meet him.'

Sentence (2) is taken from a folklore text about a young Cora woman whose betrothed was murdered just prior to the celebration of a prominent Cora ritual at which her parents requested her attendance. The sentence connector **'áh** "and then" links this sentence to the previous one in the text, which states that, while she was alone in the house after her parents had left, she heard a voice calling out and recognized it as that of her beloved. Sentence (2), introduced by **'áh**, presents her immediate response. The grammaticalized use here of **'áh** "off out in the slope" to indicate the temporal succession of one event by another is one piece of evidence that suggests that a pervasive Cora Conceptual Metaphor is TIME MARCHES UPHILL.

2.2. Topographic adverbs

The Cora locative system includes a set of sixty-six topographic adverbs whose conventionalized usages are based on a schematic model of the lay of the land, encompassing the domains of the river and the hill (Casad 1982; 1988: 352). As we shall see in this analysis, these usages are a special kind of reference point construction and, by virtue of that, are metonymic in nature (Radden, Kövecses 1999: 19). Sentence (3) presents a typical usage of a topographic adverb that designates an UP location within the hill domain.

(3) n^{y}i p-eyún a'-u-h -n^{y}éh
Q you-PROCOMP ABL-inside-slope -pass by

y-ú-h-t^{y}ivi
here-inside-slope-uphill

'Did you not previously come by here, going back uphill along the trail? (from the observer's position)

Sentence (3) also nicely illustrates the elaborate embroidery of the Cora locative system throughout the grammar. A procomplement of location, a verbal prefix sequence and a topographic adverb all figure

in the specification of locational information conveyed by this example. The repeated use of **u-** "inside" with each of the three locatives in (3) highlights in a striking way the reference point nature of the entire construction. The use of the procomplement **p-eyún** "you-thus:inside" equates the speaker's conceptualization of his interlocutor's path to an earlier conceptualized path that the speaker also has in mind. This usage of **p-eyún** reflects a conflation of two mental spaces in the sense of Fauconnier and Sweetser 1996 (cf. Rubba 1996: 234; Sweetser 1996: 320ff.).

The use of the adverb **y-ú-h-t^{y}ivi** "back up here in the hill" places the entire scene with the domain of the hill. This is reinforced by the use of the locative prefix sequence **a'-uh** "DISTAL-inside-slope" which is conventionally understood in its maximal extension to track a path from a horizontal base straight up a slope to the top of a hill. This mental tracking (cf. Matsumoto 1996: 137) is illustrated clearly by example (4).

(4) ú-h-t^{y}ee
inside-slope-long
'It is a long ways straight up the slope'

Sentence (5) is taken from a historical text about events of the Cristero period in Nayarit which followed on the heels of the Mexican revolution of 1910.

(5) á	mú	véhli'i	híra='**an-ta-kí'** i-ka				
there	they	close	circle=**top-across**-go:PL-SIM				
			around			MODE	
y-ú-h-t^{y}ap^{w}a			k^{w}áaša'ata		áihna	i	
here-inside-slope-upriver			San Francisco		DEM	ART	
Jesus	Pineeda	ahtá	aihná	i	a'ati	Lorenzo	Estraada
P.N.	P.N.	CNJ	DEM	ART	person	P.N.	P.N.

'That guy Jesús Pineda and that other fellow Lorenzo Estrada kept coming around, right upriver here, near the town of San Francisco.'

In example (5), the locative prefix sequence **an-ta-** contributes schematic information about the spatial setting within which the event occurred, but does not of itself tell the listener anything about the specific geographical region. Its usage presents the activity as repeatedly going from side to side across the "onstage" area of the speaker's focus of attention. The specific information about the

geographical setting is conveyed instead by the discontinuous locative particle + topographic adverb phrase that means "back off there upriver in the village of San Francisco." The adverb **yúht^{y}ap^{w}a** provides us with another version of UP: UP is THE DIRECTION FROM WHICH THE RIVER FLOWS.

2.3. Locative verbal prefixes

Cora and its closest Uto-Aztecan neighbor, Huichol, both display an elaborate verb morphology that includes a complex set of prefixes of location and direction which combine either singly or in various combinations with adjectives, nouns and verbs (Grimes 1964; McMahon, McMahon 1959; Casad 1977, 1982). This degree of elaboration of spatial elements in the verb word is unusual for Uto-Aztecan; it is highly reminiscent of the data discussed by Friedrich for Tarascan, whose wider linguistic affiliation is still unknown (Friedrich 1969, 1971).

Taken together, the prefixes in each sequence provide an integrated abstract characterization of the spatial domain within which a particular 'scene' is enacted (Casad 1993: 598). This scene is structured by the verbal process or stative relation signalled by the verb, adjective, or nominal stem that appears in the overall construction. The verbal process or state itself is the trajector of the relation that each component prefix denotes. It is also the profile determinant for the entire verb word at the semantic pole (cf. Casad 1982, Section 5.2). Typical examples of these prefixes are given in (6a and b). These examples display prefixes that relate in distinct ways to UP.

(6a) t^{y}i-n^{y}áh-šɨ wé'ira'a-ra'an hece
up-arrive-PAST flesh-his at
'A rash broke out in his skin.'

(6b) a-n-ká-kun
outside-top-down-hollow
'There is the mouth of a cave that drops straight down.'

To begin, the use of the prefix **t^{y}i-** "up" in (6a) indicates that a condition that had not previously been observed suddenly came into appearance somewhere on the surface of a person's skin. This usage can be called an 'Up and Out' version, carrying the implication that the prior state of the rash was localized inside the person's body and that some change brought the rash out into the open. This usage, of

course, is likely based on the BODY is a CONTAINER Conceptual Metaphor.[2] The version of UP illustrated by **an-ka-** "top-downward" in (6b) reflects the conceptual metaphor UP is at the TOP looking down. The scene in mind here is that of someone looking down at the edge of a vertically dropping cave such as those common in the Pyrenees.

3. Location versus direction

Particular variants of UP conceptual metaphors rest on the distinction between static location and dynamic motion. In several of my previous works discussing the semantics of the Cora locative prefixes, on intuitive grounds, I have asserted that each of the prefixes has both static locational meanings and dynamic directional ones.[3] I have also made the assumption that these meanings reside in the semantics of the locative prefix itself and are not derived from the meanings of the verb stem with which it combines. The following data, I believe, suggest that such directional meanings are inherent in the meanings of all of the locative prefixes and prefix sequences.

3.1. The verb stem -n^{y}éera 'to look at X'

Here I consider the semantics of the verb stem **–n^{y}éera** 'to look at X' and a pair of the locative prefix sequences that occur with it. The scenario is that of some person instructing another person which way to cast his/her gaze. This verb stem has a directionality of its own as part of its meaning, an inherent directionality that is not the entire story, but rather must be related to the conventional viewing stance of the one giving the instructions, who, in this case is viewing a given scene objectively (cf. Langacker 1990). The examples in (7a-b) were given as instructions during an eye examination.

(7a) **a-tá-**n^{y}eeri-či
outside-straight-look-IMP
'Look straight ahead.'

---

[2] Further motivation for this conceptual metaphor can be deduced from other Cora data that I do not consider here.

[3] This section is adapted from a longer paper (Casad 2000).

(7b) **a-h-tá-**n^{y}eeri-či
outside-slope-straight-look-IMP
'Look off to one side.'

Note first, as the respective glosses suggest, that the patient was being instructed to look in a particular direction and not at a specific object located in a particular position. Had the patient been instructed to look at a particular object, the interpreter would have used a very different Cora expression.

The use of these particular locative prefixes is thoroughly motivated: they provide the necessary clues to the hearer (= patient being examined) that he/she is to look off in a particular direction, rather than simply look all around. In effect, the implicit directionality in the meaning of the verb stem itself is not sufficient to tell the hearer where to look; instead, the hearer needs to have this directional information specified differentially for each distinct situation. This explains why there is no bare stem imperative form ***n^{y}éeriči**. In particular, the locative prefix combination **a-tá-** "outside-straight" is conventionally understood as designating a path straight ahead of the human being as he/she is in a standing position and faces off in some direction or another.[4] On the other hand, **a-h-tá-** "outside-slope-straight" is conventionally understood as designating a path leading off to one side of the person involved. The use of the locative prefix **h-** "slope" in this prefix sequence places the path of visual perception at right angles to the canonical vertical orientation of the body of the human observer, i.e. this path goes right over the observer's shoulder, from either a standing or a sitting position.

3.2. Paths, directions and locations

Other variations of UP can be pinpointed in terms of the orientation of a path along a verticality dimension (cf. Serra Borneto 1996: 466–467). In its most concrete form, a path can be thought of as how one

[4] Supporting this analysis is the observation that the Cora expression meaning 'to be awake' obligatorily includes the locative prefix sequence **a-ta-**, as in (i)

(i) a-tá-n^{y}ee
outside-across-see
'he is awake'

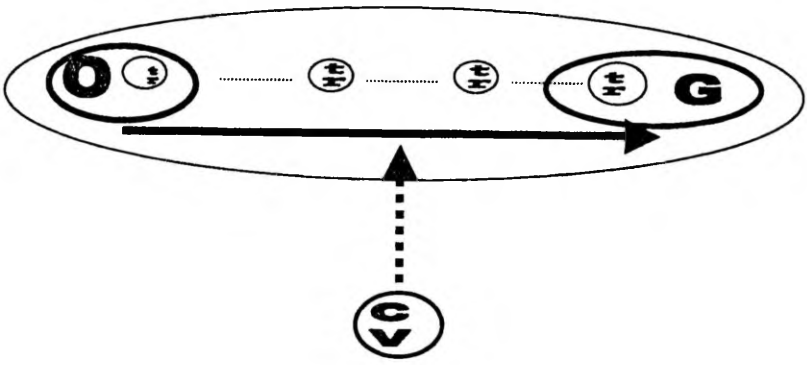


Figure 1. The Directed Path Schema

gets from one point in three dimensional space to a second point in that same space. I take this characterization to be the prototypical version. In a more schematic sense, I take the term to refer to a conceptual schema (or, in Lakoff's terms, an image schema (1987: 275)) that has salient endpoints, the Origin, or Source, and the Goal, and a set of intermediate points that constitute the extension of that path over an unspecified base (cf. Casad 1993; Hawkins 1984; Talmy 1975: 182; Radden 1996; Smith 1993, among others). The differential end points give the path an inherent orientation that has its basis in a set of related concepts such as the natural movement of an entity through space or the typical posture that a perceiving entity assumes for viewing a scene. The image schema, then, is given in Figure 1.

The initial point on the path is labelled 'Origin' (O), whereas the final point is labelled 'Goal' (G). In addition, the path itself is situated within an unspecified Base, and the canonical viewing point is located at a neutral point at the middle, but spatially removed from the path itself. The movement of a trajector along that path is indicated by the series of three circles labelled 'tr' and the dotted lines between them are construed as indicating continuous motion along the path.

Paths, of course, go in many directions, including UP. The first example I cite of an upwards going path is given in (8). It comments about a man climbing a hill.

(8) 'á-h-tʸi tɨ a'-u-h-méh
outside-slope-uphill SUBR ABL-away-slope-go from
'The man who is going uphill along the trail.'

The DISTAL **a'-** in (8) situates both the directionality and goal of the path the trajector is following in a region that is spatially removed from the speaker's location. The origin is effectively backgrounded and in this instance is anchored to the location of the speaker. The topographic adverb **'áhtʸi** is an 'outside' form that indicates that the climber's path is going upwards across the side of the hill in contradistinction to a path right straight up the slope from the viewer. The form of the verb stem indicates that this is an ongoing process, i.e. it is giving us an imperfective view of the scene. The adverb itself gives us another version of UP: UP is IN THE SLOPE HEADING TOWARDS THE TOP OF THE HILL.

Figure 2 presents the contrast between the speaker's line of sight right up the slope (the pair of arrows at the middle of the Canonical Viewer's field of vision) and his/her line of sight obliquely across the slope. This idealized viewing position is indicated by the letters CV in the oval located in the horizontal section of the schema. Side views of the 'hill' are indicated by the arrows leaving the Canonical Viewer's position at oblique angles.

In contrast to (8), the adverb **m-eh-tʸi** in (9) signals the prominence that the speaker is placing on the proximity of the the climber in his particular location along the path that he is following. This construal is also an imperfective one. The speaker is simply reporting the event as an ongoing upwards climb without any salient reference to either the starting place or the end point. An ablauted form of the NARRATIVE MODE prefix **hí'i-** is seen here in contrast to the use of the DISTAL **a'-**, given above in (8).[5]

(9) m-eh-tʸi tɨ hu-'u-h-mé
there-slope-uphill SUBR NARR-inside-slope-go
'the one who is climbing up by the trail'

[5] In other contexts, NARRATIVE MODE **hí'i** means 'BE', as seen in (i).
(i) hí'i tevi
BE-person
"he is a person"

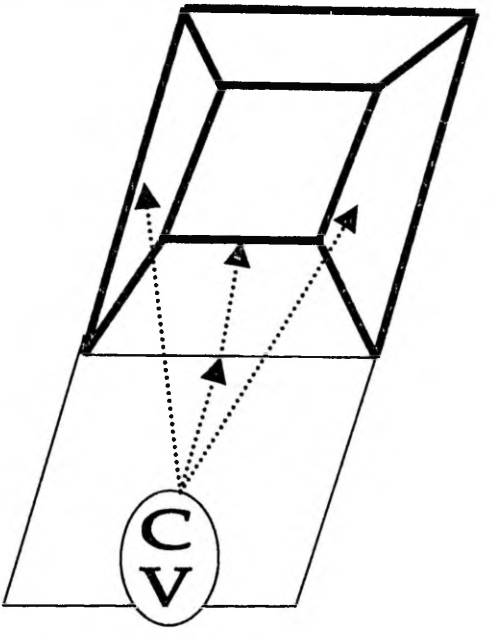


Figure 2. Looking upslope vs. looking across the slope

On the other hand, Cora speakers can specifically focus on either the starting point or the end point of an upwards oriented path. When focus is placed on the end point of such a path, the speaker can draw on the locative particle **an** to signal that end point. In this case, the version of UP reflects the conceptual metaphor UP is AT THE TOP OF THE HILL, as illustrated in (10).

(10) án nú=a'-u-h-n^{y}éh
top I=away-inside-slope-arrive
'I arrived up there on top by way of the slope.'

Sentence (10) presents us with a perfective view of the scene and its focus on the goal of the path the speaker followed. The upwards oriented path with its highlighted final position at the top of the hill is illustrated in Figure 3. The trajector's path, starting in the flat at the foot of the hill, crosses the face of the slope obliquely and stops in the area that constitutes the head of the slope. The overt indication of the trajector and his path underscores Talmy's point that the set of basic elements in the motion situation includes a mover (Talmy

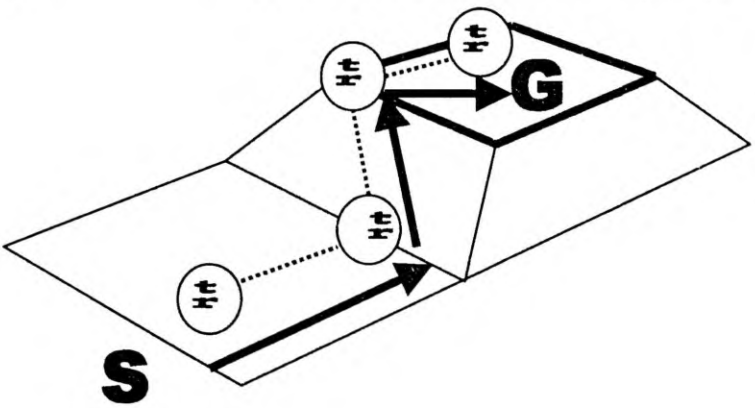


Figure 3. Up at the top of the hill

1975: 182; 198–199); this is one instantiation of the force vector that determines the orientation of the Path in the Source-Path-Goal Schema (Radden 1996: 436).

The directional usages of the Cora locative prefixes reflect a variety of situations beyond that of physical motion of discrete animate entities through three dimensional space. Some directional usages clearly are based on what has come to be termed 'Abstract' or 'Fictive' Motion (cf. Langacker 1991: 149ff.; Matsumoto 1995; 1996; Talmy 1996; Matlock 2001), an idea that links the perception of an ordering of sucessive positions in physical motion to an analogous mental ordering of successive states or relations in other domains and that allows speakers to construe processes wholistically. This enables them to build up the conception of an entire trajectory that constitutes a single gestalt, a psychological entity composed of a continuum of component states that can be accessed simultaneously (Langacker 1991: 153). This is very much what has to be done for a Cora speaker to make an evaluation of a static situation such as that described earlier in (4) and now in (11).

(11) mú=há'-uh-kun t^{y}ásta'a
there=away-slope-hollow cave
inside
'There is an upwards-going hole off there in the corner.'

The directionality of the cave is clearly expressed by the locative prefix sequence **á'-uh-** "DISTAL-inside:upward" in (11). The prefix

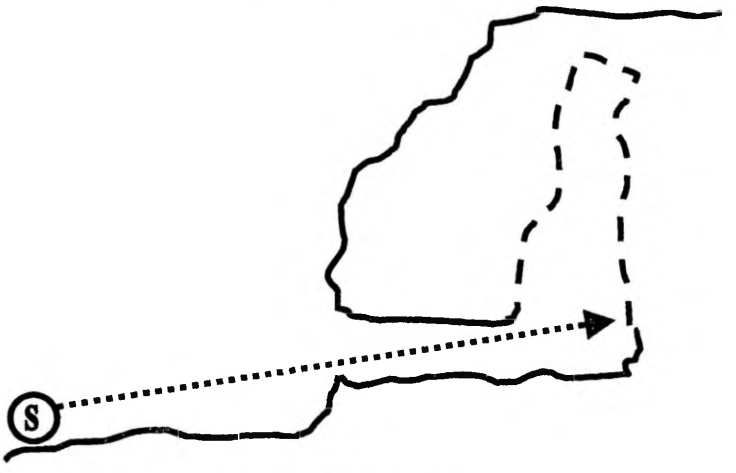


Figure 4. mú á'uh-kun ' cave going upwards'

sequence **uh-** is conventionally understood to reflect a path that begins at some point within a horizontal surface, follows it for an indefinite distance and then makes an abrupt turn into a vertically oriented path. The predicate adjective root *kun* designates a bounded area devoid of material content, an area that prototypically is globe shaped and thus is neutral with respect to any axial orientation. The use of the Medial form of the sentence initial locative particle again indicates the deictic nature of the situation in which the speaker is singling out for comment a specific location within eyeshot. The use of the distal *á'*- indicates that the upwards going hole is outside of the speaker's immediate neighborhood and is thus unaccessible in visual terms. All this is depicted diagrammatically in Figure 4.

The Speaker's canonical viewing position from outside the cave is indicated by the circle labelled S. The speaker's line of sight that displays the extent of the scene that is visually accessible is indicated by a dotted line with an arrowhead that designates the directionality of that line of sight. The cave itself and its enclosing ground area are given in cross profile, with the part of the cave that extends upwards given as a broken line. In short, the speaker cannnot really see the upwards extension of the cave, but knows on other grounds that it is oriented upwards. This usage is obviously based on metonymic relationships defined in terms of cultural Idealized Cognitive Models, as discussed by Radden and Kövesces (1999: 22).

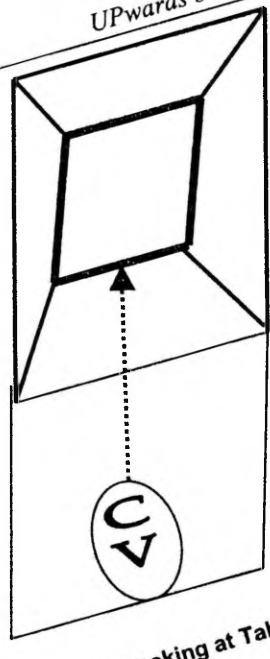


Figure 5. Looking at Table Top Mountain

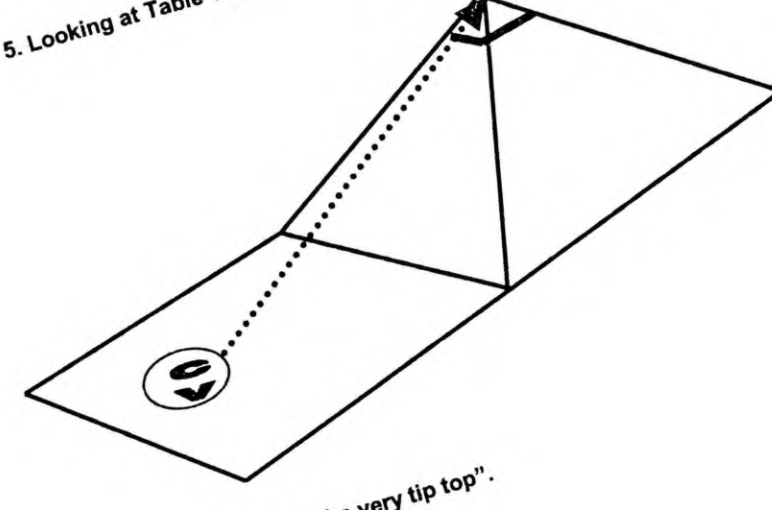


Figure 6. ant^{y_i}-: "Up at the very tip top".

3.3. Combinations of distinct locatives

Additional varieties of UP in Cora accrue from the combination of distinct locative morphemes. For example, the prefix sequence **a-** plus **n-** plus **t^{y}i-** "up" can be glossed as "there at the very tip top of" (or: "at highest point on top of"). Typical **an-t^{y}i-** locations include the top of a hill or cliff, as in (12a–b).

(12a) ta-'a-n-t^{y}i-n^{y}éh
we-outside-top-up-come:PAST
'We came to the top of the hill.'

(12b) a-n-t^{y}í-t^{y}ee
outside-top-up-long
'It is a long ways up from observer's position at foot of cliff to the top of the cliff.'

These two usages of **an-t^{y}i-** reflect physical motion (12a) and abstract motion (12b), respectively. The augmentative role of **t^{y}i-** "up" results in a meaning for the prefix sequence of "up at the very tip top" in the domain of the topography in clear contrast to **an** "on top of". The difference perceptually is that **an** could designate the entire expanse of the top of a mountain such as Cape Town's Table Mountain (Figure 5), whereas **an-ti-** would be fully appropriate for referring to the top of the Matterhorn.

The base for **an-t^{y}i-** differs from that of Figure 5 in that the 'head' of the slope area is typically rather prominent in the entire configuration. The shape of the configuration is such that **an-t^{y}i-** locations are typically peaks, points, or tips. Prototypically, these prominent peaks are oriented vertically, as in Figure 6.

It is easy to see how the configuration in Figure 6 fits the situations described by **an-t^{y}i-n^{y}éh** (12a) and **an-t^{y}í-t^{y}ee** (12b). In each case the progress of the conceived motion is plotted against a path that goes from the speaker's canonical position at the foot of the slope to the landmark position at the highest point at the head of the slope.

4. UP in the domain of the human body

As we have already seen, the human capacity to categorize entities and relationships within numerous conceptually distinct domains partially motivates many of the varieties of UP. Beyond the results

from selecting different domains, examples already discussed in this paper have also begun to illustrate the role of metonymies in motivating Cora conceptual metaphors. In many cases, metonymic relations figure in the emergence of specific cultural models (cf. Barcelona 2000: 6; Panther, Radden 1999: 2). These relationships, of course, obtain within a single domain (Croft 1993; Gibbs 1999: 62; Ruiz de Mendoza I. 2000: 115).

In this section of the paper I illustrate these points by turning to the domain of the human body. Crucially, the human body provides a handy framework for exploiting a wide variety of metonymic relations and undergirding cultural conceptual metaphors. The human capacity for focussing attention on an area of specific scope and using a wider scope as the frame for characterizing smaller regions and configurations within that frame accounts for many of the Cora usages within this domain for elaborating the concept of 'upness' that Cora speakers notice.

4.1. Overview of bodily UPs

There are actually several kinds of UP locations in the domain of the human body. These are signalled by both locative particles and locative prefix sequences. The locative particles map out over the human body, as it is viewed in a prototypical standing position. They collectively form a four-fold division of body areas organized in part by the basic 'outside' versus 'inside' distinction that is so central to the Cora locative system (Casad 1982: 376–380; Casad, Langacker 1985). This mapping also includes two slope-oriented notions: the 'face of the slope' versus the 'head of the slope' (Casad 1982, Ch 7). In turn, the locative prefix sequences allow the speaker to shift his focus of attention to particular parts of the body, construing spatial relationships in a variety of ways.

In the locative particle + possessed bodypart construction, the proximal, medial, and distal forms of the locative particle correlate with first, second, and third person, respectively, as illustrated by (13a–c).

(13a) y-é-n n^{y}a-muu-ce-'e
here-outside-top my-head-ABS-on
'right here on the top of my head'

(13b) m-á-n a-muu-ce-'e
there-outside-top your-head-ABS-on
'right there on top of your head'

(13c) á-n muu-ce-'e-n
outside-top head-ABS-on-his
'there on top of his head'

The successive locative particles link to the prototypical speaker–hearer domain in a striking way: **yén** "here on top" designates the top of the speaker's head (13a), whereas **mán** "right there on top" designates the top of the addressee's head (13b) and **án** "off there on top" indicates the top of the head of the person being discussed (13c).

4.2. The most salient bodily UP

The examples in (13) also pinpoint the most salient UP relationship in this domain: UP is AT THE TOP OF THE HEAD. There is clearly a perceptual basis for the use of the UP on TOP locative particle to refer to the top of the head, i.e. the top of the human head, is a horizontally extended surface (cf. Casad 1998: 5, 2000). It is also located at the uppermost extremity of the verticality scale as it is matched to a person in standing position.

Other kinds of UP contrast with the head locations. Thus, there are a number of body part locations which are construed as being 'UP in the slope' locations. Sentence (14) identifies one such location.

(14) y-á-h pú=nʸa-kʷi'i y-é'h
here-outside- slope SUBJ=me-hurt here-outside-slope

nʸa-tɬ́hči hece
my-thigh at

'It hurts me right here in the thigh muscle.'

The first occurrence of the proximal form of the locative particle **yáh** in (14) locates the area of pain somewhere along the vertical alignment of the body at a point significantly enough removed from ground level to be perceived as an 'UP in the SLOPE' location. The precise identification of the area of pain is given by the complex locative phrase that follows the main verb. This consists of a second occurrence of **yáh** (modified phonologically to **yéh**) in construction

with a postpositional phrase, the object of which is the body part term **tɨ̂hči** "thigh."

4.3. The diversity of bodily UPs

Still other versions of UP arise from the conventional usages of the locative prefix sequences to designate particular body part areas. The prefix sequences themselves are not *names* for the body parts, however[6], nor are they required to be used in constructions along with the corresponding body part names. Nevertheless, ambiguities of reference do arise. As seen in (14), these are clarified by using the appropriate body part name as the object of a postposition.

The following set of examples illustrates the diversity of UP locations within the domain of the human body and the specific locative prefix sequences that designate those UP locations. The 'head of the slope' locations include the top of the head (15), the tip of the nose (16), the tip of the finger (18a) and the buttocks (18b), the mouth (19) and upper lip (20), the chest (21a), and navel (21b), as well as the upper surface of the hand (22). Additional locations are, the beltline (23), the fingernail (24), and, finally, the chin (25). All these correlate with **yan**, **man**, or **an**. For the sake of brevity, I cite only the first person singular possessor forms. To highlight the role of the distinct locative prefix sequences, I draw on the same verb stem for all of these examples, i.e. **tu'a** "to hit" and I cite it only in its simple past tense usage.

(15) na-'a-vá'a-tu'a y-én nya-muuce-'e
me-outside-back-hit here-top my-head-on
'It hit me right here on top of my head.'

As I have already noted, the most obvious UP location in this domain is the top of the head. This is illustrated again in (15). The typical locative prefix sequence to designate this UP location is **a-va'a-** "outside-back:surface."

[6] This is an interesting typological parallel between the Cora prefixes of location and the locational suffixes of Tarascan, discussed by Friedrich (1969, 1971).

4.4. An UP twist at the nose

A second UP on TOP location is given in (16).

(16) na-'a-n-t^{y}i-tú'a y-én n^{y}a-cú'u hap^{w}a
me-outside-top-up-hit here-top my-nose on
'It hit me right here on the tip of my nose.'

The relevant UP location in (16) is the tip of the nose, as is indicated by the postpositional phrase. The use of the locative prefix sequence **an-t^{y}i-** links up with the conceptual metaphor UP is AT THE VERY TIP TOP, but the linkage involves an interesting twist of its own. To see this, I cite two additional usages of **an-t^{y}i-** from a different pair of domains.

The examples in (17) show extensions of **an-t^{y}i-** in which it is not obvious how its topographical meaning relates to particular usages in other domains.

(17a) ra-'a-n-t^{y}i-véihči
DISTR:SG-outside-top-up-cut
'He is going to cut the log up into sections.'
(The log can be either lying down or standing up.)

(17b) láapi tɨ́ kɨ́n íra-'an-t^{y}í-pi-t^{y}e-'e-n
pencil SUBR with coil-top-up-sharp-CAUS-APPLIC-PRTC
'That with which pencils are sharpened.'

To begin, **ra'ant^{y}ívéihči** (17a) refers not to the tip of a tree top, but rather to the accessible, and hence, prominent, point of the tree or log where the cutting is taking place. Crucially, the result of the cutting entails the complete separation of one section of the tree trunk or log from the rest. The notion "at the highest point" fades out to more like "the most prominent point", i.e. the point on the section of the log that marks where it was separated from the rest.

The idea of a point is especially relevant to **ira'ant^{y}ípit^{y}e'en** (17b). However, the orientation of **an-t^{y}i-** to the verticality scale fades out in this case. Within the scope of the base object pencil, the point is at one end of an axis. Generally, the perceived long axis of such configurations may become assimilated to the verticality scale with the result that **an-t^{y}i-** comes to designate the tips of all kinds of projections, whether vertically inclined or not (Friedrich 1969: 20; 1971: 28). This line of extension is illustrated pictorially in Figure 7 (a–c).

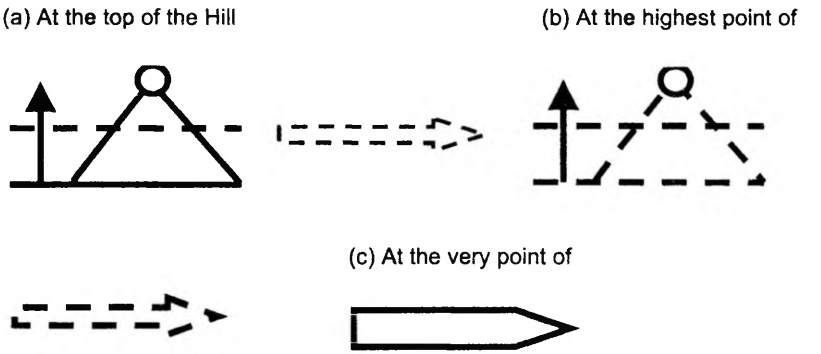


Figure 7. Extensions of an-t^{y}i-

Figure 7(a) gives a cross-section pictorial representation of a scene in which an entity is located at the apex of a triangle, which I use here to represent a hill. This is paired with a verticality scale that consists of a horizontal base to which a perpendicular, upwards oriented arrow is attached. The horizontal broken line crossing both the verticality scale and and the hill at roughly the midpoint separates the hill domain into UPPER and LOWER sectors, so that UP in the hill domain is taken to be substantially removed from the horizontal base level.

Figure 7(b) gives a pictorial representation of the semantic extension of **an-t^{y}i-** from its meaning of "UP at the Very Tip Top" to its new meaning of "UP at the HIGHEST POINT." This extension is motivated in part by the shift from the topographic domain to a wide range of other domains. It is augmented by semantic specializations that subsequently arise through conventional usage. The schematic nature of the particular entity invoked in any of these domains is indicated by using a broken line for the triangular base. Crucially, the verticality scale remains salient to the meaning of **an-t^{y}i-** in all of these usages.[7]

Figure 7(c) represents pictorially an even further attenuated semantic extension of **an-t^{y}i-**. In this version, the notion of verticality has completely faded from the picture and **an-t^{y}i-** can here be glossed as "At the very tip of." Diagrammatically, I model it as an

[7] This is one detail that I have changed from my original formulation of this semantic extension, given in Casad (1988: 357).

elongated horizontal mass which has a pronounced narrowing to a sharp point at one end. This entire narrowed area is construed as the appropriate **an-t^{y}i-** location.

The extension of **an-t^{y}i-** to mean "at the point of" is the link to two seemingly unusual usages within the domain of the body. These are the fingertips (18a) and the buttocks (18b).

(18a) na-'a-n-t^{y}i-tú'a y-én n^{y}a-šit^{y}é hece
me-outside-top-up-hit here-top my-finger on
'It hit me right on the tip of my finger.'

(18b) na-'a-n-t^{y}i-tú'a y-én n^{y}e-ḱ́ca hece
me-outside-top-up-hit here-top my-buttocks on
'It hit me right on the buttocks.'

The usages of **an-t^{y}i-** in (18a and b) do not to appear to be readily explicable in terms of only assuming a normal vertically oriented standing position for the human body. However, in the usage designating the fingertips, the extension of **an-t^{y}i-** from "at the very tip top of" to "at the very extremity of" is straightforward. It simply requires us to focus our attention on the region of the hand and forearm, noting the configuration that would result from the speaker holding his hand up to indicate the site of impact. In (18b), in normal standing position, the buttocks do not protrude prominently from the vertical axis of the general profile of the body, but in squatting position, they decidedly do so. In any event, the use of the locative particle **y-én** "here-on:top" contributes the information that the contact event was localized in an exterior, visually accessible surface.

4.5. UP at the mouth

The mouth is also an UP on TOP location, as the postpositional phrase in (19) clearly shows.

(19) na-'a-n-káa-tu'a y-én n^{y}e-t^{y}en^{y}e-'e
me-outside-top-down-hit here-top my-mouth-on
'It hit me right here in the mouth.'

This usage is at least partially motivated by another conceptual metaphor that we have already mentioned, that is, UP is AT THE TOP LOOKING DOWN. Again, the shift in domain from the topography to the human body entails a concomitant shift in the way that the situation is construed. Instead of the UP on TOP location being where

someone is standing, the UP on TOP location is where the food is placed in order to chew and swallow it; the direction that the ingested food subsequently follows to the gullet, of course is a downward one, as the use of the prefix sequence **an-kaa-** indicates.

The use of the locative particle + postpositional phrase construction **y-én n^{y}e-víɨ hece** here-top my-lip-on "right here up on my upper lip" in (20) shows that the upper lip is also construed as an UP on TOP location.

(20)	na-'a-n-táa-tu'a	y-én	n^{y}e-víɨ	hece
	me-outside-top-across-hit	here-top	my-lip	on
	'It hit me right here on the upper lip.'			

The conceptual metaphors that partially motivate this usage include UP is ON TOP and UP GOES ALL THE WAY ACROSS. In this case, the perception of the vertical ordering of the upper lip *vis à vis* the lower one and the perceived total lateral extension of the lips, viewed with the mouth closed allows the Coras to extract an ACROSS relation which is paired with the construal of the mouth itself as being an ON TOP location. In short, two conceptual metaphors apparently work together to motivate this usage.

4.6. Middle UPs

Yet another conceptual metaphor at least partially motivates the usage of the prefix sequence **an-t^{y}a-** "UP-in the middle" to designate the chest area in (21a) and the belly area in (21b).

(21a)	na-'a-n-t^{y}áa-tu'a	y-én	n^{y}a-taviice-'e
	me-outside-top-middle-hit	here-top	my-chest-on
	'It hit me right here in my chest.'		
(21b)	na-'a-n-t^{y}áa-tu'a	y-én	n^{y}e-sipuuce-'e
	me-outside-top-middle-hit	here-top	my-navel-on
	'It hit me right on the belly button.'		

The locative particle plus postpositional phrase construction in (21a) shows that the chest is construed as another UP ON TOP location. This is reinforced by the use of the locative prefix **an-** "on top" in the prefix sequence **an-t^{y}a-**. The second locative prefix **t^{y}a-** invokes a complex notion of "in the middle," a relationship between a more inclusive bounded area and a smaller region totally residing with the confines of the larger one. For the human body, the full range of **t^{y}a-** extends along the profile of the standing human body from just

above the ankle to the shoulder. The fact that the UP location occupied by the chest is also an IN THE MIDDLE location reinforces my hypothesis that the conventionalized usages of the locative prefixes to designate body part locations generally assume that the speaker is viewing the human person as being in a normal standing position. The fact that the chest and belly (21b) area is a relatively extensive region suggests that the conceptual metaphor at work here is best characterized as UP is AN AREA IN THE MIDDLE OF THE SLOPE.

4.7. UP at the hand

The hand is also taken to be the base for specifying an UP on TOP relation, as seen by the use of the locative particle **y-én** 'here on top' with the postpositional phrase in (22).

(22)	na-'a-náa-tu'a	y-én	nʸa-mʷáhka'a	hece
	me-outside-perimeter-hit	here-top	my-hand	on
	'It hit me right here on my hand.'			

This usage clearly shows the way that seemingly contradictory models can be combined into an overall coherent model. From the perspective of the human body as a whole, the hand is an 'on the periphery' location, as evidenced by the use of the locative prefix sequence **a-naa-** in this example. However, when we narrow the focus of our attention and place it on the configuration of the hand itself, at least in an English speaking mode, we distinguish between the 'outer' surface of the hand and the 'inner' surface. In the Cora view, this outer surface is construed as an UP on the TOP location, implying the conceptual metaphor UP is on the OUTER SURFACE of the HAND[8]. This particular conceptual metaphor comes into play in motivating other usages that I discuss here also.

Earlier, with respect to (18a) I discussed another 'up on top' prefix sequence that related to the hand, i.e. the use of **an-tʸí-** to designate the fingertips. Here we turn to the last example that relates to the hand. As the locative particle plus postpositional phrase construction in (23) shows, the exposed surface of the fingernail is also an UP on TOP location.

---

[8] For the Cora, the inner surface of the hand is designated by the prefix sequence **u-ii-tʸa-** "inside-Facing:toward-in:middle", which we can paraphrase as "face to face in the middle".

(23) na-'a-vá'a-tu'a y-én n^{y}a-kúcape'e hece
me-outside-back-hit here-top my-fingernail on
'It hit me right here on top of the fingernail.'

In example (23), the point of contact involved in the hitting action is signalled by the locative prefix sequence **a-va'a-** which carries its static locational meaning of 'outside-cover'. This usage indicates a perceptually accessible non-frontal upper surface (cf. Casad 1982; 1998; 2000). This usage invokes the conceptual metaphor UP is AN OUTER SURFACE OF THE HAND, which appears to be a special case of a much more general conceptual metaphor UP is AN OUTER SURFACE.

4.8. UP is at the beltline

The beltline is also construed as an UP on TOP location, as illustrated in (24).

(24) na-'a-n-ku-rá'a-tu'a y-én
me-outside-top-around-corner-hit here-top

n^{y}á=ta-'a-vá'a-n-hɨ'ɨke-'e
I:SUBR=SUBR-outside-back-REFL-bind-in
'It hit me right here on my beltline.'

The extensiveness of this UP on TOP location is indicated by the triple locative prefix sequence **an-ku-ra'a-** "top-around-back-and:forth", which can be glossed as "going all the way around the top." In this example, the precise nature of this UP on TOP location is spelled out by the adverbial clause which we can gloss as "where I bind myself up." The locative prefix sequence employed here is **a-va'a-** "outside-coming:toward" and its usage here may well indicate the action of pulling a belt into an increasingly restricted closed curve as one tightens the belt to its proper fit. In short, it seems clear that at least partially motivating the usage of both the locative particle **y-én** "here on top"' and the locative prefix **an-** in **an-ku-ra'a-** is an additional conceptual metaphor, i.e. UP is AT THE BELTLINE.

4.9. UP and DOWN at the mouth

The postpositional phrase in (25) indicates that the lower jaw region of the face is also an UP on TOP location.

(25) na-'an-káa-tu'a y-én nʸa-'ayáɨ hece
me-top-down-hit here-top my-jawbone on
'It hit me right here on the chin.'

The joint usage of **an-ka-** "on:top-downwards" with **y-én** "here-on:top" invokes at least two conceptual metaphors. The UP on TOP is AN OUTER SURFACE metaphor is invoked by the image schema that encapsulates the notion of physical contact that is associated with the action of hitting (cf. Vandeloise 1996: 541; 546). The use of the locative prefix sequence **an-káa-** invokes a specialization of the UP is AT THE TOP LOOKING DOWN, i.e. UP is TRACKING DOWNWARD FROM THE MOUTH. In other words, this usage invokes the notion of abstract motion as part of its meaning.

To summarize this section, we have examined a variety of what the Coras construe as UP locations within the domain of the human body. These varieties are designated in the grammar and lexicon of Cora by body part terms, locative particles and verbal prefixes. The usages of these grammatical elements, upon analysis, display a family of conceptual metaphors, some of which themselves are reflections of even more basic conceptual metaphors. The overall domain, as we have seen, is a conceptuial frame for defining the particular metonymies that partially motivate the particular metaphors (cf. Panther, Radden 1999: 9; Koch 1999: 146, 151).

Particular conceptual metaphors that we have discovered include the following: UP is AT THE TOP OF THE HEAD; UP is IN THE SLOPE; UP is AT THE VERY TIP TOP; UP is AT THE HIGHEST POINT; UP is AT THE VERY TIP; UP is AT THE TOP LOOKING DOWN; UP is ON TOP and UP GOES ALL THE WAY ACROSS; UP is AN AREA IN THE MIDDLE OF THE SLOPE; UP is ON THE OUTER SURFACE OF THE HAND; UP is AT THE BELTLINE and UP is TRACKING DOWNWARDS FROM THE MOUTH.

5. Conclusion

To conclude, in this paper we have seen the ubiquity of UP in Cora from several angles. In section 2, we examined a variety of locative morphemes that can be glossed as UP. These included locative particles, topographic adverbs and locative verbal prefixes. Section 3 substantiated the claim that Cora locative prefixes have both static locative senses of UP as well as dynamic directional ones. In Section 4, we examined a variety of what the Cora construe as UP locations within the domain of the human body. These varieties are designated in the grammar and lexicon of Cora by body part terms, locative particles and verbal prefixes. The usages illustrate numerous points crucial to Cognitive Linguistics, exemplifying further the usefulness of constructs such as conceptual metaphors, domain shifts, subjectivity and objectivity, focus of attention, figure and ground, lexical extension, physical motion, fictive or abstract motion, prototypes, the role of the speaker's construal of situations that he/she chooses to discuss and the encyclopedic nature of meaning, among others. Section 5 took us into the domain of time and and lengthened our list of Cora conceptual metaphors that constitue a complex radial category (Lakoff 1987) which is organized around the conception of UP.

The analyses presented here are crucially based on concepts highlighted in works such as Croft's (1993) study of domains in relation to metaphor and metonymy, as well as more recent work by Barcelona (2000) and Panther and Radden (1999). Also crucial to these analyses is Talmy's (1996) and earlier work on Fictive Motion, Langacker's (1990, 1999) work on the Subjective/Objective distinction, Matsumoto's (1995, 1996) work on subjective motion and Langacker's (1993) work on reference point constructions. In addition, my theme here is reminiscent of the work by Lindner on UP and OUT in English (Lindner 1981), more recent work on the verticality dimension by Serra Borneto (Serra Borneto 1996) and on the PATH notion by Talmy (1972, 1975) and Radden (1996), among others. Finally, this paper represents one more effort to mesh both the approach of Langacker and that of Lakoff into a unified account of Cora spatial language (cf. also Casad 1997). In particular, I have pinpointed a family of Conceptual Metaphors in the UP category.

To begin, Conceptual metaphors from the domain of topography include UP is THE DIRECTION FROM WHICH THE RIVER FLOWS, UP is IN THE SLOPE HEADING TOWARDS THE TOP OF THE HILL, UP is AT THE TOP OF THE HILL and UP is AT THE VERY TIP TOP OF THE HILL. Conceptual metaphors from the domain of the human body include the following: UP is AT THE TOP OF THE HEAD; UP is IN THE SLOPE; UP is AT THE VERY TIP TOP; UP is AT THE HIGHEST POINT; UP is AT THE VERY TIP; UP is AT THE TOP LOOKING DOWN; UP is ON TOP and UP GOES ALL THE WAY ACROSS; UP is AN AREA IN THE MIDDLE OF THE SLOPE; UP is ON THE OUTER SURFACE OF THE HAND; UP is AT THE BELTLINE and UP is TRACING DOWNWARDS FROM THE MOUTH. Several of these relate to the more general the BODY IS A CONTAINER metaphor, including UP is COMING OUT INTO THE OPEN FROM UNDER THE SKIN.

References

Barcelona, Antonio (ed.) 2000. Metaphor and Metonymy at the Crossroads: A Cognitive Perspective. Berlin, New York: Mouton de Gruyter.

Casad, Eugene H. 1977. Location and direction in Cora discourse. – Anthropological Linguistics 19, 216–241.

Casad, Eugene H. 1982. Cora Locationals and Structured Imagery. La Jolla, CA.: University of California, San Diego Ph.D. Dissertation.

Casad, Eugene H. 1984. Cora. – Southern Uto-Aztecan Grammatical Sketches. Ed. by Ronald W. Langacker. Arlington, TX: The University of Texas and the Summer Institute of Linguistics. 152–459.

Casad, Eugene H. 1988. Conventionalization of Cora locationals. – Topics in Cognitive Grammar. Ed. by Brygida Rudzka-Ostyn. Amsterdam, Philadelphia: John Benjamins. 345–378.

Casad, Eugene H. 1993. Locations, paths and the Cora verb. – Conceptualizations in Natural Language Processing. Ed. by Richard A. Geiger, Brygida Rudzka-Ostyn. Berlin, New York, Amsterdam: Mouton de Gruyter. 593–645.

Casad, Eugene H. 1997 Many goofs: Exploiting a non-prototypical verb structure. – Proceedings of LP'96. Ed. by Bohumil Palek. Prague: the Charles University Press. 233–250.

Casad, Eugene H. 2000. Where do the senses of Cora va'a- come from? – Polysemy in Cognitive Linguistics. Ed. by Hubert W. Cuyckens

and B. Zawada. Papers from IVth ICLA Conference, Amsterdam, the Netherlands. Berlin, New York, Amsterdam: Mouton de Gruyter.

Casad, Eugene H.; Langacker, Ronald W. 1985. "Inside" and "Outside" – Cora Grammar. IJAL 51, 247–281.

Croft, William 1993. The role of domains in the interpretation of metaphors and metonymies. – Cognitive Linguistics 4, 335–70.

Fauconnier, Gilles; Sweetser, Eve (eds.) 1996. Spaces, Worlds and Grammar. Chicago and London: The University of Chicago Press.

Friedrich, Paul. 1969. On the Meaning of the Tarascan Suffixes of Space. Indiana University Publications in Anthropology and Linguistics, Memoir 23.

Friedrich, Paul 1971. The Tarascan Suffixes of Locative Space. Language Science Monographs, Vol. 9. Bloomington: Indiana University.

Gibbs, Raymond W. 1999. Speaking and thinking with metonymy. – Metonymy in Language and Thought. Ed. by Klaus-Uwe Panther, Günter Radden. Amsterdam, Philadelphia: John Benjamins. 61–76.

Grimes, Joseph E. 1964. Huichol Syntax. London, The Hague, Paris: Mouton.

Hawkins, Bruce W. 1984. The Semantics of English Spatial Prepositions. San Diego: UCSD Doctoral Dissertation.

Koch, Peter 1999. Frame and contiguity: On the cognitive bases of metonymy and certain types of word formation. – Metonymy in Language and Thought. Ed. by Klaus-Uwe Panther, Günter Radden. Amsterdam, Philadelphia: John Benjamins. 139–167.

Langacker, Ronald W. 1987. Foundations of Cognitive Grammar I: Theoretical Prerequisites. Stanford, CA.: Stanford University Press.

Langacker, Ronald W. 1990. Subjectification. – Cognitive Linguistics 1, 5–38.

Langacker, Ronald W. 1991. Concept, Image, and Symbol: The Cognitive Basis of Grammar. Cognitive Linguistics Research 1. Berlin, New York: Mouton de Gruyter.

Langacker, Ronald W. 1993. Reference point constructions. – Cognitive Linguistics 4, 1–38.

Langacker, Ronald W. 1999. Grammar and Sonceptualization. Cognitive Linguistics Research 14. Berlin, New York: Mouton de Gruyter.

Lakoff, George. 1987. Women, Fire and Dangerous Things. Chicago and London: The University of Chicago Press.

Lindner, Susan Jean 1981. English Verb-Particle Constructions with UP and OUT. La Jolla, CA.: University of California at San Diego: Ph. D. Dissertation.

Matlock, Teenie 2001. How Real is Fictive Motion? Ph.D. Dissertation, Department of Psychology, University of California at Santa Cruz.

Matsumoto, Yo 1995. Subjective motion and English and Japanese verbs. – Cognitive Linguistics 7, 183–226.

Matsumoto, Yo 1996. Subjective-change expressions in Japanese and their cognitive and linguistic bases. – Spaces, Worlds and Grammar. Ed by Gilles Fauconnier, Eve Sweetser. Chicago and London: The University of Chicago Press. 124–156.

McMahon, Ambrosio y María Aiton de McMahon (compiladores.) 1959. Vocabulario cora. Serie de vocabularios indígenas Mariano Silva y Aceves, 2. Instituto Ling□ístico de Verano en cooperación con la Dirección General de Asuntos Indígenas de la Secretaría de Educación Pública. México, D.F.

Panther, Klaus-Uwe; Radden, Günter (eds.) 1999. Metonymy in Language and Thought. Amsterdam, Philadelphia: John Benjamins.

Radden, Günter 1996. Motion metaphorized: the case of coming and going. – Cognitive Linguistics in the Redwoods: The Expansion of a New Paradigm in Linguistics. Ed. by Eugene H. Casad. Berlin, New York: Mouton de Gruyter. 423–458.

Radden, Günter; Kövecses, Zoltán 1999. Towards a theory of metonymy. – Metonymy in Language and Thought. Ed. by Klaus-Uwe Panther, Günter Radden. Amsterdam, Philadelphia: John Benjamins. 17–59.

Rubba, Johanna 1996. Alternate grounds in the interpretation of deictic expressions. – Spaces, Worlds and Grammar. Ed. by Gilles Fauconnier, Eve Sweetser. Chicago and London: The University of Chicago Press. 227–261.

Ruiz de Mendoza Ibáñez, Francisco José 2000. The role of mappings and domains in understanding metonymy. – Metaphor and Metonymy at the Crossroads: A Cognitive Perspective. Ed. by Antonio Barcelona. Berlin, New York: Mouton de Gruyter. 109–132.

Serra Borneto, Carlo 1996. Liegen and stehen in German: A study in horizontality and verticality. – Cognitive Linguistics in the Redwoods: The Expansion of a New Paradigm in Linguistics. Ed. by Eugene H. Casad. Berlin, New York: Mouton de Gruyter. 459–505.

Smith, Michael B. 1987. The Semantics of Dative and Accusative in German: An Investigation in Cognitive Grammar. La Jolla, CA.: University of California at San Diego: Ph. D. Dissertation.

Smith, Michael B. 1992. The Role of Image Schemas in German Grammar. – Leuvense Bijdragen 81:3, 385–410.

Smith, Michael B. 1993. Cases as conceptual categories: Evidence from German. – Conceptualizations in Natural Language Processing.

Ed. by Richard A. Geiger, Brygida Rudzka-Ostyn. Berlin, New York, Amsterdam: Mouton de Gruyter. 531–565.

Sweetser, Eve 1996. Mental spaces and the grammar of conditional sentences. – Spaces, Worlds and Grammar. Ed. by Gilles Fauconnier, Eve Sweetser. Chicago and London: The University of Chicago Press. 318–333.

Talmy, Leonard 1972. Semantic Structures in English and Atsugewi. Ph.D. Dissertation, University of California:Berkeley.

Talmy, Leonard 1975. Semantics and syntax of motion. – Syntax and Semantics, Vol. 4. Ed. by J. Kimball. New York: Academic Press. 181–237.

Talmy, Leonard 1996. Fictive motion in language and "ception." – Language and Space. Ed. by Paul Bloom, Mary A. Peterson, Lynn Nadel, and Merrill F. Garret.Cambridge, Mass.: MIT Press.

Vandeloise, Claude 1996. Touching: a minimal transmission of energy. – Cognitive Linguistics in the Redwoods: The Expansion of a New Paradigm in Linguistics. Ed. by Eugene H. Casad. Berlin, New York: Mouton de Gruyter. 541–566.

Когнитивные сценарии диалогических событий

Светлана Дикарева
Таврический национальный университет им. В. И. Вернадского
Крым, Украина

В статье показывается влияние идей Халдура Ыйма и Тартуской лингвистической школы на формирование концептуального аппарата когнитивной лингвистики, **сцены** и **сценарии** рассматриваются в качестве категорий, позволяющих описать динамическую модель диалога, взятого как последовательность речевых шагов, объединенных в диалогические события.

1. Когнитивная природа языковой коммуникации

Состояние лингвистики сегодня характеризуется стремительным изменением научных ценностей – от "языка в себе" мы перешли к изучению языка в широкой социальной и культурной среде. Объектом изучения лингвистов стало уже не слово, предложение или связный текст, а **дискурс** – семиотический процесс, погруженный в контекст жизни.

Изучение дискурса состоит в анализе **коммуникативных событий**, под которыми понимают языковые сообщения в совокупности с социокультурными, психологическими и другими экстралингвистическими факторами, включая среду общения, время и место. Отличительной чертой семантических и прагматических исследований дискурса является анализ **когнитивных структур** – знаний, мнений и установок участников коммуникации, концептов "веры" и "предубеждений", когнитивных сценариев осуществления коммуникативных событий.

В когнитивной лингвистике исследуются прямые и косвенные соответствия между когнитивными событиями и лингвистическими знаками, которые основываются на эмпирическом исследовании языковых знаков и документализации психологических явлений. При этом онтология языка рассматривается как многопричинный сценарий (*multicausal scenario*), обусловленный такими параметрами, как коммуникация, культурные и социальные знания, привычки речевого

сообщества, а также параметры взаимодействия биологических факторов и познания (Schulze 2000).

На этот факт уже в 70-е годы указывал Халдур Ыйм. "Понимание текста, – пишет Х. Ыйм, – требует не только понимания значений слов и грамматического строения. Оно требует, как правило, и определенных знаний о том фрагменте действительности, о котором что-то сообщается. Смысл текста для слушателя всегда представляет собой результат взаимодействия непосредственного содержания сообщения и знаний слушателя" (Ыйм 1978: 118).

В этом высказывании, на наш взгляд, содержатся 2 постулата, определяющих когнитивный подход к анализу языковой коммуникации.

Первый акцентирует внимание на **экстралингвистических знаниях**, необходимых для понимания текста: "Понимание текста ... требует, как правило, и определенных знаний о том фрагменте действительности, о котором что-то сообщается".

Второй подчеркивает **диалогичность** языковой коммуникации, в которой центральная роль отводится слушающему и его знаниям: "Смысл текста для слушателя всегда представляет собой результат взаимодействия непосредственного содержания сообщения и знаний слушателя".

2. Когнитивные модели диалога

В статье "Тайна диалога" В. З. Демьянков совершенно справедливо отмечает проблематичность и загадочность абсолютно всех моментов диалога: "Таинственно то, что собеседники не видят перед собой партитур, но интерпретируют свои собственные и чужие высказывания так, как если бы эти партитуры были. Их диалог обладает внутренней динамикой, драматическим ходом, напоминающим ход исполнения музыкального дуэта, где сходства и расхождения уже заданы композитором. Неужели же в каждом собеседнике уже заложена партитура, которой он следует более или менее корректно?" (Демьянков 1991: 14).

Для исследования "тайн диалога" в работах Халдура Ыйма разрабатывается теория "наивной модели употребления языка", интуитивная, наивная модель повседневной коммуникации: "Language in its generality is based on a naive conception, a naive

picture of the world that it is used to talk about and to cooperate upon" (Õim 1996: 195). При этом различаются наивная теория ментального действия (*folk theory of mental action*) и наивная теория взаимодействия (*folk theory of common interaction*), которая включает в себя функциональную модель партнера, способность предсказывать его коммуникативное поведение.

Народная мудрость отражена в фольклорных текстах. Фольклорные тексты типа пословиц представляют собой рекомендации, заготовленные опытом народа на тот случай, когда человек стоит перед выбором своего поведения, включая и речевое поведение.

Например, особенности русского речевого общения отражены в пословицах, зафиксированных в словаре Владимира Ивановича Даля. *Не всякую речь (правду) сказывай, Лучше недоговорить, чем переговорить, Во многословии не без пустословия, Не спеши языком, торопись делом, Говори с другими поменьше, а с собою побольше, Меньше бы говорил, больше бы слушал, Один говорит – красно, двое говорят – пестро* и т.п.

Эти пословицы представляют собой советы говорящему и определяют условия успешности диалогического взаимодействия, которые как бы построены на соединении Принципа Кооперации и Принципа Вежливости (ср. *Не всякую речь (правду) сказывай, Лучше недоговорить, чем переговорить* и др.)

Отдельную группу составляю пословицы, предназначенные для оценки не своего, а чужого речевого поведения, т.е. речевого поведения собеседника, например, *Говорит, как река льётся* (красноречие), *Красно говорит, а слушать нечего, Говорил день до вечера, а слушать нечего* (пустословие), *И невелика беседа, да честна* (искренность речи), *В глаза льстят, за глаза ругают* (неискренность речи).

По выражению В. И. Даля, пословица "не сочиняется, а вынуждается силою обстоятельств, как крик или возглас, невольно сорвавшийся с души" (Даль 1957: 18). И в этом плане каждую из приведенных пословиц можно рассматривать как рекомендацию, формирующую наивную модель мира, совет, который заготовлен на тот случай, когда человек выбирает стратегию своего речевого поведения.

В пословицах содержатся предписания относительно такого речевого поведения, которое приводит к коммуникативному успеху.

3. Сцены и сценарии диалогических событий

К числу наиболее полезных для моделирования диалога категорий относятся, как нам представляется, понятие **сцены** и **сценария**.

В 1979 году был опубликован перевод на русский язык популярной книги Марвина Минского "Фреймы для представления знаний" (Минский 1979). В этой работе выдвигается гипотеза о восприятии человеком зрительных образов в виде фреймов – структур знаний, анализируются ситуации распознавания образов на основе сцен и субфреймов, обсуждается модель понимания естественного языка как процедуры, включающей в себя эпизоды и сложные сценарии (Минский 1979).

Под фреймом обычно понимается структура знаний о типизированном объекте или ситуации, т.е. фрейм является декларативным способом представления знаний. В отличие от фрейма сценарий – процедурный способ представления стереотипного знания, которое описывается в виде алгоритма или инструкции, например, *сценарий торговой сделки, сценарий посещения ресторана.*

Понятие сцены связывается со зрительным восприятием объекта или ситуации, например, пространственно-временные координаты события, положение точки зрения субъекта по отношению к объекту восприятия, дейктический центр, фон и фокус и др (Талми 1999).

Как показывают когнитивные исследования лексики и грамматики разных языков, структурирование, которое осуществляется языковыми средствами, соответствуют структурированию в других когнитивных системах, таких, как зрительное восприятие, логическое мышление, кинетическое мышление. Главной функцией структурирования является концептуальная связность (*coherence*).

В языке эта функция имеет 2 формы реализации: **связность пространства** в пределах сцены и **связность во времени** в пределах сценария.

Разрозненные фрагменты связываются некоторым образом в виде единого гештальта в пределах предложения. Вместе с тем в ходе дискурса множество разных понятий сменяют друг друга, создавая потенциальную проблему потери связности. Однако при помощи системы лексических и грамматических средств связности в дискурсе поддерживается когнитивная непрерывность, а с течением времени формируется связный гештальт. Например, в русском языке такие мета-реплики, как *Да-да, Ну конечно! Да неужели? Вы подумайте!.* направляют иллокутивный поток, специфицируют логическую канву диалога, прорисовывают его риторический каркас (ср. понятие регулятивных высказываний в статье Pajusalu 1996).

4. Динамическая модель диалога

Как показано в предыдущем разделе, главной функцией структурирования дискурса является концептуальная связность (*coherence*), которая в диалоге имеет 2 формы реализации: **связность пространства** в пределах сцены и **связность во времени** в пределах сценария. Хореография диалога (динамическая модель) состоит в описании речевых шагов и ритма взаимодействия собеседников, в анализе того, каким образом происходит объединение шагов в диалогические события (Дикарева 1994). Для иллюстрации динамики диалога, приведем 3 фрагмента из "Балаганчика" Александра Блока.

Диалог 1. Он в голубом, она в розовом, маски – цвета одежд. Они вообразили себя в церкви и смотрят вверх в купола.

Она Милый, ты шепчешь – "склонись..."
Я, лицом опрокинута, в купол смотрю.
Он Я смотрю в непомерную высь –
Там, где купол вечернюю принял зарю.
Она Как вверху позолота ветха.
Как мерцают вверху образа.
Он Наша сонная повесть тиха.
Ты безгрешно закрыла глаза. <...>

Диалог 2. В середину танца врывается вторая пара влюбленных. Впереди – она в черной маске и вьющемся красном плаще. Позади – он – весь в черном, гибкий, в красной маске и черном плаще. Движения стремительны. Он гонится за ней, то настигая, то обгоняя ее. Вихрь плащей.

Он Оставь меня! Не мучь, не преследуй!
Участи темной мне не пророчь!
Ты торжествуешь свою победу!

Снимешь ли маску? Канешь ли в ночь?
Она Иди за мной! Настигни меня!
Я страстней и грустней невесты твоей!
Гибкой рукой обними меня!
Кубок мой темный до дна испей! <...>

Диалог 3. В среде танцующих обнаружилась третья пара влюбленных. Они сидят посреди сцены. Средневековье. Задумчиво склонившись, она следит за его движениями. – Он, весь в строгих линиях, большой и задумчивый, в картонном шлеме, – чертит перед ней на полу круг огромным деревянным мечом.
Он Вы понимаете пьесу, в которой мы играем не последнюю роль?
Она (как тихое и внятное эхо) Роль.
Он Вы знаете, что маски сделали нашу сегодняшнюю встречу чудесной?
Она Чудесной. <...>

Контрастность языковых структур, используемых в этих 3 диалогах, достаточно очевидна. Важно отметить также, что в каждом из этих диалогов ощущается гармония между движениями партнеров и языковыми выражениями, цветовой гаммой и стилем костюмов.

Общая дискурсивная характеристика

Диалог 1. Использованы приглушенные краски (голубой и розовый). Движения первой пары влюбленных спокойно-тихие ("Наша сонная повесть тиха"), они "вообразили себя в церкви и смотрят вверх в купола".

Диалог 2. Тревожно-контрастные краски (черная маска и вьющийся красный плащ, красная маска и черный плащ). Вторая пара влюбленных "врывается в середину танца", "движения стремительны", "он гонится за ней, то настигая, то обгоняя ее", "вихрь плащей".

Диалог 3. Строгие линии костюма, шлем. Третья пара сидит посреди сцены. "Задумчиво склонившись, она следит за его движениями", Он "чертит перед ней на полу круг огромным деревянным мечом".

Языковая характеристика

Диалог 1. Тихое, "приглушенное" диалогическое взаимодействие, со-переживание, со-любование, обмен речевыми шагами со слабым коммуникативным стимулом (ассертивы, экспессивы,"слабые" директивы).

Диалог 2. Диалог-конфликт, встречные речевые шаги (в основном, директивы) с сильным коммуникативным стимулом (повеление – отказ, обвинение – несогласие, категоричное требование – согласие).

Диалог 3. Двухголосный монолог, встречных речевых шагов нет.

Таким образом, в контексте общего дискурса движения партнеров (неречевое поведение) находятся в соответствии с коммуникативным диалогическим движением (речевое поведение). Реплики характеризуются семантической и прагматической сочетаемостью, образуя связный гештальт.

5. Заключительное замечание

Лингвистическая проблематика, связанная с изучением диалога, обширна и многоаспекта. В нашем эмоциональном мире поиск путей эффективного диалога все более актуален. Когнитивные исследования диалога направлены на поиск универсальных структур (стереотипов поведения, сценариев) и способов коммуникативной реализации этих структур. Однако подлинный диалог – это всегда общение с неизвестным, тайна, заключенная в противостоянии Я и Ты.

Литература

Вежбицкая, Анна 1996. Язык. Культура. Познание. Отв. ред. М. А. Кронгауз. М.: Русские словари.

Герасимова, И. А. 2000. Танец: эволюция кинестезического мышления. – Эволюция. Язык. Познание. Под общей ред. И. П. Меркулова. М.: Языки русской культуры. 84–112.

Демьянков, В. З. 1991. Тайна диалога. – Диалог: теоретические проблемы и методы исследования. М. : АН СССР. 10–42.

Дикарева С. С. 1994. Прагматические условия диалога. Симферополь: Изд-во Симферопольского университета.

Дикарева С. С. 2001. Сцены и сценарии Web-коммуникации. – Доклады конференции CSC'2001 Когнитивные сценарии языковой коммуникации. 24–28 сентября, Украина, Крым, Партенит. 25–28.

Koit, Mare; Õim, Haldur 1998. The concept of communicative strategies: A Theoretical model and an implementation. – Proceedings of

Cognitive Strategies for Language Communication. Eds. S. Dikareva, V. Ronginsky, L. Bessonova. 21–25 September, Crimea, Ukraine. 14–21.

Минский М. 1979. Фреймы для представления знаний. Пер. с англ. О. Н. Гринбаума. Под ред. Ф. М. Кулакова. М.: Энергия.

Pajusalu, Renate 1996. Regulative utterances in Estonian literary dialogues and radio interviews. – Estonian in the Changing World. Ed. by H. Õim. Tartu: University of Tartu. 133–162.

Schulze W. 2000. Cognitive Linguistics Meets Typology. The Architecture of a "Grammar of Scenes and Scenarios" http://www.lrz-muenchen.de/~wschulze/cog_typ.htm.

Талми Л. 1999. Отношение грамматики к познанию. – Вестник Московского университета. Сер. 9. Филология, № 1, 4 и 6.

Ыйм, Х. 1978. Язык, значение и знание. – Проблемы моделирования языковой интеракции. Ч. II, Тарту: ТГУ. 117–129.

Õim, Haldur 1996. The need for a theory of folk theories in cognitive semantics: A Review and a discussion. – Estonian in the Changing World. Ed. by H.Õim. Tartu: University of Tartu. 193–210.

Õim, Haldur 1996. Naïve theories and communicative competence: Reasoning in communication. – Estonian in the Changing World. Ed. by H. Õim. Tartu: University of Tartu. 211–231.

The dialogue engineering life-cycle[1]

Laila Dybkjær and Niels Ole Bernsen
Natural Interactive Systems Laboratory, Odense, Denmark

Abstract

Software engineering life-cycle models have been around for about 30 years. Various models have been proposed, refined and adapted to the needs of different sub-areas of systems development. Dialogue engineering is a fairly new sub-area of software engineering, commercial spoken language dialogue systems (SLDSs) having been in the market place for only about a decade. While an iterative software engineering life-cycle model may apply to SLDS development at a general level, the model is insufficient at lower levels of detail. To better support SLDSs developers in industry and research, there is a need for a model which specialises software engineering life-cycle modelling to the development and evaluation processes which are specific to SLDSs and their components. This paper describes and illustrates dialogue engineering life-cycle modelling and how to tailor it to the field of SLDSs.

1. Introduction

Even if software engineering life-cycle modelling is not new, a continuing variety of general models having appeared over the last 30 years, the field remains of major interest to industry and research. For industry, a life-cycle model is an important tool to efficiently control the development and evaluation process and ensure quality. In research, professional engineering practice is becoming increasingly important as researchers are moving from component exploration to advanced systems development. In several sub-areas of systems development, general models have been refined and adapted to better suit the particular needs of those areas and thus provide better support. An example is safety-critical systems where evaluation is emphasised much more than in many other areas of software development. One

[1] We gratefully acknowledge valuable comments from Hans Dybkjær on a draft version of this paper.

reason why life-cycle modelling is a research topic in itself is that the adaption process has not yet happened in all sub-areas. Dialogue engineering is one such sub-area in need of a model which specialises software engineering life-cycle modelling to the development and evaluation processes that are specific to SLDSs and their components.

This paper proposes a dialogue engineering life-cycle model which is tailored to SLDSs development and evaluation. Section 2 reviews general software engineering life-cycle models and briefly describes the main models or their main representatives, i.e. the waterfall model, the iterative model, prototype development, the spiral model, and extreme programming. Section 3 presents the general DISC dialogue engineering model which specialises iterative life-cycle modelling to the special needs of dialogue engineering. Section 4 describes and illustrates the DISC dialogue engineering model in detail, including life-cycle issues, integration of so-called 'grid' issues, addition of evaluation criteria, and description of the resulting dialogue engineering development and evaluation life-cycle.

2. Software engineering life-cycle models

The first software engineering life-cycle model was introduced about 30 years ago. Before that time developers used what has been called the *tunnel development model* (Muller 1997). Projects were initiated and, some day later, maybe, a system resulted. There was no model of the development process itself. But there was a clear need to get some structure to the process, e.g. to better understand software development and justify the resources spent on development.

This first model (Royce 1970) became known as the *waterfall model* because it basically describes the development process as a linear execution of four steps followed by a maintenance phase when the software has been put into operation, cf. Figure 1. The five steps are known by different names. We shall refer to them as analysis, design, construction, integration, and maintenance, respectively. The model was well received because it provides structure and visibility to the development process.

The identification of major activities and the shifts in overall focus as development proceeds was the prime contribution of the waterfall model and is still valid today. However, the model is too simplistic and rigid. The strictly separated phases lead to costly han-

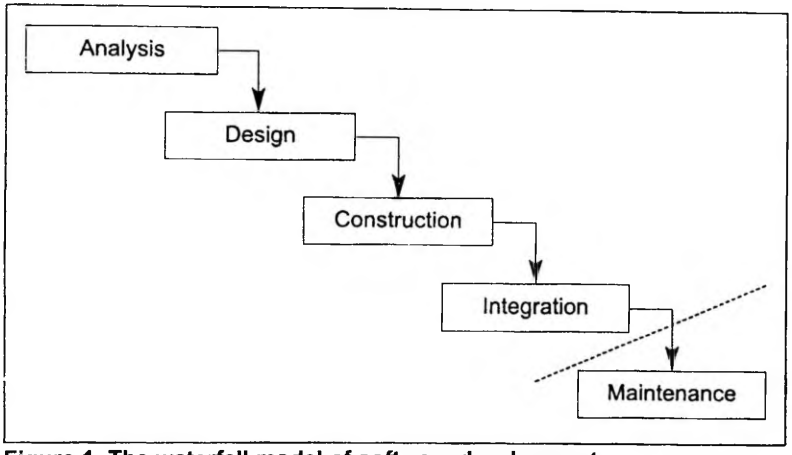


Figure 1. The waterfall model of software development

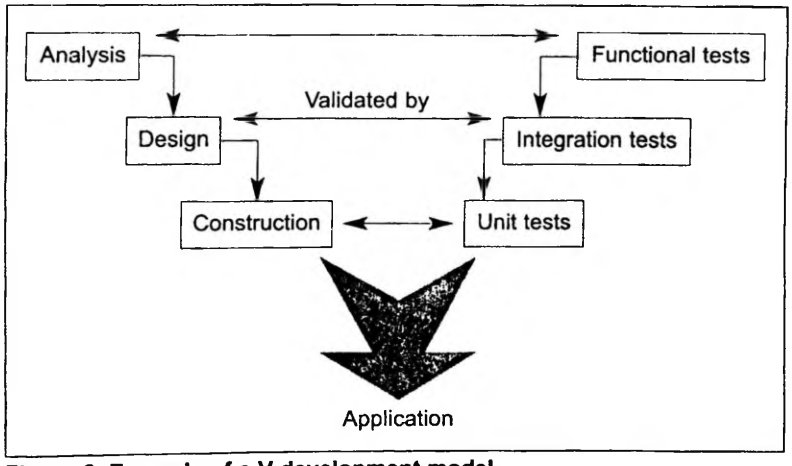


Figure 2. Example of a V development model
Tests are developed in parallel with the software (Muller 1997)

dling of changes and poorly reflect human cognitive and social processeses.

The first problem identified in the model was the assumption that a system or component is only being evaluated in the integration phase, i.e. towards the very end of the development process. Until then, only documents are available for validation. This means that even serious problems may only be discovered at a late stage when the cost of making changes is high. A second problem is that the

model assumes that a particular phase is being fully completed before the next phase is initiated. This means that, e.g., all requirements must be known and made explicit in the analysis phase before proceeding to the design phase, which is often in conflict with the reality of software engineering practice.

One way to address the evaluation problem is to introduce elements of evaluation at earlier stages in the life-cycle. Later versions of the waterfall model are thus sometimes presented in the form of the letter V, indicating that the development of test plans and test data is done synchronously with software development. Even if, in the waterfall model, actual evaluation is only carried out once the software is ready, the *V-model* is likely to improve developers' understanding of the life cycle stages and the software being developed by encouraging them to decide on detailed evaluation plans and data at each stage. Figure 2 shows an example of a V model in which functional tests are specified during analysis, integration tests during design, and unit tests during the construction phase.

However, the V-model remains rigidly sequential and puts too little focus on the human processes. Several alternative, general life-cycle models have been proposed in the attempt to overcome the problems of the waterfall model and meet the needs of complex systems development, see e.g. Pressman 1997; Sommerville 1992 or later for an overview. Some of these models have not been used much in practice, such as the formal transformation model (Pressman 1997; Sommerville 1992). Other models appear under different names but are, in fact, quite similar and may be viewed as variations on the same model, e.g. exploratory programming and prototyping. In this brief overview we shall only mention what we consider the main models or their main representatives, i.e. the iterative model, prototype development, the spiral model, and extreme programming. The overview focuses on process structure. Other important aspects that have been identified in the 1980s and 1990s are not included, such as user involvement, organisational and social implications of introducing new systems, teamwork, or design rationale representations.

A frequently used model is the *iterative model* which has many variations depending on, e.g., project size and domain complexity (Muller 1997). The iterative model incorporates the idea of iterating one or more of the first four phases of the waterfall model, cf. Figure 3. Moreover, phases may overlap and sometimes a breakdown into

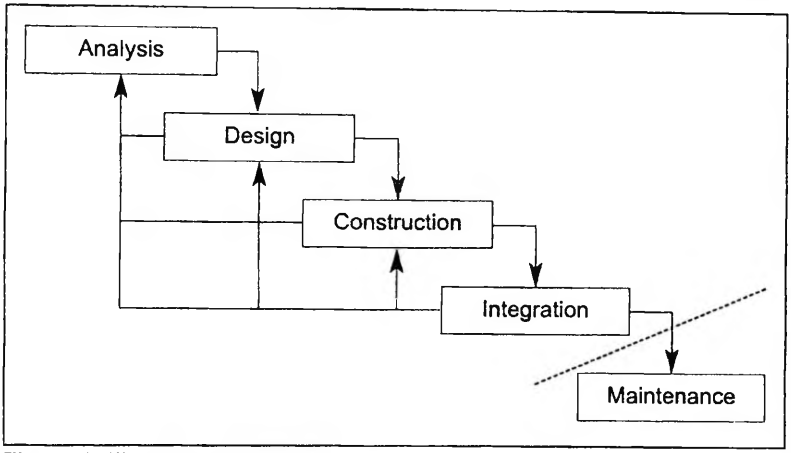


Figure 3. The iterative model

more detailed phases is introduced. Today, the trend is towards shorter cycles, focusing on prototypes. In fact, both prototype development, the spiral model and extreme programming are iterative models and may be viewed as further developments of the first iterative software engineering life-cycle models.

Prototype development has been used widely since the heyday of AI 20 years ago. Prototyping is a useful method when it is not clear in advance what the system should be like, as may be the case with, e.g., systems exploring new forms of interaction. Prototypes are also well suited for rapidly constructing an early version of the intended system. Prototyping requires good planning and process control. One risk is that programmers do not want to discard code from the prototype even if the final system could be improved if (part of) the prototype code were re-implemented.

The *spiral model*, cf. Figure 4, was introduced in the mid-1980s (e.g. Boehm 1988). "It has two main distinguishing features. One is a cyclic approach for incrementally growing a system's degree of definition and implementation while decreasing its degree of risk. The other is a set of anchor point milestones for ensuring stakeholder commitment to feasible and mutually satisfactory system solutions" (Boehm and Hansen, http://www.stsc.hill.af.mil/crosstalk/2001/may/-boehm.as). Roughly speaking, the model combines prototype development with the use of the waterfall model for each step. The spiral model is intended to help manage risk. The system is defined and

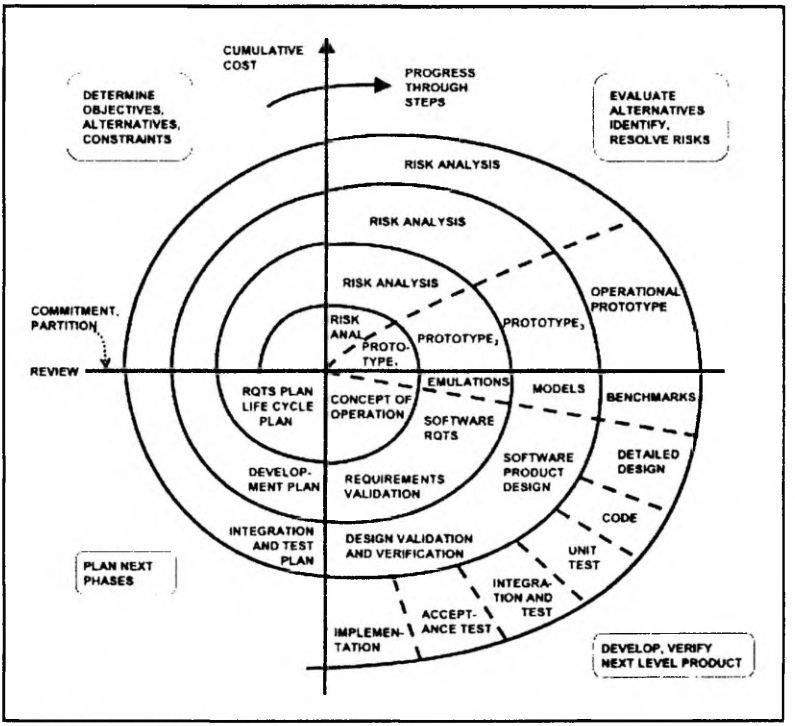


Figure 4. Spiral development diagram
(http://www.stsc.hill.af.mil/crosstalk/2001/may/boehm.asp)

developed stepwise, corresponding to cycles in the spiral. Each cycle is in principle ended by answering the question: "Should we continue?" either negatively or positively depending on the risks involved. Each new cycle starts by an analysis to determine the best way in which to deal with the current cycle. Basically, each cycle in the spiral includes the following four steps: determine objectives, analyse and evaluate, develop and test, and plan next phase. The spiral model has strongly influenced later systems development methods, such as DSDM (Dynamic Systems Development Method) (www.dsdm.org).

Extreme programming is a fairly recent model. It basically prescribes to do everything (analysis, design, etc.) all the time, i.e. using very rapid iterations. Programming is done pairwise, i.e. programmers sit together in pairs when writing code, providing real-time analysis and code review. A limited-functionality prototype is developed very early in the project. For instance, the prototype may be able to handle

a single key functionality issue only. Every day, every week, or whatever may be the turn-over time, an improved version of the prototype is made ready. In this way, the prototype is eventually extended to have the functionality required of the final system. Thus, the prototype keeps changing rapidly, and components are being updated independently (e.g. Beck 1999a, 1999b).

The two major problems in the waterfall model, i.e. (too) late evaluation and strict phase seriality, cf. above, are both addressed in the iterative model and its variations discussed above. By being iterative, those models by definition overcome the problem that a particular phase is being fully completed before the next phase is initiated. Moreover, by consequence of their iterative nature, the models share the idea that system or component evaluation is not only done once towards the end of the development process but is done throughout development. The main difference among the models as regards evaluation would seem to be the time it takes to make an iteration, with extreme programming assuming the faster turn-around time.

Software engineering life-cycle issues have remained an important research area ever since the first model was presented, and their scope includes both custom-made and off-the-shelf software. Using an adequate life-cycle model is of key importance to efficient and successful software development. Nevertheless, too little attention is often paid to this fact by development teams. This may partly be due to ignorance but an other important reason is probably that detailed software life-cycle models are still missing in many areas or are still at the research stage. An area in point is that of spoken language dialogue systems (SLDSs). A specialised dialogue engineering life cycle model is discussed in more detail in the following sections.

3. The DISC dialogue engineering model

The first, very simple commercial SLDSs were introduced only about 10 years ago. As increasingly advanced and complex SLDSs are being developed in research and industry, the need has emerged for a detailed dialogue engineering life-cycle model.

At a general level, the iterative life-cycle model in Figure 3 also applies to SLDSs development. However, a model which specialises in the development and evaluation processes which are specific to SLDSs and their components could provide far better support to SLDS developers. Also, there is a need for a model which not only focuses on

software development but also on its evaluation, on the development and evaluation of documentation, and on the continuous evaluation of factors which may influence the development and evaluation process at any time throughout the life-cycle.

In the European DISC project (www.disc2.dk) on best practice in the development and evaluation of SLDSs, we developed a draft dialogue engineering model for SLDSs and components. The model includes a life-cycle model, a so-called 'grid' (see below), and a set of evaluation criteria. The draft model, and in particular its life-cycle part, has been further developed and refined after the DISC project ended in early 2000. The resulting dialogue engineering model is described below using dialogue management for illustration.

The current version of the DISC dialogue engineering model is based on (i) first ideas in (Bernsen *et al* 1998), (ii) the DISC approach to dialogue engineering, (iii) analysis by the DISC partners of the actual life-cycles and properties ('grid' issues, see below) of a large number of different SLDSs and components, and (iv) the draft DISC model with subsequent elaborations. All DISC results are available at www.disc2.dk.

In the DISC approach, an SLDS has six aspects: speech recognition, speech generation, natural language understanding and generation, dialogue management, human factors, and system integration. In simple systems, the natural language understanding and generation aspect may be non-existent but the five other aspects probably must be present for the system to be an SLDS at all (even low quality human factors are human factors). From the point of view of best practice, an SLDS should be the result of (a) correct choices among the available options, technological and otherwise, within each aspect and (b) correct development (including evaluation) practice.

Based on analysis of 25 existing SLDSs and components (Bernsen *et al* 1999), DISC has developed a 'grid' best practice guide per aspect. Each grid defines a space of aspect-specific issues which the developer must, or may have to, address, primarily depending on the complexity of the SLDS or component to be developed. When developing a dialogue manager, for instance, the developer should consider if the dialogue manager should provide some form of graceful degradation in order to handle cases of repeated user-system communication error. For each issue, the currently available options are laid out in the grid together with the pros and cons for choosing a par-

ticular option (cf. (a) above). For instance, feedback to the user is an important issue in dialogue management. The grid offers two options, i.e. process feedback and information feedback. Process feedback informs the user that the system is still working even if it takes some time to, e.g., query the database. Information feedback allows the user to verify that spoken input has been correctly understood by the system. Both types of feedback can be provided in several different ways which are presented as options together with their pros and cons.

In addition to the aspect-specific grids and again based on the analysed exemplars, DISC has developed a life-cycle best practice guide per aspect (cf. (b) above). The current DISC dialogue engineering life-cycle model has two interrelated levels. At the overall level, and assuming a general iterative software engineering life-cycle model, the model is used in developing and evaluating entire SLDSs as well as individual SLDS components. At the more detailed level, additional support is provided for addressing a particular aspect of an SLDS by specialising the life-cycle to this aspect, i.e. by taking into account the particular grid issues and evaluation criteria which are relevant to that aspect, such as dialogue manager development.

4. Life-cycle issues, grid issues and evaluation criteria

This section describes the current status of the DISC dialogue engineering model. We take the life-cycle issues as point of departure, integrate grid issues, add evaluation criteria, and describe the resulting dialogue engineering development and evaluation life-cycle.

4.1. Life-cycle issues

The dialogue engineering life-cycle may be depicted following the V-model as having two legs. One leg addresses the development phases. The grid issues must be addressed in these phases. The second leg shows what is being evaluated. Particular evaluation criteria, which depend on the aspect considered, are derived from the chosen grid issues and used at specified points during evaluation. Evaluation is an integral part of the development process and is carried out throughout the life-cycle. What is being evaluated depends on the particular life-cycle phase.

Figure 5 shows the five major development phases (the maintenance phase is not shown) which are typical to SLDSs, and the major

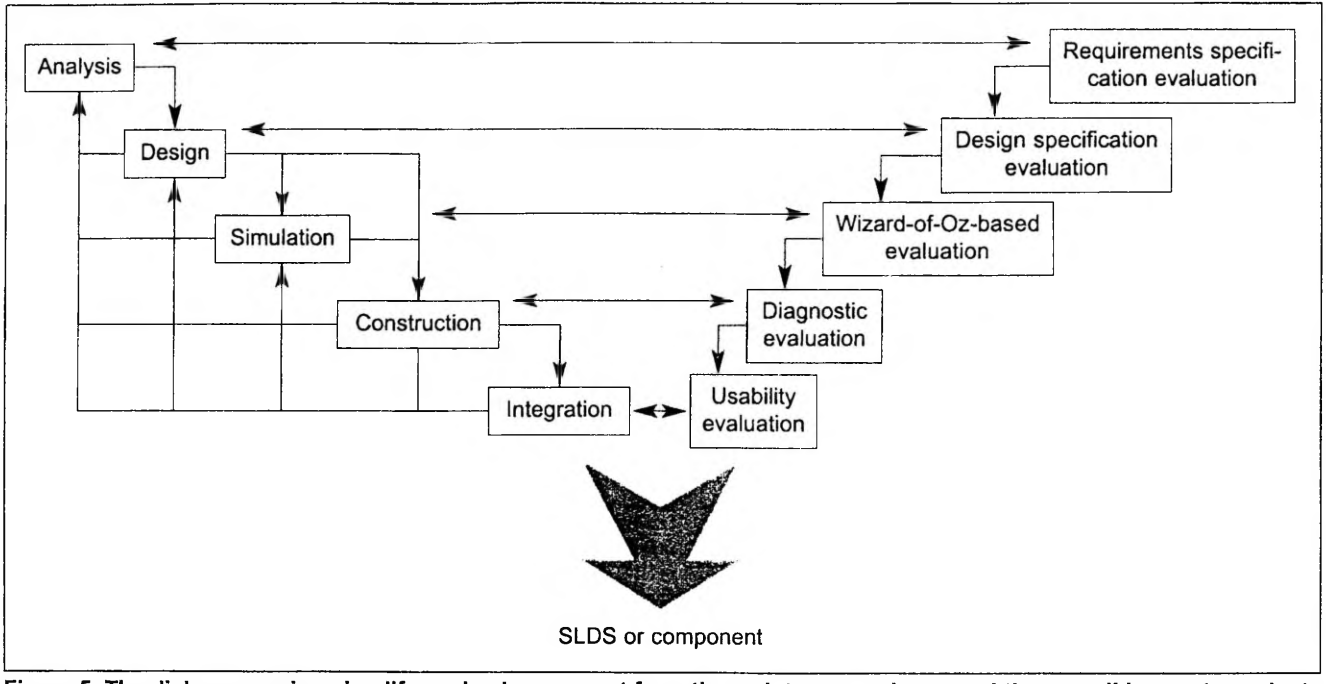


Figure 5. The dialogue engineering life-cycle phases apart from the maintenance phase, and the overall issues to evaluate

Life-cycle phase	System / component evaluation	Document evaluation		Management factors evaluation
Analysis	Requirements specification evaluation	Requirements specification documentation	Development and evaluation process description	Development time Personnel resources Mastery of the development and evaluation process Problems during development and evaluation Management quality
Design	Design specification evaluation	Design specification documentation		
Simulation (optional)	Wizard-of-Oz-based evaluation	Wizard-of-Oz documentation		
Construction	Diagnostic evaluation	System/component documentation		
Integration	Usability evaluation	User manual and guide		

Figure 6. The dialogue engineering life-cycle model (apart from maintenance)

evaluation activities in focus during those phases. It should be noted that simulation is frequently used in the development of (advanced) SLDSs. This is why simulation is mentioned as a separate (optional) phase. Unlike the V-model in Figure 2, Figure 5 shows what to evaluate in a particular phase but does not show when a particular type of evaluation is planned.

Figure 5 focuses on software development phases and evaluation. Documentation is not considered even though this is an important part of the development and evaluation process. Also, the figure does not show or relate to parameters, such as resources, skills and unexpected problems which may strongly influence development and evaluation at any point in the life-cycle, and which should therefore constantly be monitored.

Figure 6 shows the dialogue engineering life-cycle model, including life-cycle phases, system/component evaluation, document evaluation, and monitoring of other key factors. Documentation is assumed to develop along with the system or component during the individual life-cycle phases as described in more detail below.

A number of life-cycle issues pertain to one or more of the life-cycle phases in the left-hand column of Figure 6. Similarly, what is included in the development process must also be evaluated and thus reflected in the evaluation criteria to be used (see below). Figure 7

Figure 7. Issues addressed by the dialogue engineering life-cycle model

General

- OVERALL DESIGN GOAL(S) define the general objective(s) of the development process.

Constraints

- HARDWARE CONSTRAINTS refer to any a priori constraints on the hardware to be used in the design process.
- SOFTWARE CONSTRAINTS refer to any a priori constraints as regards use of software (e.g. development platform or pre-existing modules or components).
- CUSTOMER CONSTRAINTS are constraints imposed on the system/component by the customer, if any. Customer constraints may include hardware or software constraints.
- ORGANISATIONAL ASPECTS address if and how the system/component will have to fit into some organisation or other.
- OTHER CONSTRAINTS than hardware, software and customer constraints may be imposed on the system/component and influence the development process.

Ideas and preferences

- DESIGN IDEAS are ideas which developers may wish to realise in the development process, such as to develop a generic dialogue manager.
- DEVELOPER PREFERENCES may impose constraints on system/component development which were not dictated from elsewhere. Developer preferences may relate to many different issues, e.g. preferred programming language.

Criteria

- REALISM CRITERIA describe if the system/ component will meet real user needs, will meet them better, in some sense to be explained (cheaper, more efficiently, faster, other), than known alternatives, or if the dialogue manager is "just" meant for exploring specific possibilities to be explained, or if there are other realism criteria which should then be explained.
- FUNCTIONALITY CRITERIA concern which functionalities the system / component should have.
- USABILITY CRITERIA concern the usability aspects of the system/component. Usability criteria may be seen both from a system developer's and from an end-user's point of view. Usually the main focus will be on usability for end-users.

CONTINUATION. Figure 7.

Documentation and references

- REQUIREMENTS SPECIFICATION DOCUMENTATION addresses how the requirements specification is documented.
- DESIGN SPECIFICATION DOCUMENTATION addresses how the design specification is documented.
- WIZARD-OF-OZ DOCUMENTATION addresses the documentation of the (partially) simulated system/component.
- SYSTEM DOCUMENTATION addresses the documentation of the implemented system/component.
- USER MANUAL AND GUIDE concerns the description of the system/component to its users.
- DEVELOPMENT AND EVALUATION PROCESS DESCRIPTION (including references) is used to capture specifications, choices and decisions made, their justifications and their consequences and results. The description should include, e.g., information on how requirements were established, how the system/component was developed and evaluated, implementation issues, and tests done on the system/component.

Management factors

- DEVELOPMENT TIME addresses the time planned for development and the actual time used.
- PERSONNEL RESOURCES address the amount of person months allocated to, and actually spent on, the development of the system/component.
- MASTERY OF THE DEVELOPMENT AND EVALUATION PROCESS addresses the question of which parts of the process the development team has sufficient mastery in advance and of which parts they do not have such mastery.
- PROBLEMS DURING DEVELOPMENT AND EVALUATION should be described and handled.
- MANAGEMENT QUALITY concerns the way in which the project is managed.

Post-development issues

- PORTABILITY addresses ease of portability.
- MODIFICATIONS concern what is required if the system/component is to be modified.
- ADDITIONS, CUSTOMISATION addresses how additions to, and customisation of, the system/component can and should be carried out, e.g. if there is a strategy for resource updates and if there is a tool to enforce that the optimal sequence of update steps is followed.

shows a set of life-cycle issues which are general and which are assumed by the dialogue engineering model presented here. The issues all relate to the software being developed (general, constraints, ideas and preferences, and criteria), the documentation relating to the software (documentation and references), or the management factors which influence the development process (management factors). The last group of issues in Figure 7 (post-development issues) relate to the maintenance phase and will not be discussed any further. Evaluation criteria will be discussed later.

Some life-cycle issues are important to several development phases whereas others are restricted to a single phase as explained in more detail below. For each life-cycle issue a number of details must be considered, including the phase(s) to which it primarily belongs, evaluation methods to apply, the influence of the issue on the SLDS or component, the importance of taking it into account and possible effects of not addressing it, particular difficulties relating to the issue, and people with major influence on the issue. Space does not allow us to address all of these details. Only the two first-mentioned points will be discussed below.

We distinguish three categories of stakeholder in development: the procurer or customer, the users, and the developers. Most SLDSs are custom-made software (although several of their components are not) and therefore the customer also plays an important role. In case of off-the-shelf software, the development process roughly remains the same but the input from the customer is not available, so developers have to manage without it, contribute the information themselves or obtain it from representative users.

4.2. Grid issues specialising in life-cycle issues

As mentioned in Section 3, we shall use the dialogue management aspect for grid illustration. Dialogue management grid issues may be structured into the following categories covering the issues listed in parentheses (Bernsen and Dybkjær 2000a):

- goal;
- system varieties (multimodal systems including speech, multilingual systems, and multi-task, multi-user systems);
- system speech and language (are the speech and language layers OK, do the speech and language layers need support from the dialogue manager, and real-time requirements);

- getting the user's meaning (task complexity, controlling user input, who should have the initiative, input prediction/prior focus, sub-task identification, and advanced linguistic processing);
- communication (domain communication, meta-communication, other forms of communication such as greetings, expressions of meaning, i.e. how is the meaning of what has to be conveyed to the user expressed by the dialogue manager, error loops and graceful degradation, feedback, and closing the dialogue); and
- history, users, implementation (dialogue histories, novice and expert users, user groups, other relevant user properties, and implementation issues).

Jointly with the system/component-specific grid issues, the first four categories of life-cycle issues in Figure 7 (general, constraints, ideas and preferences, and criteria) determine the software that is being developed. The life-cycle issues are general but, taken together with the grid issues, they aim at a particular aspect of an SLDS. Let us look at some examples. The dialogue management grid issue categories from the above list are referenced in parentheses. The grid issues are described in detail at http://www.disc2.dk/slds/dm/DMgrid.html and more comprehensively in (Bernsen, Dybkjær 2000a).

The life-cycle issue of customer constraints covers constraints imposed on the system/component by the customer. Customer constraints will typically include grid issues. Many different customer constraints may be imposed on dialogue manager development. For example, the customer may want the system-provided service backed up by human operator fallback during the company's opening hours.

Design ideas are ideas which the developers want to realise in the development process. For dialogue management, examples of design ideas involving particular grid issues could be exploring internal dialogue manager modularity (grid: implementation), investigating different ways in which the dialogue manager can support natural language understanding (system speech and language), exploring system co-operativity in dialogue (communication), experimenting with different dialogue control strategies (getting the users meaning), or exploring ways in which to exploit contextual information to improve the dialogue (history, users, implementation).

Usability criteria may be viewed both from a system developer's and from an end-user's point of view. Usually the main focus will be on usability for end-users. End-users will not experience the dialogue

manager as a stand-alone component but only as part of an entire system. To a large extent, however, it is the dialogue manager which is responsible for how satisfactory the SLDS is to use. Grid issues which may generate usability criteria include, e.g., natural, flexible and robust dialogue, which has to do with initiative, feedback, error loops and graceful degradation, histories, user properties, etc. (getting the users meaning, communication, and history, users and implementation), sufficient meta-communicative facilities (communication), and users' backgrounds (history, users, implementation).

The first four categories of life-cycle issues in Figure 7 and the grid issues which become subsumed by them, are typically in focus already in the analysis phase or, at the latest, in the design phase, depending on whether they are brought in as requirements or as part of the design specification. Some grid issues may be involved in the analysis phase as part of a life-cycle issue whereas others are added to that life-cycle issue in the design phase. For instance, meta-communication may be a grid issue brought in as a usability constraint in the requirements specification, i.e. in the analysis phase, whereas, e.g., error loops and graceful degradation may be added as another usability constraint in the design phase as part of the design specification. Overall goals and customer constraints are usually specified in the analysis phase whereas developer preferences typically belong to the design phase. The remaining life-cycle issues (first four categories) may be specified during analysis and/or during design, depending on their importance in the actual development process. All the life-cycle issues from the first four categories continue to play a role not only in analysis and/or design but also in later phases, i.e. in simulation, construction, and integration. These later phases serve to progress the development of a system or component in accordance with the requirements and design specifications developed during analysis and design.

The documentation issues, on the other hand, are related to a particular phase, cf. Figure 6, except for the last one which is an accumulating document throughout development across phases and iterations.

The management issues in Figures 6 and 7 do not belong to any specific phase but must be monitored throughout development. Development time and personnel resources may appear as, e.g., customer constraints and be part of the requirements specification. Mas-

tery of the development and evaluation process may be a factor which influences the resources set aside in terms of time and personnel and which therefore also influences the requirements specification in an indirect way. Throughout the development process, it is important to monitor that the resource budget (time and personnel) holds and, if not, take the necessary actions immediately. Problems during development and evaluation may influence the resource budget. It is therefore important to keep an eye on such problems and consider solutions and consequences as soon as they are spotted. Management quality is determined by how well monitoring and management of the development and evaluation process is done, whether the appropriate actions are taken, and whether this is done in a timely fashion.

4.3. Evaluation criteria derived from grid issues

Since evaluation is an integral part of the development process, we also need to take a look at evaluation criteria and the evaluation process. Evaluation criteria are aspect-specific in the same way as grid issues are aspect-specific.

An interesting observation made in DISC is that, based on the grid issues, it is possible to derive a set of evaluation criteria per aspect. Suppose that, for instance, the dialogue management grid includes 24 issues for consideration by dialogue manager developers, such as which types of dialogue histories to include in a particular application. If the SLDS to be developed is a relatively simple one, not all of the 24 issues are likely to be relevant, so the developers select options within, say, 14 of the issues and ignore the remaining issues because these are relevant only to more sophisticated dialogue managers than presently needed. In this case, the developers must apply evaluation criteria to 14 chosen dialogue manager options in order to do a complete evaluation of the dialogue manager aspect of the application. Process and results of generating a complete set of evaluation criteria for human factors in SLDSs are presented in (Dybkjær, Bernsen 2000).

Knowing *what* to evaluate, is not enough, however. *How* to evaluate is just as important. To follow best development practice, developers have to evaluate their solution with respect to a chosen option at the right time(s) and in the right way(s). Thus, how to evaluate is a matter of applying a particular evaluation criterion correctly at the right stages during the development life-cycle.

Given that the developer knows, per SLDS aspect and for the particular application at hand, what to evaluate, such as how well the dialogue manager handles error loops and graceful degradation, focus can shift to how to do the evaluation. In DISC, we have iteratively developed an evaluation template to support the 'how' of evaluation. The template is a model of what the developer needs to know in order to apply an evaluation criterion to a particular property of an SLDS or component, such as the histories used by the dialogue manager. This knowledge is specified by the template's ten entries including what is being evaluated, system part evaluated, type of evaluation, method(s) of evaluation, symptoms to look for, life-cycle phase(s), importance of evaluation, difficulty of evaluation, cost of evaluation, and tools. Details on the evaluation template can be found in (Bernsen, Dybkjær 2000b). Examples of filled templates can be found at the DISC website, e.g. http://www.disc2.dk/slds/dm/DMevaldetail.html.

4.4. Evaluation integrated into the development life-cycle

This section briefly discusses evaluation and evaluation methods as part of the development life-cycle. It should be noted that the evaluation criteria described above which were derived from the grid issues, only form part of what has to be evaluated. The grid-derived evaluation criteria only relate to the software and not to, e.g., documentation or the development process itself.

In the *analysis phase,* the overall goals of the system/ component as well as the most important constraints, ideas and preferences, and criteria, cf. Figure 7, are established and described in a requirements specification. This is done in collaboration with the procurer (if any) and the system end-users. An important activity is to specify the evaluation criteria which the final system must satisfy to be accepted by, e.g., the procurer. A first evaluation is made of the feasibility of the requirements specification given constraints and resources. The requirements specification and any additional documentation is produced and evaluated as to sufficiency and clarity.

During the *design phase,* a design specification based on the requirements specification and other sources is worked out, adding additional constraints, ideas, preferences and criteria, and detailing these to a level sufficient to form the basis of simulation or construction. In parallel and closely interacting with this activity, a design analysis evaluation takes place. Design and design analysis evaluation

involve using experience and common sense, thinking hard when exploring the design space, doing walkthroughs of models, comparing with similar systems, browsing the literature, applying existing theory, guidelines and design support tools, if any, involving experts and future users, etc. The completeness of the design specification may be judged by checking whether all relevant entries in the DISC "grid(s)" have been considered. Design analysis evaluation also consists in checking whether the design goals and constraints are sound, non-contradictory and feasible given the resources available. The documentation produced in this phase includes the design specification and any additional documents which have been used or produced. For instance, literature should be referenced and walkthroughs and their analysis should be documented. Documentation evaluation consists in judging whether the design specification is appropriately represented and whether all relevant documents have been included.

The *simulation phase* is an optional part of SLDSs development. Typically, simulations are made using the Wizard-of-Oz (WOZ) simulation methodology in which the system or some of its components as specified during design are being simulated by one or more humans with subjects who should preferably believe that they are interacting with a real system (Fraser, Gilbert 1991; Bernsen *et al* 1998). The purpose is to gather data early on concerning how well the system or component might work, which means that analysis and evaluation of WOZ data is an important part of this phase. The advantage of early simulation is that, if done extensively and analysed carefully, a large number of problems with the design concepts as evidenced by observed phenomena in the simulated human-system interactions can be spotted, diagnosed, and removed early in the development process. The disadvantage is the cost of setting up and running the simulations, and of analysing the generated data. WOZ data gathering often includes use of questionnaires and interviews for investigating subjects' opinions of the simulated SLDS or component. These may provide crucial insights into the users' perception of the system or component and help capture user observations which might have implications for virtually any kind of deficiency. The documentation produced in this phase addresses the preparation and set-up of the WOZ simulation, the actual experiments carried out, the analysis of the collected data, and the implications for the design specification and the next WOZ iteration (if any). Documentation evaluation focuses on whether the WOZ process is adequately represented.

In the *construction phase,* the specified system or component is being implemented and debugged. During debugging, two typical types of diagnostic test are glassbox tests and blackbox tests. In a glassbox test, the internal system representation is inspected. The evaluator should ensure that reasonable test suites, i.e. data sets, are constructed which will activate all loops and conditions of the program being tested. In a blackbox test, only input to, and output from, the program are available to the evaluator. Test suites are constructed in accordance with the requirements and design specifications, and along with a specification of the expected output. Expected and actual output are compared and deviations must be explained. In general, either there is a bug in the program or the expectation is incorrect. Bugs must be corrected and the test run again. It may be added that the blackbox test may also suggest that the expected output, even if forthcoming, is flawed, leading to partial redesign and illustrating the iterative nature of development. Test suites should include fully acceptable input as well as borderline cases to test if the program reacts reasonably and does not break down in case of input error. The documentation produced in the construction phase includes comments in the code, proper descriptions of the implemented system or component architecture, modules and interfaces, and test data and results. Documentation evaluation consists in checking whether this information is provided in an acceptable form.

The outcome of the *integration phase* is the final system or component. The final integration of system or component parts is done in this phase. The integrated system is tested thoroughly and, when judged to satisfy the requirements, delivered to the customer. The tests involve interaction between the system and real users, either in controlled experiments with selected users and scenarios which they have to perform, or in field studies in which the SLDS or component is being exposed to uncontrolled user interaction. The collected user-system interaction data is analysed and used to evaluate the system or component. The user-system interaction data may be complemented by data from questionnaires and interviews. The main difference is that in the integration phase there should be far fewer problems to diagnose and solve. The final test will often be the acceptance test in which the procurer (if any) tests the system or component according to the evaluation criteria specified early on (cf. above) to verify that it meets the agreed requirements. The documentation in this phase

includes a user manual and guide (if any), descriptions of preparations and set-up of controlled user tests and/or field tests, and of the actual tests carried out, analysis of the collected data and possibly of the implications for the implemented system. Documentation evaluation consists in checking whether the documentation is provided and adequate.

Maintenance is the final (and longest) phase in the life-cycle. The system or component is in actual use during this phase and maintenance includes, e.g., the correction of errors, functionality improvements, and extensions to provide new services.

During the life-cycle phases listed in Figure 6, a global description of the development and evaluation process should be worked out, cf. Figure 7. This document should reference all materials produced and provide a description of the process. The document is useful during maintenance by providing easy access to information, and it can be invaluable when planning new projects. It is also useful for new members joining the development team after the project has started.

Moreover, parameters which cannot be fully controlled in advance and which may negatively influence the development and evaluation process at any time in the life-cycle, should be monitored regularly throughout and action taken as early as possible when needed, cf. the management factors in Figure 7.

Any of the phases in Figure 6 may be iterated as needed. For example, the simulation phase will typically iterate with the design phase a number of times in order to get the design specification right, and the integration phase will normally reveal problems which require (minor) adjustments to the code, thus iterating with the construction phase. Moreover, phases my overlap, so that, e.g., WOZ planning and simulation may be initiated whilst the design specification is being worked out.

5. Conclusion

We have presented an extended DISC dialogue engineering model which aims to support developers of SLDSs and components by providing a life-cycle model which is tailored to the needs of SLDS development and evaluation. The life-cycle model takes a general, iterative life-cycle model as its starting-point and specialises the model through aspect-specific grid issues, i.e. properties of SLDSs and their components, as well as evaluation criteria and methods. The model

presented might be transferable to specialised software engineering models in other areas of speech and language engineering and beyond. The model's dependence on a state-of-the-art 'grid' means that the model has to be continuously updated in order to take into account novel technical developments as well as their usability implications. In the area of SLDSs, new technical developments will include, among others, multimodal dialogue systems which include spoken dialogue, web-based spoken dialogue systems, and the taking of a major step beyond task-oriented SLDSs towards domain-oriented systems which no longer enable particular tasks but allow free-form conversation about any topic in a domain.

References

Beck, K. 1999a. Extreme Programming Explained. Embrace Change. Addison-Wesley.

Beck, K. 1999b. Embracing change with extreme programming. – IEEE Computer, October. 70–77.

Bernsen, N. O.; Dybkjær, H.; Dybkjær, L. 1998. Designing Interactive Speech Systems. From First Ideas to User Testing. Springer Verlag.

Bernsen, N. O.; Dybkjær, L. 2000a. From single word to natural dialogue. – Advances in Computers, Vol. 52. Ed. by Marvin V. Zelkowitz. London: Academic Press. 267–327.

Bernsen, N. O.; Dybkjær, L. 2000b. A methodology for evaluating spoken language dialogue systems and their components. – Proceedings of the Second International Conference on Language Resources and Evaluation (LREC 2000), Athens. 183–188.

Bernsen, N. O.; Dybkjær, L.; Heid, U. 1999. Current practice in the development and evaluation of spoken language dialogue systems. – Proceedings of Eurospeech'99, Budapest, Hungary. 1147–1150.

Boehm, B. W. 1988. A spiral model of software development and enhancement. – IEEE Computer 21:5, 61–72.

Dybkjær, L.; Bernsen, N. O. 2000. Usability issues in spoken language dialogue systems. – Natural Language Engineering. Special Issue on Best Practice in Spoken Language Dialogue System Engineering 6:3-4, 243–272.

Fraser, N. M.; Gilbert, G. N. 1991. Simulating speech systems. – Computer Speech and Language 5, 81–99.

Muller, P.-A. 1997. Instant UML. Canada: Wrox Press.

Pressman, R. 1997. Software Engineering: A Practitioner's Approach, European Edition. McGraw Hill.

Royce, W. W. 1970. Managing the Development of Large Software Systems. Proc. WESTCON, San Francisco CA.

Sommerville, I. 1992. Software Engineering. Fourth Edition, Addison-Wesley (1992 or newer edition).

Vaja on veel üht eesti keele grammatikat

Mati Erelt, Tiit Hennoste
Tartu ülikool

Eesti keele teadusliku grammatika ilmumisest on möödas juba õige mitu aastat. Seda grammatikat on selle aja jooksul jõutud nii kiita kui ka kritiseerida. Kritiseeritud on rohkem grammatika esimest osa, eriti foneetika-fonoloogiaosa puudumist, morfoloogia liigset keerukust jm (nt Hint 1997, 1998). Süntaksist on kirjutatud-räägitud leebemas toonis. Retsensiooni vormis on EKG süntaksiosa ainsana analüüsinud Haldur Õim (1996), kelle märkustest on olnud kasu EKG kasutamisel ülikooliõppes. Grammatika esimese osa põhioponent Mati Hint on asunud koos kolleegidega Tallinna Pedagoogikaülikoolist koostama alternatiivset sõnagrammatikat. Tulemust näeme loodetavasti mõne aasta pärast. Et süntaksi suhtes rahulolematust eriti väljendatud ei ole, siis pole räägitud ka vajadusest uue süntaksi järele. Ega nii suurt tööd, nagu seda on akadeemilise grammatika koostamine, iga mõne aasta tagant ette võtta saagi, sest eeldab ju suure ülevaateteose koostamine küllaltki pikaajalisi süvauuringuid.

Iseasi on muidugi see, kui on tekkinud vajadus põhimõtteliselt uut tüüpi grammatika järele, mis ei asendaks, vaid pigem teisendaks või täiendaks olemasolevat mingist aspektist, ning kui on olemas ka eeldused selle vajaduse realiseerimiseks. Niisugune olukord on meie arvates kätte jõudnud ning ettevalmistustöödega võiks pihta hakata. Kuna me ei tea, mis nägu on tulevasel pedagoogikaülikooli sõnagrammatikal, siis oleks mõistlik alustada süntaksist. Põhjusi uue süntaksi tegemiseks on vähemalt kolm.

Esiteks, vajaduse olemasolevat süntaksit teisendava-täiendava lauseõpetuse järele tingib kõigepealt EKG ja üldse eesti süntaksi kohati **suurevõitu eestikesksus**. Ehkki EKG süntaks on traditsioonilise eesti keele lauseõpetusega võrreldes suur samm edasi rahvusvaheliselt rohkem levinud kirjelduse poole, on selles ikka veel üht-teist sellist, mis on suurema pingutuseta arusaadav ainult meile endile või vanast rasvast ka mõnele meie soome kolleegile.

Osaliselt tundmatu teistele lauseuurijatele on näiteks meie grammatika lauseliikmete süsteem: predikatiivi (*Laps on* haige) ja

predikatiivadverbiaali ehk EKG järgi seisundimääruse (*Laps jäi* haigeks. *Laps läks* haigena *kooli*) lahkulöömine, seotud ja vabade laiendite kooskäsitlemine määruse all, lisandi pidamine täiendi eriliigiks jpm.

Ka oleme oma kirjeldustes mõningaid tavalisi süntaksinähtusi liiga vähe esiplaanile tõstnud. Nii jäävad EKGs omaette käsitluseta niisugused põhilised süntaktilised nähtused nagu ühildumine ja sõnajärg. Neid on küll vaadeldud ühes, teises ja kolmandas kohas, kuid mitte ühe tervikliku nähtusena, mida juhivad kindlad seaduspärasused (ühildumise puhul näiteks ühildumishierarhia, sõnajärje puhul näiteks moodustaja pragmaatiline staatus, moodustaja "raskus" jm). Ei ole eraldi käsitletud valentsi muutvaid süntaktilisi protsesse (kausativatsiooni, tõstet jm), mille kohta mõningat infot küll leiame, kuid grammatika eri nurgist.

Kui lehitseda ühe või teise süntaksinähtuse tüpoloogilisi ülevaatekäsitlusi, aga ka süntaksi või morfosüntaksi küsitluskavu, mis on kokku pandud andmete hankimiseks mingi(te) süntaktilis(t)e nähtuste avaldumise kohta konkreetsetes keeltes,[1] siis näeme, et neid küsimusi, millele ei ole EKG põhjal kerge kohe vastust leida, on üksjagu. Vastused võivad grammatikakirjelduses ju olemaski olla, kuid peidetud kujul, sest vajadust nende eksplitseerimise järgi pole küsimuste puudumise tõttu lihtsalt olnud.

Seega: et osaleda eesti keele süntaksiuurimustega rahvusvahelises süntaksidiskussioonis, tuleks senine eestikeskne süntaksikirjeldus kõigepealt mujal kasutatavaga suuremasse kooskõlla viia. Niisugune restruktureerimine ning ühtlasi eesti ja muu vahelise seose näitamine peaks olema nimelt eraldi grammatika ülesandeks, mitte iga uurija eralõbuks, nagu see seni on olnud. Rõhutagem, et me ei taha ilmtingimata üht kirjeldusmalli teisega asendada, vaid ehitada silda eestikeskse süntaksikirjelduse ja laiemalt levinud funktsionaaltüpoloogilise süntaksikirjelduse vahele. Õigupoolest on silla ehitus juba ammu alanud, töö tuleb lihtsalt lõpule viia.

Teine põhjus koostada veel üks grammatika on vajadus tuua grammatikasse uus kirjeldusaspekt – **näidata keelekasutuse, antud juhul**

[1]Üks põhjalikumaid küsitluskavasid on B. Comrie ja N. Smithi kava (Comrie, Smith 1977). Selle alusel on koostatud mitmete keelte süntaksiülevaateid, sealhulgas ka soome keele kohta (Karjalainen, Sulkala 1996) ja väike ülevaade isegi eesti keele kohta (Voitk 1987).

siis süntaksi varieerumist allkeeliti. Senised grammatikad kõnelevad eesti keelest kui millestki väga ühtsest, tuues vaid harva ja juhuslikult näiteid erinevustest eri allkeelte vahel. Erinevusi on mainitud põhiliselt sõnavara juures, erinevatest stiilikihtidest rääkides. Mujal maailmas on hakanud juba ilmuma sellised grammatikad, mis kirjeldavad keelt allkeelte süsteemina ning toovad välja eri allkeelte kvalitatiivsed ja kvantitatiivsed sarnasused ja erinevused. Ennekõike nimetatagu uut inglise keele grammatikat – Douglas Biber jt "Longman Grammar of Spoken and Written English" (1999) ja esialgu veel pooleli olevat soome keele akadeemilist grammatikat (peatoimetaja Auli Hakulinen). Aeg on hakata koostama sellist grammatikat ka eesti keele kohta.

Keele jagunemine allkeelteks on tingitud keelekasutajate ning kasutussituatsioonide erinevustest (vt Hennoste 2000; teisi lähenemisi allkeeltele vt kogumikust Hennoste toim 2000). Kasutajate erinevuse (sugu, vanus, territoriaalne kuuluvus, sotsiaalne staatus jms) arvestamine eesti keele süntaktilises kirjelduses ei ole esmatähtis, need erinevused avalduvad keeles ennekõike sõnavaras, morfoloogias ning häälduses, mitte süntaksis.

Kasutussituatsiooni erinevused seevastu kajastuvad süntaksis hoopis suuremal määral. Kõige rohkem erinevad omavahel kaks põhilist situatiivset allkeelt ehk metaregistrit – kirjalik register, mis oma prototüüpsel kujul on ühtlasi redigeeritud, avalik ning monoloogiline, ning suuline register, mis prototüüpsena on spontaanne, argine ning dialoogiline, kandes nime argivestlus. Nende kahe allkeele erinevused on ilmselt suurimad, mis keeles olemas on, ja nad põhinevad suhtlussituatsioonide olemuslikul ja kaotamatul erinevusel.

Argiregistrit tajutakse ühiskonnas tavaliselt ühe eraldi allkeelena, kirjalik põhiregister jaguneb aga mitmeks allregistriks, millest olulisimad on ilukirjandus, ajakirjandus, asjaajamine ja teadus. Neist omakorda on kultuuris ja ühiskonnas kesksed kaks – ajakirjandus ja ilukirjandus, mis on ühtlasi üldkeelelised registrid. Teadus ja asjaajamine esindavad üldjuhul ametikeele registreid, mille kasutus ja levik on palju kitsamad ja mis sisaldavad palju terminoloogiat. Eesti kultuuris on nii ilukirjandus kui ajakirjandus olnud pikka aega normingulised registrid. Sealjuures on kalded nendest olnud lubatud üksnes ilukirjanduse dialoogis kindlates suundades ja kindlatel eesmärkidel.

Eesti keele senine süntaksikirjeldus põhineb peaaegu täielikult kirjalikul allkeelel. Samal ajal on tegelik grammatikakirjeldus käsitlenud kirjalikku registrit jagamatu tervikuna, ehkki jagunemine allregistriteks (mida on nimetatud ka funktsionaalstiilideks) on ammu teada asi ja igas grammatikas ka mainitud. Suulise keele erijooni grammatikates tavaliselt isegi ei mainita. Nende uurimine on kestnud ka väga lühikest aega. Praeguseks on olemas ainult rida üksikuurimusi ja Tiit Hennoste ülevaade suulise eesti keele erijoontest (Hennoste 2000–2001). Viimane tegeleb aga valdavalt nende joontega, mida kirjalikus keelekasutuses peaaegu ei esine, jättes käsitlemata mõlemas registris esinevate süntaksinähtuste kasutuserinevused.

Kuivõrd kirjaliku registri süntaktiline ühtsus on võrreldamatult suurem kui lahknevused eri allkeelte vahel, siis jääb kavandatava süntaksi põhiosa ikkagi suuresti edasi senise süntaksi taoliseks. Küll aga peaks sellele lisanduma vähemalt kolme allkeele – argivestluste keele ning kirjalikust registrist vähemalt ilu- ja ajakirjanduse keele kvalitatiivsete ja kvantitatiivsete erinevuste käsitlus.

Kolmas oluline põhjus uue süntaksi kirjutamiseks lähtub uurimises kasutatavast keelematerjalist. Et saada adekvaatset pilti keelenähtuste tegelikust kasutusest, **peab süntaksikirjeldus toetuma piisavalt suurtele tekstikorpustele.** Senine grammatikakirjeldus on toetunud eelkõige eri tekstidest pärit juhunäidetele ja kirjutajate introspektsioonile. Sellisel moel saab kätte ainult osa tegelikust keelekasutusest ja pole üldse võimalik teha eri allkeelte võrdlusi. Ainult korpuste põhjal on võimalik teha nii keelekujude siseseid kui ka nende vahelisi kvantitatiivseid võrdlusi, eristada tüüpilist mittetüüpilisest ning kirjeldada muutumistendentse.

Inglise ja soome keele grammatikate koostamise kogemus näitab, et soliidse 1000–2000-leheküljelise deskriptiivse grammatika koostamiseks on piisav umbes 5 miljoni sõne suurune tekstikorpus iga allkeele jaoks. Liiga suured korpused muutuvad praktiliselt analüüsimatuks. Nende keelenähtuste kirjeldamiseks, mis tulevad välja alles väga suurte korpuste puhul (näiteks on nähtusi, mis tulevad välja alles 100 miljoni sõne suurusest korpusest), tuleb loota muudele ainestiku hankimise viisidele – informantide kasutamisele ning juhunäidetele. Informantide kasutamine on paratamatu juba sellegipärast, et tuvastada struktuuride vahelisi ekvivalentsussuhteid.

Eesti keele kohta on olemas 20. sajandi kirjaliku keele korpus, mis koosneb eri kümnendite allkorpustest. Kõik kümnendikorpused

peale 1980. aastate oma sisaldavad ilukirjandust ja ajakirjandust (1980. aastate nn baaskorpus ka muude allkeelte korpusi). 1990. aastate ilukirjanduskorpuses on umbes 600 000 sõnet, ajakirjanduskorpuses umbes 400 000 sõnet (vt http://www.cl.ut.ee/ee/corpusb/). Suulise keele litereeritud korpus sisaldab 350 000 sõnet, millest umbes poole moodustab argivestlus (vt Hennoste jt 2000). Kirjaliku registri korpuste koostamisel on lähtutud põhimõttest, mis on väikeste korpuste koostamisel üldine – valida katkeid võimalikult paljudest erinevatest tekstidest, eelnevalt prognoosides, millised tekstid on ühiskonnas olulised ja millised erinevad oma keelekasutuse poolest (vt Hennoste, Muischnek 2000).

Süntaksi koostamisel tuleks ilmselt lähtuda mitte varasematest kui 1990. aastate tekstidest, kasutades varasemate perioodide korpusi vaid 20. sajandil toimunud olulisemate süntaktiliste muutuste kirjeldamisel. Kahtlemata on aga 1990. aastate korpus oma praeguses mahus ebapiisav korraliku korpusekeskse süntaksi koostamiseks ja vajab täiendamist nii ajakirjanduse kui ka ilukirjanduse osas. Ajakirjandustekstidega ei ole probleemi. Kui tahta teha korpust suurusega 5 miljonit sõnet, siis tuleb võtta sisse kõik vähegi olulised eesti ajalehed. Kuid ilukirjanduse osas on takistuseks originaaltekstide nappus. Ka praegustes korpustes on ilukirjandustekstide kogumisse võetud igast eesti originaalsest ilukirjandusteosest üks katke, tavaliselt pikkusega ca 2000 sõnet. Seega saaks selle korpuse mahtu tõsta ainult juhul, kui suurendada oluliselt valitavate katkete pikkust, pikendada uuritavat perioodi või võtta lisaks originaalteostele sisse ka tõlked. Kuid ka siis on jõudmine 5 miljoni sõneni ilmselt võimatu ning leppida tuleb märksa vähemaga.

EKG süntaksile on vaja täienduseks uut süntaksit, see on selge. See ei pruugi olla nii mahukas kui EKG oma, rääkimata inglise või soome akadeemilise grammatika süntaksiosadest. Usume, et niisuguse süntaksi koostamine on ka võimalik. Kuid selleks on siiski vaja mõneaastast ettevalmistusperioodi, et täiendada korpusi ning teha uurimuslikke eeltöid.

Kirjandus

Biber, Douglas 1988. Variation across Speech and Writing. Cambridge: Cambridge University Press.

Comrie, B.; Smith, N. 1977. Lingua Descriptive Studies: Questionnaire. – Lingua 42, 1–72.

EKG I, II = Erelt, Mati; Kasik, Reet; Metslang, Helle; Rajandi, Henno; Ross, Kristiina; Saari, Henn; Tael, Kaja; Vare, Silvi. Eesti keele grammatika I ja II. Eesti Teaduste Akadeemia Keele ja Kirjanduse Instituut. Tallinn 1995 ja 1993.

Hennoste, Tiit 2000. Allkeeled. – Eesti keele allkeeled. Tartu ülikooli eesti keele õppetooli toimetised 16. Toim. T. Hennoste. Tartu. 9–56.

Hennoste, Tiit (toim) 2000. Eesti keele allkeeled. Tartu ülikooli eesti keele õppetooli toimetised 16. Tartu.

Hennoste, Tiit 2000–2001. Sissejuhatus suulisesse eesti keelde I–IX. Akadeemia 2000: 5, 1117–1150; 6, 1343–1374; 7, 1553–1582; 8, 1773–1806; 9, 2011–2038; 10, 2223–2254; 11, 2465–2486; 12, 2689–2710; 2001: 1, 179–206.

Hennoste, Tiit; Muischnek, Kadri 2000. Eesti kirjakeele korpuse valiku ja märgendamise põhimõtted ning kahe allkeele võrdluse katse. – Arvutuslingvistikalt inimesele. Tartu ülikooli üldkeeleteaduse õppetooli toimetised 1. Toim. T. Hennoste. Tartu. 183–218

Hennoste, Tiit; Lindström, Liina; Rääbis, Andriela; Toomet, Piret; Vellerind, Riina 2000. Eesti suulise kõne korpus ja mõne allkeele võrdlemise katse. – Arvutuslingvistikalt inimesele. Tartu ülikooli üldkeeleteaduse õppetooli toimetised 1. Toim. T. Hennoste. Tartu. 245–284.

Hint, Mati 1997. Eesti foneetika möödunud hiilgus ja möödumata viletsus. – Keel ja Kirjandus 2, 73-82; 3, 157–164.

Hint, Mati 1998. Kolmkümmend aastat hiljem. Uute eesti keele grammatikate puhul. – Keel ja Kirjandus 2, 100–112; 3, 188–201.

Longman 1999. Douglas Biber, Stig Johansson, Geoffrey Leech, Susan Conrad, Edward Finegan. Longman Grammar of Spoken and Written English. Longman.

Sulkala, Helena; Karjalainen, Merja 1996. Finnish. London, New York: Routledge.

Voitk, Malle 1987. Estonian Syntax. Answer to some of the Questions in Lingua Descriptive Studies Questionnaire. Stockholm University, Department of Linguistics. (käsikiri)

Õim, Haldur 1996. Sissejuhatavalt eesti akadeemilisest grammatikast. – Keel ja Kirjandus 12, 852–854.

Puheentutkimus ja keskustelunanalyysi Suomessa

Auli Hakulinen
Helsingin yliopisto

1. Taustaksi: puheentutkimuksen koko kuva

Elokuussa 2001 julkistettiin Opetusministeriön rahoituksella laadittu valtakunnallinen esiselvitys puheentutkimuksen resursseista Suomessa (Miettinen, Toivanen 2001). Esiselvityksen tarkoituksena oli kartoittaa, missä yliopistollisissa ja muissa tutkimuslaitoksissa tällä hetkellä tutkitaan puhetta, millä resursseilla ja miltä kannalta. Toiseksi tarkoitus oli ehdottaa toimenpiteitä, joilla nykyistä niin maantieteellisesti kuin metodisesti hajanaista tilannetta voitaisiin parantaa.

Selvityksestä ilmenee, että tavalla tai toisella puhuttujen aineistojen parissa työskenteleviä laitoksia on Suomessa vähintään 24 – tämä määrä instituutioita vastasi kyselyyn, joka oli lähetetty peräti 44 potentiaalisesti puhetta tutkivalle laitokselle. Selvityksessä ei kerrota niiden laitosten nimiä, joilta vastausta ei ole saatu. Vastanneista laitoksista valtaosa toimii yliopistoissa (fonetiikan ja kielten laitokset) ja teknisissä korkeakouluissa (esim. neuroverkkotutkimus, kognitiivinen teknologia, tietojenkäsittelytiede); mukana on neljä kaupallista tutkimusyksikköä sekä lähinnä äänikirjojen kehittämiseen keskittyvä Näkövammaisten keskusliitto.

Tieteellinen kiinnostus puhuttua kieltä kohtaan on kasvanut eri tahoilla nopeasti parin viime vuosikymmenen aikana. Syyt ovat monet, ja vain osa syistä on puhtaasti tieteensisäisistä intresseistä lähtevää. Lähinnä humanistis-yhteiskuntatieteellisillä aloilla kiinnostusta ohjaavat tieteen sisältä nousevat kysymykset; mainittakoon tässä jo perinteikäs alueellisen ja sosiaalisen variaation tutkimus sekä uudempana tulokkaana kielitieteen ja käyttäytymistieteiden välimaastossa harjoitettu keskusteluntutkimus. Tästä ei kuitenkaan välttämättä seuraa, että perustutkimuksella ei olisi sovellusarvoa erinäisiin käytännön tilanteisiin.

Teknologian puolella tutkimusta ohjaavat myös tai joskus yksinomaankin käytännön tavoitteet, kuten tiedonhakujärjestelmien kehittäminen, ihmisen ja (tieto)koneen vuorovaikutus tai puhelimen välityksellä käytettävien informaatiopalveluiden kehittäminen. Yhteistä näille on pyrkimys kehittää puheen ja puhujan tunnistamis-

menetelmiä ja luoda laajoja, monipuolisesti käyttökelpoisia puhetietokantoja.

Teknologiapainotteiseen perustutkimukseen luettavia hankkeita ovat erilaiset aivotutkimushankkeet kuten ns. puhuva pää–hanke, jonka tavoitteena on kartoittaa aivojen audiovisuaalisia integraatioalueita, taikka hanke, joka yrittää selvittää mm. sitä, miten äännejärjestelmä kehittyy vastasyntyneen aivoissa. Näilläkin hankkeilla näyttää olevan myös kytkentöjä telekommunikaation suuntaan.

Resurssit eivät kaiken kaikkiaan ole kansainvälisesti katsoen suuret. On kuitenkin huomattava, että kun ihmistieteissä tutkimusta tehdään opinnäytteinä tai opetuksen oheistyönä, puheteknologian alalla on tähän verrattuna jopa suuria tutkimushankkeita (useita päätoimisia tutkijoita), joita rahoittavat Suomen Akatemian lisäksi esimerkiksi Tekniikan edistämiskeskus, Ilmavoimat tai Keskusrikospoliisi.

2. Puhetietokannat ja puhutun kielen korpukset

Varsinkin erilaisia puheteknologisia sovelluksia tavoitteleva kehitystyö tarvitsee luonnollisesti tuekseen tietyllä tavalla esikäsiteltyjä puhetietokantoja eli puhekantakorpuksia. Voimassa oleva ristiriita eri tutkimussuuntauksien välillä aiheutuu paitsi tallennusten teknisestä monenkirjavuudesta ja siis keskinäisestä yhteensopimattomuudesta myös siitä, että eri tarkoitusperiin tarvitaan ja kelpuutetaan ainakin tällä hetkellä kovin erityyppistä puheaineistoa. Tiettyjä puheentunnistamistarpeita varten voidaan pitkään tulla toimeen vaikkapa vain yksinkertaisilla merkityksettömillä foneemisekvensseillä. Korkeatasoinen äänen laatu on tietyssä tutkimuksen vaiheessa tärkeämpi kuin puheen luonnollisuus ja keskustelullisuus. Tällöin korpukseen kelpuutetaan hyvissä äänitysoloissa tuotettuja sanoja, korkeintaan yksittäisiä lauseita tai yhtäjaksoisempaa ääneen luentaa eli lukupuhuntaa. Mutta kun yksinkertaiset tunnistusongelmat on ratkaistu, alkaa puheteknologian tutkijoiden kiinnostus kääntyä myös kohden pitempiä puhejaksoja sekä selvästi erottuvien äännesegmenttien kuvaamisesta toisiinsa sulautuneiden äänteiden muodostamiin prosodisiin hahmoihin, jollaista puhe on luonnollisessa, aidossa keskustelussa. Tässä tilanteessa ollaan nyt Suomessa. Seuraava haaste puheteknologialle on luonnollisen, ihmistenvälisen keskustelun kohtaaminen.

Ihmistieteiden puolella korpukset ovat myös syntyneet jotain tiettyä, joskin yleensä edellisiä laaja-alaisempaa tarkoitusta varten. Kattavin, n. 16 000 tuntia käsittävä puhutun kielen kokoelma on 1950-luvun lopussa Helsingin yliopistoon perustettu Suomen kielen nauhoitearkisto, jonka nykyinen sijaintipaikka on Kotimaisten kielten tutkimuskeskuksessa. Arkiston perustamisen syynä oli vanhojen kansanmurteiden saaminen talteen, ja tavoitteena mahdollisimman kattava otos eri pitäjien murretta. Arkistosta sanotaan, että se on kerätty "kielentutkimuksen tarpeiksi" (Yli-Paavola 1970: 17), mutta sillä on varmasti suureksi osaksi museaalista merkitystä ja sitä lienee tähän mennessä eniten käytetty murteiden äänne- ja muoto-oppia koskeviin tutkimuksiin. Koska äänitteet ovat valtaosin tutkijan ja kielenoppaan välisiä haastatteluja, niiden keskustelunomaisuus on aika kyseenalaista. Toisaalta nauhoitteiden laatu on korkeatasoisempi kuin vapaissa puhetilanteissa kerättyjen nauhoitteiden, sillä aineistoa kerättäessä on lähtökohtaisesti pidetty silmällä sitä, että nauhoitus tapahtuisi mahdollisimman hälyttömässä ympäristössä.

Voi kuitenkin osoittautua, että ainakin nuorempi murreaineisto on käyttökelpoista sellaisen puheentutkimuksen tarpeisiin, joka pyrkii ottamaan huomioon puhujien murretaustasta johtuvan variaation ääntämistavassa. On siksi mainitsemisen arvoista, että tätä valtavaa murreaineistoa on viime vuodesta alkaen siirretty digitaaliseen muotoon, joten se tulee olemaan hieman nykyistä paremmin muidenkin kuin murteentutkijoiden käytettävissä. Ennalta arvaamatonta käyttöä voi kenties olla myös niin ikään haastattelutekniikalla kerätyillä sosiolingvistisillä aineistoilla, joita 1970-luvun lopulla kerättiin Nykysuomalaisen puhekielen murros–hankkeeseen neljästä Suomen suurimmasta kaupungista (tästä lähemmin ks. Mielikäinen 1980).

Uusimpia tulokkaita puhutun kielen tutkimuksen alalla ovat keskusteluntutkijat. Ensimmäinen keskusteluarkisto alkoi sattumalta, Suomalaisen keskustelun keinoja–hankkeen (1985–1989) saaman nauhalahjoituksen pohjalta (Hakulinen 1989). Lahjoitus käsitti kasvokkaisia, koetilanteessa käytyjä, osin spontaanejakin ryhmäkeskusteluja. Niitä täydentämään alettiin 1980-luvun lopulta kerätä puhelinkeskusteluja, ja viimeksi kuluneiden vuosien aikana arkistoon on kertynyt myös videoitua aineistoa. Keruu on pääosin tehty opiskelijavoimin: kielellisen vuorovaikutuksen tutkimiseen erikoistuminen edellyttää omaa keruu- ja litterointityötä. Tämä keskustelun-

tutkimuksen arkisto sijaitsee Helsingin yliopiston suomen kielen laitoksen yhteydessä. Arkisto ei kuitenkaan ole mikään systemaattinen, suunnitelmallisesti rakennettu korpus, vaan kvalitatiivisen tutkimuksen sivutuotteena koko ajan karttuva nauhavarasto (siitä lisää ks. 3.). Puheteknologisen hyödyntämisen kannalta tällainen aineisto on aika kaukana optimaalisesta. Äänitteiden samoin kuin kuvanauhojen taso vaihtelee suuresti, aineistosta on vain pieni osa litteroitu, eikä ääntä ja litteraattia ole toistaiseksi saatettu samalla kertaa käytettävään muotoon. Kvalitatiivisen, intensiivisen keskustelunanalyysin kannalta aineistossa sen sijaan on jo nykyisellään rajattomasti tutkittavaa.

3. Aineiston ja tutkimuksen suhde keskustelunanalyysissa

3.1. Aineiston lajeista

Kuten edellä mainitsin, Keskusteluntutkimuksen arkisto on syntynyt tutkimuksen ja opinnäytteiden sivutuotteena puolentoista vuosikymmenen aikana. Vasta vuoden 2001 kesällä keskusteluaineistot on arkistoitu systemaattisesti, ja siitä on laadittu tietokonemuotoinen kortisto, joka on tarkoitus seuraavaksi sijoittaa myös internetiin suomen kielen laitoksen kotisivulle. Näin arkistoa voi jatkossa käyttää myös laajempi tutkijayhteisö. Parhaillaan aineistoa ollaan digitoimassa, mikä takaa äänitteiden säilymisen ja helpottaa aineiston lainaamista. Toisin kuin esimerkiksi Tarton puhekielen korpus aineistomme on luokiteltu pikemminkin käytettävyyden kuin vaikkapa variaationtutkimuksen kannalta vain muutaman harvan muuttujan perusteella. Tällaisia ovat mm. tilannetyyppi, kanava, keskustelijoiden määrä sekä arvioitu murre ja ikä. Aineiston suurimman osan muodostavat arkikeskustelut, kaikkiaan 267 signumia eli erillistä nauhoitetta. Varsinkin viime vuosina kerätyistä kasvokkaiskeskusteluista melko suuri osa (130 kortistosignumia) on tallennettu sekä ääni- että kuvanauhalle. Nämä valinnat ovat osin metodisia: lähtökohtana on edelleenkin, että arkikeskustelu on nimenomaan lingvistisesti suuntautuneelle tutkimukselle sopivinta, koska se on monipuolisinta. Videointi taas helpottaa huomattavasti kasvokkaiskeskustelun selväämistä ja tutkimista ja mahdollistaa myös katseen, eleiden ja liikkeiden tutkimisen yhdessä kielenilmiöiden kanssa. (Katseen litteroimisesta tutkimusta varten ks. Seppänen 1995.) Yhtenä tärkeänä aineiston osana on mainittava ns. S2-aineisto (87

signumia) eli keskustelut, joissa yhden tai useamman puhujan äidinkieli on muu kuin suomi; keskustelukielenä on kuitenkin suomi. Aineisto on kertynyt suomi vieraana kielenä –oppiaineen tutkielmien sivutuotteena.

Arkistoon kuuluu kuitenkin myös jonkin verran institutionaalisia keskusteluja (187 signumia), joista suuri osa on peräisin erilaisista opetustilanteista (oppitunteja, urheiluvalmennustilanteita, vastaanottokeskusteluja). Tilannetyyppien suhteen arkisto onkin tekemässä työnjakoa Kotimaisten kielten tutkimuskeskuksen kanssa. Jälkimmäisessä on meneillään asioimispuheen hanke, jossa keskitytään keräämään institutionaalista aineistoa eri puolilta Suomea: näitä on mm. R-kioskilta, Kansaneläkelaitoksesta, neuvoloista ja kampaamoista (Sorjonen 2001). Suuri osa Keskusteluntutkimuksen arkiston aineistosta on vapaasti lainattavissa eikä ole sidottu mihinkään erityiseen hankkeeseen: sitä mukaa kuin opiskelijoiden tutkielmat valmistuvat, heidän aineistonsa luovutetaan yhteiseen käyttöön, jos siihen on informanteilta pyydetty lupa. Tässä suhteessa aineisto eroaa määrähankkeisiin kerätystä materiaalista, joka on useimmiten suljettua. Esimerkiksi Peräkylän ja Sorjosen johtaman lääkärin ja potilaan vuorovaikutusta koskevan tutkimuksen (Sorjonen, Peräkylä, Eskola 2001) aineisto on vain tutkimusryhmän omassa käytössä, samoin Peräkylän hankkeen Hoitoideologia ja vuorovaikutus aineisto on eettisistä syistä vain rajoitetussa käytössä.

Silmäänpistävä aukko arkistossa on mediapuheessa: vain 13 signumia on peräisin radiosta tai televisiosta. Tämä johtuu pitkälti vaatimuksesta, että opiskelijoiden on ollut kerättävä aineistonsa "kentältä", siis muutoin kuin omasta televisiostaan omassa tuolissaan istuen. Toisaalta mediapuheeseen on erikoistunut suomalais–saksalainen tv-puheen hanke (Nuolijärvi, Tiittula 2000), jonka aineisto sijoittunee osaksi Kotuksen institutionaalisen puheen arkistoa.

Suomen kielen keskusteluarkiston käytettävyyden pullonkaulana on, kuten lähes kaikissa keskusteluarkistoissa, että aineistoa ei ole kauttaaltaan litteroitu. Nauhoitteita on kaikkiaan n. 330 tuntia, josta tähän mennessä litteroitua vajaat 80 tuntia. On kuitenkin muistettava, että keskustelunanalyyttiseen metodiin kuuluu tärkeänä työvaiheena omakohtainen litterointi, joka on paras tapa perehtyä aineistoon kunnolla. Silloin kun tutkija ei itse litteroi, hänen on vähintäänkin pidettävä sekä äänitettä että litteraattia analyysinsa pohjana. Hyödyllistä kuitenkin olisi saattaa äänitteet ja litteraatit

yhdeksi tietokonemuotoiseksi korpukseksi. Tätä voidaan pitää yhtenä arkiston etätavoitteena.

3.2. Keskustelunanalyysin "tutkimusintensiivisyydestä"

Ensimmäiset maisterin tutkintoon kuuluvat opinnäytteet valmistuivat vuonna 1988. Tätä kirjoitettaessa on keskusteluaineiston pohjalta valmistunut 55 pro gradu –tutkielmaa. Niiden kirjo on valtava: toisessa päässä on yksittäisten partikkelien kuten *mm* tai *niinku* kuvaaminen, useissa käsitellään toimintatyyppejä kuten erimielisyyttä, väärinymmärryksiä tai direktiivejä; vajaat kymmenen työtä käsittelee aikuisen ja lapsen vuorovaikutusta. Vielä huomattavampi todiste keskustelunanalyyttisen metodin voimasta ja ekspansiivisuudesta on, että vuosina 1996–2001 alalta on ilmestynyt peräti 14 väitöskirjaa (ks. *Liite*), niistä neljä alkuperäisen kielitieteellisen tutkijayhteisön ulkopuolella: kolme sosiologiassa (Arminen, Koskinen ja Ruusuvuori) ja yksi kasvatustieteessä (Vehviläinen). Nämä temaattisesti tiettyyn tilannetyyppiin sitoutuneet tutkimukset eivät luonnollisesti ole voineet perustua Keskusteluntutkimuksen arkiston aineistoon, vaan tutkijoiden omiin keräelmiin, Ruusuvuoren tapauksessa lääkäri-potilas – hankkeen yhteiseen aineistoon. Myös afaatikkojen puhetta tutkivat väitöskirjat (Klippi, Laakso) perustuvat tietysti näiltä kerättyyn terapia-aineistoon. Metodiselta kannalta kaksi, Helasvuon ja Lauryn väitöskirjat edustavat pikemmin Santa Barbaran yliopiston puhutun diskurssin tutkimusta, mutta nekin ovat syntyneet kiinteässä vuorovaikutuksessa yhteisöön sekä ovat käyttäneet hyväkseen yhteistä keskusteluaineistopohjaa. Parhaillaan on tekeillä ainakin toinen mokoma väitöskirjoja, valtaosa edelleenkin suomen kielen alalta.

Koska tilanne tutkimuksessa on näin vilkas, voidaan kysyä, kumpi on tulevalle monitieteiselle puheentutkimukselle suuremmaksi hyödyksi, teknisesti melko alkeellinen joskin ilmiöiden kannalta rikas keskustelukorpus vai keskustelun strategioiden ja rakennepiirteiden kuvauksesta kumuloituva tutkimustieto – ja tutkijayhteisössä kumuloituva tietotaito ("knowhow").

Kirjoitelmani alussa mainitussa raportissa (Miettinen, Toivanen 2001: 32–33) asiantuntijana haastateltu Lauri Carlson huomautti toisaalta, että dialogin mallinnuksessa tehdyt kokeilut ovat jo vahvistaneet keskusteluntutkimuksessa tehtyjä "havaintoja". Toisaalta hän arveli, että juuri laajoilla aineistoilla olisi mahdollista ylittää

keskustelunanalyysin metodisia rajoituksia kuten sen "tulosten kvalitatiivisuus ja vaikea toistettavuus". Jos litterointi ja haku olisivat nykyistä vaivattomampia, on varmaa että ainakin yksittäinen tutkija voisi rakentaa kuvauksensa nykyistä vankemmalle perustalle. Kvalitatiivisen tutkimuksen ja nimenomaan keskustelunanalyysin ideana on kuitenkin, että tutkija käyttää maksimaalisesti hyväkseen omaa "pääomaansa", tietoaan kulttuurista sen jäsenenä ja asettautuu hermeneuttisesti selvittämään sitä, mihin kielellisiin ja tilannepiirteisiin keskustelijat orientoituvat. Tämä on viime kädessä aina työlästä ja hidasta.

Aukoista olen Carlsonin kanssa yhtä mieltä. Ehkä pahin aukko on siinä, että missään tutkimussuuntauksessa ei ole riittävästi tietoa prosodian vaikutuksesta ilmausten kokonaismerkitykseen ja ylipäätään siihen, miten ilmaukset kontekstissaan tulevat tulkituksi. Prosodia onkin eri tahoilla selvästi nousemassa tutkimuksen kohteeksi. Sukset ovat tälläkin alueella ristissä: keskustelunanalyytikot olettavat edelleen, että äidinkielenpuhujan kuulohavainto on tutkimuksen lähtökohta, jota "objektiivinen" laitteilla mittaus voi tukea mutta ei korvata.

3.3. Tulosten kumuloitumisesta

Vaihtoehto suurten aineistojen pyörittelylle voikin olla riittävän suuri tutkimusyhteisö, sillä sellaisen analyysikapasiteetti ja asiantuntemus on paljon rikkaampi kuin kirjoiksi ja artikkeleiksi työstetyt konkreettiset tutkimustulokset. Yritän seuraavaksi nostaa esiin muutaman alueen ja ilmiön, joihin kohdistunut tutkimus on tuottanut tai tuottamassa kasautuvaa tietoa.

Ensimmäiseksi tulee mieleen erilaisten indeksaalisten, siis kontekstista käsin tulkittavien kielenpiirteiden tutkimus. Näitä ovat ennen muuta pronominit ja partikkelit. Pronomineista on toistaiseksi valmistunut kaksi väitöskirjaa (Laury 1996; Seppänen 1998), jotka ovat selvästi lisänneet ymmärtämystämme niiden toiminnasta keskustelussa ja korjanneet kielioppeihin kiteytyneitä stereotypioita. Esimerkiksi Seppänen on osoittanut, että perinteisesti ajateltujen toisen ja kolmannen persoonan pronominien välissä on tilaa ja että suomessa pronomini *tämä* ihmisestä käytettynä asettuu johonkin tuossa välimaastossa. Edelleen hän on selvittänyt toisen "demonstratiivisen" ilmiön, katseen suhdetta pronominin käyttöön. Laury puolestaan on näyttänyt, miten *se*-pronomini on sadan vuoden aikana kerännyt

itseensä artikkelimaisia piirteitä, vaikka se ei vieläkään ole täysin artikkelinomainen. Toiseksi hän on voinut osoittaa, miten pronominien käyttö on enemmän kuin heijastusta spatiaalisesta tilanteesta: niillä luodaan oletuksia referenttien tunnettuudesta vaikka puhuttaisiinkin aiemmin mainitsemattomasta asiasta. Näiden tutkijoiden linjaa jatkaa *tämä*-pronominia koskevassa tekeillä olevassa väitöskirjassaan Marja Etelämäki; hieman toiselta kannalta anaforisia pronomineja puheen arkkitehtuurin näkökulmasta omassaan Outi Duvallon.

Partikkelit ovat olleet pitkään keskustelunanalyysissa huomion kohteena, mutta suomen kielestä on tehty varsinkin dialogipartikkeleista uraauurtavaa tutkimusta (ks. Sorjonen 1997). Juuri dialogipartikkelit (esim. *niin, joo, ai, no*) ovat ilmiöryhmä, joka sopisi monitieteisen tutkimuksen kohteeksi monesta syystä. Ne ovat morfosyntaktisesti yksinkertaisia, muodostavat yksin kokonaisen lausuman ja niiden yhteydessä prosodisten piirteiden ja eleiden tutkiminen olisi suhteellisen hyvin kontrolloitavissa. Puheen tunnistamisen kannalta ne ovat sikäli ensiarvoisia, että ne ovat nimenomaan indeksejä siitä, miten edellisen puhujan vuoro on vastaanotettu ja tulkittu.

Toinen ilmiöryhmä, josta tieto kumuloituu, on vuorovaikutuspuheen rakenteelliset yksiköt. Kun puheteknologit operoivat yleensä kirjakielen yksiköillä ja tarkastelevat sanoja ja lauseita, keskusteluntutkijat käyttävät näiden lisäksi sellaisia analyysikäsitteitä kuten lausuma, vuoron rakenneyksikkö, vuoro, perättäisistä vuoroista koostuva vieruspari tai laajempikin sekvenssi. Tieto näiden rakenteellisista ja prosodisista ominaisuuksista lisääntyy (ks. esim. Helasvuo 1997), joskin aika hitaasti.

Yksi perustavanlaatuinen ero keskustelunanalyysin ja filosofis-kielitieteellisen pragmatiikan tutkimuksen välillä on suhtautuminen toimintaan. Pragmatiikassa ainakin aluksi pyrittiin eristämään luonnollisesta puheestakin puheakteja ja puheaktityyppejä, enemmän tai vähemmän menestyksellisesti. Keskustelunanalyysi on syntynyt sosiologiassa, ja sen sisäänrakennettuna ideologiana on vuorovaikutus toimintana ja nimenomaan yhteistyönä. Tässä lähestymistavassa katsotaan, että myös ns. vastaanottaja toimii merkityksenantajana, ei pelkkänä puhujan merkityksen selvääjänä. Tämän vuoksi käytetäänkin juuri nimitystä vastaanottaja kuulijan sijasta, koska se luo puheen vastaanottamisesta kyllin aktiivisen mielikuvan.

Jotkut toiminnot ovat melko helposti tunnistettavissa, ainakin päältä katsoen. Tällaisia ovat kysymykset ja vastaukset. Näitä toimintoja koskeva tietämys on keskustelunanalyyttisten tutkimusten myötä valtavasti kasvanut verrattaessa sitä perinteiseen kieliopillis-rakenteelliseen tietoon. Suomessa on tehty paitsi arkikeskusteluun rakentuvaa perustutkimusta (Raevaara 1996; Hakulinen 2001) myös tuotettu runsaasti eri tilannetyyppeihin liittyvää tietoa niin kysymysten kuin vastaustenkin käytöstä ja tulkinnasta (esim. Kajanne 2001; Raevaara 2001; Ruusuvuori 2000; Vehviläinen 1999). Aina ei kysyminenkään tukeudu interrogatiiviseen lausetyyppiin, muodon ja merkityksen suhde ei ole yksiyksinen. Kysymyksellä voidaan tehdä samalla kertaa muutakin kuin etsiä tietoa. Ja samaa muotoa voidaan käyttää useaan tarkoitukseen. Esimerkiksi tarkistuskysymys tai korjausaloite *täh*? voi olla osoitus siitä, että vastaanottaja ei ole kuullut, ymmärtänyt tai uskonut sitä mitä edellisessä vuorossa on esitetty, tai hän siirtyy tällä ilmauksella päivittelemään kuulemaansa.

Mutta on paljon toimintoja, joiden tunnistaminen ei perustu niinkään vuoron kielellisiin kuin ei-kielellisiin piirteisiin. Yksi tällainen keino on nauru tai naurahdus, jolla puhuja voi paitsi merkitä vuoronsa sellaista asiaa sisältäväksi, jolle voidaan yhdessä nauraa, myös tarjota vastaanottajalle tulkintavihjeen siitä, että vuoro sisältää tavalla tai toisella ongelmallista asiaa (Haakana 1999, 2001). Nämä harvat esimerkit osoittavat mielestäni jo, että keskustelunanalyyttinen tutkimus tuottaa perinteistä lingvististä tietämystä täydentävää tietoa.

Viitteet

Haakana, Markku 2001. Kielen ohessa? Näkökulmia vuorovaikutuksen ei-kielellisiin keinoihin. – Keskustelunanalyysin näkymiä. Toim. M. Halonen, S. Routarinne. Kieli 13. Helsingin yliopiston suomen kielen laitos. 70–88

Hakulinen, Auli 1989. Keskustelun tutkimisen tavoitteista ja menetelmistä. – Suomalaisen keskustelun keinoja I. Toim. A. Hakulinen. Kieli 4. Helsingin yliopiston suomen kielen laitos. 9–40

Hakulinen, Auli 2001 Minimal and non-minimal answers to yes–no questions. – Pragmatics 11:1, 1–15.

Mielikäinen, Aila (toim.) 1980. Nykysuomalaisen puhekielen murros. Jyväskylän osatutkimus. Jyväskylän yliopiston Suomen kielen ja viestinnän laitoksen julkaisuja 20.

Miettinen, Manne; Toivanen, Juhani (toim.) 2001. Puheentutkimuksen resurssit Suomessa. CSC – Tieteellinen laskenta Oy.

Nuolijärvi, Pirkko; Tiittula, Liisa 2000. Televisiokeskustelun näyttämöllä. Televisioinstitutionaalisuus suomalaisessa ja saksalaisessa keskustelukulttuurissa. Helsinki: SKS.

Raevaara, Liisa 1996. Kysymyksen paikka ja tulkinta. – Suomalaisen keskustelun keinoja II. Toim. A. Hakulinen. Kieli 10. Helsingin yliopiston suomen kielen laitos. 23–46

Raevaara, Liisa 2001. Vastaamisesta institutionaalisena toimintana. – Keskustelunanalyysin näkymiä. Toim. M. Halonen, S. Routarinne. Kieli 13. Helsingin yliopiston suomen kielen laitos. 47–69

Seppänen, Eeva-Leena 1995. Vuorovaikutus paperilla. – Keskustelunanalyysin perusteet. Toim. L. Tainio. Tampere: Vastapaino. 18–31

Sorjonen, Marja-Leena 2001. Asioimispuheen tutkimushanke. Suunnitelma vuosiksi 2002–2005. Käsikirjoitus. Kotimaisten kielten tutkimuskeskus.

Sorjonen, Marja-Leena; Peräkylä, Anssi; Eskola, Kari (toim.) 2001. Keskustelu lääkärin vastaanotolla. Keskustelunanalyyttinen tutkimus potilaan institutionaalisista tehtävistä. Helsinki: SKS.

Yli-Paavola, Jaakko 1970. Vuosikymmen kielennauhoitusta. Suomen kielen nauhoitearkiston toimintaa v. 1959 – 1968. Tietolipas 60. Helsinki: SKS.

Liite: Keskusteluntutkimuksen väitöskirjat 1996–2001

Arminen, Ilkka 1998. Therapeutic interaction. A study of Mutual Help in the Meetings of Alcoholics Anonymous. The Finnish Foundation of Alcohol Studies.

Haakana, Markku 1999. Laughing matters: a conversation analytical study of laughter in doctor–patient interaction. Helsingin yliopisto, Suomen kielen laitos.

Helasvuo, Marja-Liisa 1997. When discourse become syntax: NPs and clauses as emergent syntactic units in Finnish conversational discourse. Julkaisematon väitöskirja, UCSB. Tulossa John Benjamins 2002.

Kajanne, Milla 2001. Kansalaiset kysyjinä. Yleisökysyminen vuorovaikutuksen osana television EU-keskusteluissa. Helsinki: SKS.

Kangasharju, Helena 1998. Alignment in disagreement: building alliances in multiperson interaction. Helsingin yliopisto, Suomen kielen laitos.

Klippi, Anu 1996. Conversation as an achievement in aphasics. Helsinki: SKS.

Koskinen, Ilpo 1999. Managerial Evaluations at the workplace. Helsinki: Hakapaino.

Laakso, Minna 1997. Self-initiation of repair in conversations of fluent aphasic speakers. Helsinki: SKS.

Laury, Ritva 1997. Demonstratives in Interaction. Amsterdam: John Benjamins.

Raevaara, Liisa 2000. Potilaan diagnoosiehdotukset lääkärin vastaanotolla: Keskustelunanalyyttinen tutkimus potilaan institutionaalisista tehtävistä. Helsinki: SKS.

Ruusuvuori, Johanna 2000. Control in Medical Consultation. Acta Electronica Universitatis Tamperensis 16.

Seppänen, Eeva-Leena 1998. Läsnäolon pronominit: *tämä*, *tuo*, *se* ja *hän* viittaamassa keskustelun osallistujaan. Helsinki: Suomalaisen Kirjallisuuden Seura.

Sorjonen, Marja-Leena 1997. Recipient Activities: particles *nii(n)* and *joo* as Responses in Finnish Conversations. Julkaisematon väitöskirja, UCLA. Tulossa John Benjamins 2001.

Tainio, Liisa 2001. Puhuvan naisen paikka. Sukupuoli kulttuurisena kategoriana kielenkäytössä. Helsinki: SKS.

Vehviläinen, Sanna 1999. Structures of Counselling Interaction. University of Helsinki, Department of Education.

Eesti dialoogikorpuse loomise probleemid

Tiit Hennoste, Mare Koit, Maret Kullasaar,
Andriela Rääbis, Evely Vutt
Tartu ülikool

Sissejuhatus

Inimesega loomulikus keeles suhtlevate süsteemide loomisega tegelevad nii teaduslikud kui ka tehnilised distsipliinid, sealhulgas arvutilingvistika ja intellektitehnika, kus töötatakse välja suhtlusmudeleid ning kasutatakse neid küsimus–vastussüsteemide, automatiseeritud juhtimissüsteemide, ekspert-, masintõlke- jms süsteemide loomisel. Selliste süsteemide praktiline tähtsus kasvab järjest enam koos arvutite ulatusliku levikuga.[1]

Et tavaline arvutikasutaja võiks pöörduda arvuti poole loomulikus keeles ja saada selles keeles ka vastuseid, et ta saaks arvutiga suhelda nagu teise inimesega, tuleb arvutil modelleerida keelt kasutava inimese peas toimuvaid protsesse: partneri repliikide mõistmist, vastuse planeerimist ja ülesehitamist, adekvaatset reageerimist jt inimlike suhtlusnormide järgimist. Seega on, ühelt poolt, vaja modelleerida keele (teksti või kõne) analüüsi ja genereerimist, teiselt poolt aga suhtlusprotsessi ennast.

Eksisteerib mitmeid erinevaid lähenemisviise dialoogi modelleerimisele, sh dialoogigrammatikad, plaanipõhised mudelid, ühistegevuse teooriad. Dialoogimudelite praktilise realiseerimise populaarseteks valdkondadeks on reiside kavandamine, piletite reserveerimine, ülikooliastujate nõustamine jms (Cole jt 1997).

2001. aastal käivitus Eesti Teadusfondi projekt (nr 4555) eksperimentaalse dialoogsüsteemi loomiseks, mis suhtleks kasutajaga eesti keeles ning mille abil saaks hankida informatsiooni reisiplaneerimise, sh bussi-, rongi- jm liikluse kohta. Tööd alustasime olemasolevate dialoogimudelite ja dialoogikorpuste ning eestikeelsete dialoogitekstide analüüsist, et koostada realisatsiooni aluseks sobiv dialoogimudel. Dialoogide põhiliseks allikaks on olnud eesti suulise kõne korpus (http://psych.ut.ee/~linds), osa dialooge on hangitud ka

[1] Käesolevas artiklis käsitletud probleemide uurimine ja artikli kirjutamine on saanud võimalikuks tänu ETF grandile 4555.

arvutisimulatsioonidest (nn "võlur Ozi" meetodil). Dialoogides oleme asunud märgendama dialoogiakte, et ette valmistada dialoogsüsteemi väljatöötamiseks ja testimiseks vajalikku korpust. Järgnevas käsitlemegi selle töö käigus esilekerkinud probleeme.

1. Dialoogiaktide märgendamine

1.1. Mis on dialoogiakt?

Kui inimesed suhtlevad, siis nad teevad keele abil midagi: küsivad, vastavad, tervitavad, soovitavad jne. Selliseid keele abil läbiviidavaid tegevusi on nimetatud kõneaktideks, suhtlusaktideks, dialoogiaktideks, kõnetegevusteks jne. Ühesõnaga: aktid on tegevused, mida inimene kõne abil suhtluses teeb.

Kõneaktide uurimisega on aegade jooksul tegelnud erinevad lingvistikasuunad. Teoreetiliselt on kõneaktide mõistet enim käsitlenud John Searle (Searle 1969). Esimese tuntud aktitüpoloogia, mis toetus tegelikele dialoogidele, koostasid John Sinclair ja Malcholm Coulthard, kelle süsteemi aluseks oli inglise koolitundide analüüs (vt nt Sinclair, Coulthard 1975). Seda tüpoloogiat on hiljem modifitseerinud nt Anna-Brita Stenström (Stenström 1994). Viimasel aastakümnel on praktilise aktide määratlemisega tegelnud nt korpuslingvistid, diskursuseanalüütikud, vestlusanalüütikud, viimasel ajal ka keeletehnoloogid, keda on huvitanud arvuti ja inimese suhtlus loomulikus keeles (vt nt Hakulinen 1989; Allwood jt 2000; Stolcke jt 2000; Jokinen jt 2001).

Meie töörühma huvi on samasugune, kuigi suunatud kindlale dialoogitüübile: küsimus–vastus-dialoogidele. Meid huvitab, millised dialoogiaktid on seal leitavad, kuidas neid rühmitada ja kuidas nad on seotud keelega. Kaugemaks eesmärgiks on see, et arvuti suudaks keeleliste tunnuste abil aktid ära tunda ja saaks selliselt aru, mida inimene teeb, kui ta arvutiga suhtleb.

Selleks tuleb lahendada rida probleeme, mis on suuresti seotud loomuliku suulise kõne ja suulise dialoogi eripäraga. Vaatame lühidalt olulisi probleeme ja viitame sealjuures Tiit Hennoste ülevaatele eesti keele suulisest kõnest (Hennoste 2000–2001).

Esimene küsimus on, mis on akti kandev üksus suulises dialoogis? Selleks peame kõigepealt teadma, milliste üksuste abil inimesed suulises kõnes suhtlevad.

Suuline dialoog jaguneb kõnevoorudeks, mis on ühe kõneleja jätkuv häälesolek. Voorud liigenduvad vooruehitusüksusteks ehk lausungiteks, mille piirid on eelkõige intonatsioonilised, aga ka grammatilis-pragmaatilised (vt ülevaadet Hennoste 2000–2001: 2226–2236). Ideaalis peaks akti kandev üksus olema kõige väiksem keeleline üksus, mille abil saab midagi teha. Põhiliseks ühte akti kandvaks üksuseks on lausung, kuid see pole reegel. Ühte funktsiooni võib kanda ka osalausung või mitu lausungit kokku. Kuid ka fraas või isegi sõna (*ahah, jah* jms) võib olla omaette dialoogiakt. Seega ei saa väita, et on olemas mingi kindla formaalse ehitusega akti kandev üksus, vaid aktide kandjad ja nende keelelised realisatsioonid tuleb leida konkreetsete dialoogide analüüsi abil.

Teine küsimus on, kas üks üksus täidab ainult ühte funktsiooni. Suurem osa uurijaid on leppinud ühe aktiga. Kuid praktiliselt on see võimalik ainult siis, kui akti tõlgendada väga laialt. Vähegi konkreetsema aktimääratluse puhul aga täidab üks lausung tihti mitut rolli. Seetõttu on proovitud teha ka mitmeastmelisi aktisüsteeme. Nt Sinclair ja Coulthard ning nende järel Stenström on kasutanud kaheastmelist liigendust eristades samme (*move*) ja akte (*act*). Samm on see, mida kõneleja teeb vooruvahetuse läbiviimiseks (nt küsimine, kutsumine, jutustamine, tervitamine jne). Mille poolest erineb sellest akt, pole kuigi selge. Meie oleme praegu leppinud sellega, et sama üksus võib kanda mitut akti, kuid me ei süstematiseeri esialgu akte selle tunnuse järgi hierarhilistesse rühmadesse.

Kolmandaks, alati on võimalus, et kõneleja mõtleb lausungiga ühte akti ja kuulaja tõlgendab seda teise aktina (nt kõneleja võib esitada palve, mida kuulaja tõlgendab küsimusena). Millised aktid sellisel juhul märgitakse? Kas kavandatud akt või see akt, millena kuulaja asja tõlgendas, või mõlemad aktid? Praegu oleme märgendanud aktid sellena, mida kõneleja on meie arvates nendega mõelnud. Aluseks on arusaam, et kuulajapoolsest tõlgendusest lähtumine on analüüs tagantjärele. Tegelik vastaja peab aru saama konkreetse akti ja sellele eelneva jutu põhjal, millise aktiga on tegemist.

Neljas probleem tuleneb sellest, et suulisele suhtlusele on alati iseloomulik grammatiline, häälduslik, semantiline ja funktsionaalne varieerumine. See tähendab, et samal üksusel on olemas mitmeid häälduslikke variante jne (nt sõna *kakskümmend* eri variandid) (Hennoste 2000–2001: 1555–1561). Siin on vaja eraldada kaks asja. Esiteks peab olema selge, millisel juhul on sama grammatilise või

häälduslíku konstruktsiooni puhul tegu eri aktidega. Teiseks peab olema selge, millal on tegu sama akti realiseerumisega erinevates kontekstides, st sama akti variantidega. Kindel on sealjuures, et kontekstid ei ole reeglistatavad, vaid pigem statistilised tendentsid.

Viies probleem on, et dialoogis ühenduvad voorud kahel moel. Ühed voorud ühenduvad naaberpaarideks ja teised jäävad vabadeks voorudeks. Naaberpaari mõiste pärineb vestlusanalüüsist (vt Hennoste 2000–2001: 2236–2239). Naaberpaarid on voorupaarid, mille korral esiliikme olemasolu nõuab kindlat järelliiget (nt küsimus vastust, tervitus vastutervitust jne). Järelliikmed on omakorda liigendatavad ootuspärasteks ja ootusevastasteks, mis realiseeruvad üldjuhul keeleliselt erineval kujul. Meil on vaja kõigepealt leida naaberpaaride põhitüübid ja nende võimalikud konkreetsed variandid (nt küsimus–vastus moodustab ühe naaberpaari tüübi, küsimused, mis nõuavad erinevat vastamisviisi ja erinevad vastuseliigid moodustavad konkreetsed variandid).

Kuid naaberpaar on ainult ideaalis kahest voorust või kahest lausungist koosnev üksus. Temasse võivad liituda vahesekventsid (nt paranduslõigud), eelsekventsid ja järelsekventsid. Eelkõige vahesekventsid on tihti seotud olukorraga, kus vastajal puudub piisav info teise poole jutule reageerimiseks (nt küsitakse infotelefonist infot heade restoranide kohta, kuid andmebaasis on info liigendatud hoopis restoranide paiknemise, toidu tüübi jms põhjal.) Sellisel juhul peab vastaja enne vastamist täpsustama oma võimalusi vastata või pakkuma välja omapoolse andmebaasi liigenduse, mille alusel ta saab vastata.

Kuues küsimus on selles, et kõnevoorude vaheldamine ei ole aktidega üheselt seotud. Ideaalne on, kui konkreetne voor koosneb ühest lausungist, mis kannab ühte akti. Sellisel juhul on potentsiaalne vooruvahetuskoht kergesti aimatav ning suhtlus voolab lihtsalt. Kuid see pole reegel.

Esiteks, voorus võib olla mitu lausungit ja mitu akti ning arvuti peab ka siis aru saama, millal on voor lõppenud. Selleks tuleb eristada voorus paiknemise suhtes kolme liiki akte:

- aktid, mis alustavad vooru;
- aktid mis on vooru keskel, st ei alusta ega lõpeta;
- aktid, mis lõpetavad vooru.

Teiseks, pikema vooru kuulaja võib kasutada kahte liiki akte. Ühel juhul annab ta märku, et ta soovib ise vooru üle võtta. Kuid ta

võib kasutada ka tagasisidet või jätkamissignaali, mille abil ta osutab just vastupidist: et teine pool võib jätkata (vt potentsiaalsete vooruvahetuskohtade ning tagasisidepartiklite kohta Hennoste 2000–2001: 1787–1793, 2228–2236). Seega peab ka arvuti suutma vahet teha, kumma variandiga on tegemist.

Seitsmes raskus tuleb suhtluse põhimõttelisest mittesujuvusest. Suhtluse ideaaliks on tavaliselt peetud probleemideta, sujuvat suhtlemist. Kuid praktiline suhtlus näitab, et see on vaid väga harva võimalik. Seetõttu on vajalikud mehhanismid, mis lubavad tekkinud probleeme lahendada. Probleemid võivad sealjuures olla igasugused alates keelelisest parandusest kuni mittekuulmiseni.

Suhtlusprobleemide lahendamiseks on keeles olemas parandusmehhanism, mille saab jagada vastavalt sellele, kes algatab paranduse ja kes viib läbi ning vastavalt sellele, kas toimub edasilükkamine või ümbertegemine (vt Hennoste 2000–2001: 2689–2710; Strandson 2001). Arvuti jaoks on oluline see, et ta saaks aru nt inimese eneseparandusest. Kui inimene ütleb lause: *ma sooviks teada, millal lähevad bussid Tartust Tallinna, ei Pärnusse,* siis peab arvuti aru saama, et inimene küsib infot ainult Pärnusse minevate busside kohta. Sellised parandused on praegu jäetud siiski märgendamata. Küll aga on märgendatud parandused, mida suhtlejad teevad kahepeale. Sellisel juhul algatab üks pool paranduse tihti seetõttu, et ta ei saanud teise poole jutus millestki aru. Seega peab probleemi tekitaja oma juttu niimoodi modifitseerima, et teine asjast aru saaks.

Ja kaheksandaks, eelnevad probleemid olid nö mikrotasandi probleemid, mis tulevad välja konkreetsete voorude või naaberpaaride tasandil. Lisaks on aga vestlusel olemas ka oma makrostruktuur. Laias laastus jaguneb infodialoog alustuseks, lõpetuseks ja teema arenduseks. Ja kuigi paistab, et nt infotelefonis on tüüpiliseks ühe probleemi lahendamine ühe helistamise käigus, ei ole see siiski ainuvõimalik. Sama dialoog võib sisaldada ka mitu probleemi (nt bussiajad Tallinnast Tartusse ja Pärnusse vms). Algused ja lõpud on tihti seotud vormelite kasutusega ning moodustavad omaette rutiinse suhtlusosa (vt Rääbis 2000).

See aga tähendab, et tuleb leida need aktid ja nende realisatsioonid, mille abil ühelt dialoogi osalt teisele üle minnakse, et arvuti saaks aru, millal soovib partner nt vahetada teemat või hakata dialoogi lõpetama.

Praktiliselt on aktide tüpoloogia tegemine toimunud järgmiselt. Alguses vaatasime üle erinevad kättesaadavad aktide loendid. Selles töös osalesid lisaks käesoleva artikli autoritele ka Ann Must ja Riina Vellerind. Leitud aktidest tegi Tiit Hennoste koondloendi, täiendas ja täpsustas seda ning süstematiseeris aktid. Seejärel proovisime määratleda akte erinevates loomulikes dialoogides, mis on võetud TÜ suulise kõne korpusest (vt korpuse kohta allpool). Analüüsi tulemusel selgus üldjuhul, et iga uus dialoog tõi sisse mõne akti, mida loendis ei olnud, või sundis aktide tüpoloogiat ümber tegema.

1.2. Dialoogiaktide põhirühmad

Dialoogiaktide süstematiseeritud nimestik kujutab endast nn märgendusskeemi, mis on edaspidi aluseks dialoogides esinevate aktide märgendamisel. Meie märgendusskeemis on praegu 7 aktide rühma.

- Rituaalid, kuhu kuuluvad tervitamine, tänamine ja palumine, tutvustamine jms.
- Vestluse ümberstruktureerimise aktid on sellised, mille abil kõnelejad algatavad uusi teemasid ja vahetavad vestlustüüpi (nt lobisemisest tööalaseks nõupidamiseks).
- Vooruvahetust juhtivad aktid on sellised, mille abil kasutaja palub kõnelejal jätkata ja kontrollib suhtluskontakti olemasolu.
- Parandusaktid on sellised, mille abil partnerid lahendavad suhtluses ettetulevaid probleeme.
- Direktiivid on aktid, millega antakse edasi ja võetakse vastu korraldusi, soove jms.
- Küsimused on sellised aktipaarid, milles üks pool küsib ja teine vastab.
- Seisukohavõtud on aktid, millega üks pool esitab mingi seisukoha (arvamuse, hinnangu, süüdistuse jms) ja teine reageerib sellele.
- Muud aktid sisaldab ülejäänud aktide rühmi, kuhu kuuluvad nt info andmine, põhjendamine, järeldamine, lubadused jms. See on tegelikult aktide rühm, mis on süstematiseerimata.

Kõik aktirühmad peale viimase moodustavad naaberpaare ning on seetõttu jagatud kaheks pooleks: esiliikmeteks ja järelliikmeteks. Esiliikmete abil antakse käske, esitatakse küsimusi jms, järelliikmed esitavad reaktsioone käskudele, vastuseid küsimustele.

Skeemi tervikuna me siinkohal ära tooma ei hakka, sest see on liiga mahukas (vt ühte varianti Nurmsalu 2001).

2. Dialoogi analüüs: näide

Vaatame konkreetset dialoogi ja selles märgendatud dialoogiakte. Näitedialoog on võetud eesti suulise kõne korpusest.

1. ((kutsung)) | KUTSUNG |
2. V: ´Estmar=´info, | ESITLUS | | KUTSUNGI VASTUVÕTMINE | ´Leenu=kuuleb | ESITLUS |
3. V tere | TERVITUS |
4. H: tere. | TERVITUS |
5. (0.8) {Leenu.} | KONTAKTEERUMINE |
6. V: ja? | POSITIIVNE KONTAKT |
7. (0.5)
8. H: rotilõks. | ? SOOV |
9. (1.8)
10. V: kuidas? | ÜLEKÜSIMINE |
11. H: rotilõks. | PARANDUSE LÄBIVIIMINE |
12. (0.8)
13. V: jah, rotilõks | MITTEMÕISTMINE |
14. H: {´andke ´kõik.} | ? PALVE |
15. V: kuidas? | ÜLEKÜSIMINE |
16. H: {´kõik kus ma saan ´osta.} | PARANDUSE LÄBIVIIMINE | | ? SOOV |
17. (2.2)
18. V: e ma arvan et seda saab teha majapidamistarvete ´kauplustest. | ARVAMUS |
19. (0.5) saan teile neid ´pakkuda, | PAKKUMINE |
20. soovite. | JUTUSTAV KAS-KÜSIMUS |
21. (1.0)
22. H: {---} ´majapidamistarvete kauplusest. | ÜLEKÜSIMINE |
23. V: ma arvan ´küll jah. | PARANDUSE LÄBIVIIMINE |
24. (...)
25. H: mitte ´hiirelõksu. | ? PARANDUSE ALGATUS |
26. mul on kurat=ee noh (.) päris ´võikad ´elukad. | HINNANG |
27. (0.5)
28. V: jah ma ´usun. | SAMAMEELSUSE OSUTAMINE |
29. (0.5) m:a arvan et neid saab ka ´majapidamistarvete ´kauplusest. | ARVAMUS |
30. ma ei oska teile küll midagi ´muud ´pakkuda. | TEEMA LÕPETAMISE PAKKUMINE |
31. H: noh? | ? TEEMA LÕPETAMISE AKTSEPTEERIMINE | | ? KÄSK | | ? PALVE |
32. (0.8)
33. V: jaa üks hetk? | DIREKTIIVI VASTUSE EDASILÜKKAMINE |
34. (...) ee ´Meltoni ´äri. | INFO ANDMINE | | DIREKTIIVI TÄITMINE |
35. H: jah? | JÄTKAJA |
36. (0.5)
37. V: neli kolm null, (.) kaheksa viis kaheksa. | AVATUD VASTUS: INFO ANDMINE | | DIREKTIIVI TÄITMINE |

38. (1.2) ´Eritreid. | AVATUD VASTUS: INFO ANDMINE | | DIREKTIIVI TÄITMINE |
(1.0) neli neli üks, kaheksa üks null. | AVATUD VASTUS: INFO ANDMINE | | DIREKTIIVI TÄITMINE |
39. H: kaheksa üks null. | ÜLEKÜSIMINE |
40. V: jah? | PARANDUSE LÄBIVIIMINE |
41. H: jah | PARANDUSE HINDAMINE | | ?PALVE |
42. (1.2)
43. V: ja siis (...) e ´Kauburi tööstuskaubad. | DIREKTIIVI TÄITMINE |
44. (0.8) neli kaks null, (.) kaheksa null neli. | DIREKTIIVI TÄITMINE |
45. (...) Estiko Kom´merts=ä. | DIREKTIIVI TÄITMINE |
46. (.) neli kaheksa kuus, kaheksa kuus kolm. | DIREKTIIVI TÄITMINE |
47. (...) ´Räni kauplus. | DIREKTIIVI TÄITMINE |
48. H: jah | JÄTKAJA |
49. V: neli seitse seitse, (.) kaks kolm kolm. | DIREKTIIVI TÄITMINE |
50. (1.0)
51. H: kaks | ÜLEKÜSIMINE |
52. V: kaks kolm kolm. | PARANDUSE LÄBIVIIMINE |
53. H: ää=hh
54. V: ´Raadi kauplus, | DIREKTIIVI TÄITMINE |
55. (1.5) nelisada, kolm kolm kolm. | DIREKTIIVI TÄITMINE |
56. (...) ´Ristiku kauplus, | DIREKTIIVI TÄITMINE |
57. (1.0) neli seitse üks, (.) kolm viis kolm. | DIREKTIIVI TÄITMINE |
58. (1.0)
59. H: no aitab. | ? KÄSK |
60. (.)
61. V: jaa | DIREKTIIVI TÄITMINE |
62. =palun? | PALUN |
63. H: aitäh. | TÄNAN |

Toodud näites on tegemist on helistamisega infotelefonile, milles helistaja soovib saada infot kaupluste kohta, kust saab osta rotilõksu.

Dialoog algab standardse sissejuhatusega, milles info andja esitleb ennast ja oma institutsiooni ning suhtlejad tervitavad teineteist (read 2–4). Ebatavaline on siin see, et helistaja kasutab tervitamisel ka infoandja nime (rida 5).

Seejärel esitab helistaja oma soovi (rida 8). See on esitatud nii ebamääraselt, et info andja ei saa aru, millise aktiga on tegu. Seetõttu järgneb paranduste seeria, mille käigus püütakse tekkinud probleem ühiselt lahendada (read 10–15). Alles real 16 saab helistaja aru, millest mittearusaamise tekkis, teeb oma esialgsesse infosse paranduse ning esitab soovi uuesti.

Järgneb pikk paus, infotelefoni vastus arvamuse kujul ja pakkumine (rida 18). Vastamise viisist saab aru, et info andja jaoks on küsimus raske vastata ning ta püüab leida olukorrast väljapääsu.

Vastus tekitab omakorda probleemi helistajale, kes on veendunud, et majapidamistarvete poes ei ole rotilõkse, vaid ainult hiirelõksud. Seetõttu järgneb uus paranduste seeria ridadel 22–28. Vastaja jääb endale kindlaks (rida 29). Ühtlasi pakub ta välja selle alateema lõpetamise ning mineku tagasi konkreetse info andmise juurde (rida 30). Selle aktsepteerib helistaja ning ühtlasi esitab soovi info saamiseks (rida 31).

Sellega minnakse üle uude alaosasse, mille keskmeks on info andmine / direktiivi täitmine ja info vastuvõtmine, mida kuulaja teeb pidevalt korduva jätkamisakti abil (nt rida 35). Ka sellesse ritta sekkub üks probleem (rida 39), kui kuulaja küsib numbri üle ja vastaja kinnitab eelnevat infot.

See alaosa lõpeb real 59 direktiiviga, mis on ettepanek lõpetada antud teema. Teine pool aktsepteerib selle direktiivi. Järgneb vestluse lõpetamine ridadel 62–63.

Antud vestluses on hästi näha neli suurt erineva funktsiooniga osa, mis esinevad suuremas osas seda tüüpi dialoogides:

- rituaalne algus (read 2–6);
- rituaalne lõpp (read 62–63);
- suhtluses tekkinud probleemide (mittearusaamine, mittekuulmine, info ebausutavus jms) lahendamine vastastikuses koostöös (read 8–30 ja 39–41);
- soovitud info andmine (read 31–38 ja 43–61).

Lisaks on dialoogi märgendamise näha mitu probleemi.

Üks probleemirühm tekib sellest, et praegu pole me osanud teha vahet, millise direktiiviga on tegemist ja mis eri direktiive eristab. St millal on tegu nt käsuga, millal palvega, millal sooviga (vrd read 8, 14, 16, 31, 41, 59).

Teine keerukus tuleb sellest, et mõnikord on raske teha vahet küsimusel ja direktiivil. On võimalik, et arvuti ja inimese suhtluse jaoks ei ole see vahetegemine üldse oluline. Kuid selles saab kindel olla alles siis, kui oleme märgendanud enam dialooge.

3. Märgendustarkvara

Dialoogide märgendamine võib toimuda kas automaatselt, käsitsi või nende kahe viisi kombinatsioonis. Kuna käsitsi märgendamine nõuab palju aega ja automaatne märgendamine võib tuua kaasa liiga palju vigu, on parim moodus kombineeritud märgendamine. Seetõttu

otsustasime koostada programmi, mis oleks märgendamise abivahendiks.

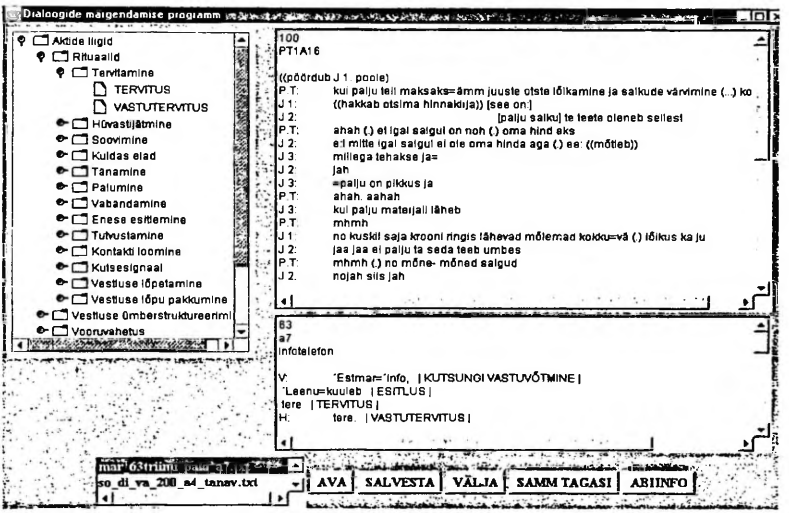
Selleks tuli esmalt tutvuda juba olemasolevate märgendamisskeemidega. Enim huvi pakkus Euroopa Liidu projekt MATE (*Multilevel Annotation Tools Engineering*, 1998–2000), mille raames on uuritud paljusid dialoogikorpuste märgendusskeeme. Projekti käigus analüüsiti rohkem kui 60 olemasolevat märgendusskeemi suuliste dialoogide annoteerimiseks erinevatel tasemetel (Dybkjær 2000).

Märgendusskeemide võrdlemiseks töötati välja hindamisjuhend, milles pöörati tähelepanu märgendamisjuhendi olemasolule, märgendajate arvule, dialoogide/lausungite/segmentide arvule, dialoogide teemale ja märgendavate nähtuste nimestikule. Samuti peeti oluliseks näidete olemasolu ja märgenduskeelt, mida kasutati nt märgendustöövahendis. Hindamisel uuriti ka skeemi kasutatavust ja võimalust kontakteeruda skeemi loojatega, et hankida lisainformatsiooni (Klein, Soria 1998). Enam kasutatavateks skeemideks osutusid Alparon, Chat, SWBD-DAMSL ja Verbmobil. Kuna SWBD-DAMSL on ainus skeem, mis pole konkreetsele ülesandele orienteeritud, siis sellest võib edaspidi kujuneda standard.

Teiseks etapiks meie märgendustarkvara loomisel oli programmi kirjutamine, milles kasutasime MATE demosüsteemi kujunduse elemente ning arvestasime ettepanekuid inimestelt, kes hakkavad selle programmiga edaspidi tööle.

Eestikeelse dialoogikorpuse märgendamisprogramm on kirjutatud keeles Java (Nurmsalu 2001). Peale programmi käivitamist kuvatakse ekraanile programmi tiitelaken, millest saab edasi liikuda põhiaknasse (joonis 1), kus toimub dialoogide märgendamine või vaadata abistavat informatsiooni programmi põhiakna ja aktide nimetuste kohta.

Märgendamise alustamiseks peab kasutaja valima esmalt faili (teksti), mida soovitakse märgendada. Dialoogiakti nimetuse lisamiseks märgendatavasse teksti peab märgendaja klõpsama tekstialal kohal, kuhu soovib paigutada akti nimetuse, ning seejärel valima vasakul asuvast dialoogiaktide puust sobiva akti. Kui soovitakse ühe lausungi järele panna kaks akti nimetust, tuleb esimese akti lõpus vajutada üks tühik ja seejärel valida teine, st uus akti nimetus. Valitavate aktide nimetused on kirjutatud suurtähtedega. Valitud dialoogiakti kustutamiseks tekstialal peab klõpsama nupul 'SAMM TAGASI'. Tekstialal saab kasutada ka kõiki tekstitoimetites kasutatavaid



Joonis 1. Programmi "Dialoogide märgendaja" põhiakna ekraanipilt

kustutamisviise. Alumisel tekstialal saab vaadata juba märgendatud faile. Faili valimiseks peab esmalt märgistama nimekirjas ühe failidest ning seejärel tegema selle faili nimel topeltklõpsu. Kui korraga on märgistatud mitu faili, siis ei ilmu uut faili tekstialale. Liigse märgistuse kaotamiseks peab lihtsalt üleliigse(te) faili(de) nime(de)l klõpsama. Märgendatud faili saab salvestada nupu 'SALVESTA' abil.

Praegu on käsil programmi testimine erinevate kasutajate peal ning täpsustamine vastavalt saadud tagasisidele.

4. Dialoogikorpus

Dialoogiaktide nimekirja täpsustamiseks, aktide analüüsiks ning dialoogimudeli väljatöötamiseks ja testimiseks on vajalik dialoogikorpus, mis sisaldaks suurel hulgal just antud ülesandele sobivaid tekste. Dialooge võib koguda mitmel viisil. Meie oleme kasutanud esialgu kahesuguseid tekste:

1) TÜ suulise kõne korpuse dialooge;
2) arvutisimulatsioonides "võlur Ozi meetodil" kogutud dialooge.

4.1. Dialoogid suulise kõne korpusest

TÜ suulise kõne korpus sisaldab praegu litereerituna ca 300 000 sõna mitmesuguseid tekste, mille tingimuseks on suulisus. Tekstide valikul dialoogikorpusesse on lähtutud järgmistest kriteeriumidest:

- lähtudes liigendusest dialoog/monoloog, on välja jäetud kõik monoloogid;
- teine piir on vahetu/vahendatud suhtlus. Tekstid saab jagada kolme rühma: silmast-silma vestlused, telefonivestlused ja massimeediatekstid. Dialoogikorpusesse on võetud vaid niisugused tekstid, milles on võimalik vahetu tagasiside. Seega on massimeediatekstid kõrvale jäetud;
- kolmas piir on argine/institutsionaalne suhtlus. Argivestlused on dialoogikorpusest välja jäetud (tekstide liigendamisest rühmadeks vt Hennoste jt 2000: 258–262).

Seega kuuluvad korpusesse põhiliselt institutsionaalsed dialoogid. Mõnel juhul on tegu polüloogiga, kui näiteks kliendiga suhtleb samaaegselt kaks ametnikku. Omaette sarja moodustavad teeküsimised tänaval, kus võõrad inimesed tegutsevad eraisikutena avalikus kohas – seega ei ole tegu institutsionaalse suhtlusega, kuna kumbki osaleja ei esinda institutsiooni. Kõik dialoogid on infovestlused, st suhtluse eesmärk ei ole suhtlemine ise, vaid kindla info hankimine.

Selliselt valitud dialoogikorpus sisaldab praegu 217 litereeritud teksti kogupikkusega 49 435 tekstisõna. Tekstide pikkus on 16–1702 sõna, keskmine pikkus 228 sõna. Dialoogikorpusesse kuulub 104 telefonikõnet ja 117 silmast-silma vestlust.

Telefonivestluste põhiosa moodustavad kõned, kus eraisik on helistanud mingisse ametiasutusse sooviga infot saada. Need võib kaheks jagada:

- helistatamine infotelefonile (küsida võib kõikvõimalikke asju);
- helistamine asutusse (nt reisibüroo, polikliinik, kauplus, bussijaam, koolitusfirma jt), kus küsimuste temaatika on piiratud.

Telefonikõnede teise, väga väikese rühma moodustavad kõned, milles omavahel suhtlevad kaks institutsiooni esindajat, näiteks helistatakse raadiost ilmajaama.

Silmast-silma vestlused jagunevad laias laastus kolmeks.

- Dialoogid kauplustes, teenindus- ja ametiasutustes (93 teksti). Siin suhtlevad omavahel müüja/teenindaja/ametnik ja klient.

- Teeküsimine tänaval (18 teksti). Suhtlejad on eraisikud.
- Küsitlused (2 teksti). Kontakti algataja ja info hankimisest huvitatud pool on ametiisik.

Dialoogid on reeglina litereeritud tervikuna. Litereerimisel on kasutatud konversatsioonianalüüsi transkriptsiooni. Iga teksti juurde kuulub taustakirjeldus, milles on andmed lindistamise, litereerimise, situatsiooni, osalejate, teema, tekstitüübi jms kohta (vt Hennoste 2000–2001: 1124–1133).

Praeguseks on sellest kogust märgendatud dialoogiaktide tasandil 20 dialoogi.

4.2. Dialoogide kogumine "võlur Ozi" meetodil

Osa dialooge on kogutud "võlur Ozi" meetodil: katseisikutel palutakse testida üht programmi, mis annab nende poolt esitatud küsimustele korrektseid ja informatiivseid vastuseid. Tegelikkuses on aga suhtluspartneriks arvutivõrgu vahendusel teine inimene (antud juhul Maret Kullasaar). Sellise nn simuleerimise teel kogutakse konkreetse valdkonna dialooge, mis eeldatavasti inimese ja arvuti vahelises suhtluses aset leida võivad, ja moodustatakse neist dialoogikorpus, mida hiljem kasutatakse dialoogsüsteemi programmeerimisel.

Mitmete uurijate varasemate kogemuste põhjal on kindlaks tehtud, et tõelähedasi dialooge on võimalik koguda vaid juhul, kui teema, mille kohta katseisik päringuid esitama peab, huvitab teda. Seetõttu oli ka meie katsete teemaks valitud reisimine, kuid seda mõningate piirangutega. Viimaseid tingisid eelkõige tõese info leidmise võimalused ja soov vastuste genereerimise aega lühendada. Liiga pikad pausid muudaksid katseisikud kahtlevateks sellise süsteemi võimalikkuses.

Dialoogide kogumisele eelnesid järgmised ettevalmistused (Kullasaar 2001).

- Võimalikele katseisikutele saadeti meil, kus paluti katsetada reisiinfot andvat programmi.
- Otsese kontakti puudumise tõttu oli nõusoleku korral järgmiseks sammuks soovitus läbi lugeda programmi kasutamisjuhend ja pakkuda ajad, millal võiks programmi testida. Kõik katsealused teadsid ette, et testitav programm tuleb enne katsetamist käivitada ja et testimise ajal viibib programmi koostaja oma arvuti taga. Kirjavahetuse käigus ilmnes, et paljud esmalt kõhklema hakanud inimesed

nõustusid kohe eksperimendis osalema, kui said teada, et ei pea midagi installeerima ja et programm ei vaja kiiret võrguühendust. Vajalik oli vaid veebilehitseja, mis toetab Java rakendusi ja veidi aega (keskmiselt kulus ühe dialoogi läbiviimiseks 20–25 minutit).

Üleskutsele reageeris 11 inimest. Katseisikuteks valitakse tavaliselt inimesed, kes pole süsteemiga tuttavad, kuid kellel on mõningaid teadmisi arutluse all oleva valdkonna kohta. See on oluline, sest katsete eesmärgiks pole süsteemiga töötamise õppimine /õpetamine, vaid antud valdkonnas võimalike suhtlussituatsioonide kogumine. Nii olid ka kirjeldatavasse eksperimenti kaasatud erineva vanuse, tegevusala ja arvutialase kogemusega inimesed. Kõigi katseisikute arvutioskus oli keskmine või sellest veidi parem. Seega ei vajanud ükski nendest arvuti kasutamisel kaasinimeste abi. Isikute keskmiseks vanuseks oli umbes 25 aastat.

Dialoogide läbiviimisel "võlur Ozi" meetodil on kasutatud mitmeid, erineva keerukusega programme. Suuremate projektide korral on selleks loodud kasutajasõbralikke ja testide läbiviija jaoks informatsioonirikkaid süsteeme. Meie eksperimendis kasutati dialoogide läbiviimiseks ja nende salvestamiseks programmi, mis töötab analoogiliselt telneti-laadse jututoaga: konkreetsel veebileheküljel on tekstiväli, millesse saab oma repliiki lisada ja partneri (katseisikute arvates arvuti) teksti lugeda. Informatsiooni esitajate eristamiseks kasutati nimesid *Arvuti* ja *Infoklient*, mille vastavad identifitseerimised viidi läbi programselt enne dialoogi algust.

Kokku on niiviisi kogutud 22 dialoogi, millest 1 jäi edasisest analüüsist välja, kuna katseisik polnud lugenud programmi kasutamisjuhendit ning seetõttu ei osanud dialoogi alustada. Samuti ei suhtunud ta eksperimenti tõsiselt ja käitus suhteliselt lapsikult, mis programmide testimisel/kasutamisel haruldane pole.

Kasutamisjuhend paiknes koos programmiga samal veebileheküljel, mis võimaldas igal hetkel abi saada.

Katsete lõppedes informeeriti katseisikuid asetleidnud simulatsioonist ning paluti neil täita ankeet, mis sisaldas mitut laadi küsimusi:

- testi läbiviimine (nt *Kas testimise käigus tekkis Teil kahtlus seesuguse programmi olemasolus?*);
- kasutatav keel (nt *Kas kasutasite lihtsamat keelt, arvates, et suhtlete arvutiga?*);

- arvuti poolt antud informatsioon (nt *Kas saite oma küsimustele ammendava vastuse?*);
- dialoogsüsteemi vajalikkus (nt *Kas Teie arvates oleks selline arvutisüsteem vajalik ning mis kasu temast Teile oleks?*).

Kuna selline eksperiment on teataval määral inimeste tüssamine ja usalduse ärakasutamine, siis sisaldas küsimustik ka märkuse: "Kui Teil on mingeid pretensioone sellise käitumise suhtes või Te ei soovi, et Teiega valminud tekstid oleks kasutatud, siis palun mind sellest teavitada. Arvestan kindlasti Teie soove. Igal juhul jääte anonüümseks, st üheski tekstis ei kasutata Teie nime." Ühelgi osalenul polnud mingeid pretensioone ning seega võis edasises töös kasutada kogu saadud materjali. Anonüümsuse tagamiseks kasutati kodeeringut, mida dekoreerida oskab vaid eksperimendi läbiviija.

Kõigis dialoogides märgendati dialoogiaktid, kasutades selleks eęspool kirjeldatud märgendamissüsteemi.

Vestlus	**Dialoogiaktid**
Infoklient: Kuidas sõita Tartus Pärnu enne kella 12 hommikul	AVATUD KÜSIMUS
Arvuti: Buss väljub kell 05.00	AVATUD VASTUS: INFO ANDMINE
Arvuti: Buss väljub kell 08.00	AVATUD VASTUS: INFO ANDMINE
Arvuti: Kas Teid huvitavad ka saabumisajad?	SULETUD KAS-KÜSIMUS
Infoklient: Jah	NÕUSTUV JAH
Arvuti: üks hetk, palun!	PALVE OODATA
Arvuti: Kell 05.00 saabub 08.20	AVATUD VASTUS: INFO ANDMINE
Arvuti: Kell 08.00 saabub 11.40	AVATUD VASTUS: INFO ANDMINE
Infoklient: Kas kella 12 ajal läheb ka praam Kuressaarde	SULETUD KAS-KÜSIMUS, TEEMA ALGATAMINE KÜSIMUSEGA
Arvuti: Buss väljub kell 08.20	INFO ANDMINE KÜSIMATA, TEEMA PAKKUMISE VASTUVÕTMINE
Infoklient: Kas Saaremaale saab kuidagi	SULETUD KAS-KÜSIMUS, TEEMA ALGATAMINE KÜSIMUSEGA
Arvuti: Mis peatusest Te sõita soovite?	TÄPSUSTAV KÜSIMUS, TEEMA PAKKUMISE VASTUVÕTMINE
Infoklient: Valjala	AVATUD VASTUS: INFO ANDMINE
Arvuti: Täpsustage sihtpeatuse nimi, palun!	VASTUSE TINGIMUSTE TÄPSUSTAMINE

Vestlus	**Dialoogiaktid**
Infoklient: Pärnu	PARANDUSE LÄBIVIIMINE
Arvuti: Selline reis meie andmebaasis puudub!	AVATUD VASTUS: INFO PUUDUMINE
Infoklient: Tahaks sõita pühapäeval Paidest Narva	AVATUD KÜSIMUS, TEEMA ALGATAMINE KÜSIMUSEGA,
Arvuti: Kas olete nõus Tallinnas ümber istuma?	TÄPSUSTAV KÜSIMUS, TEEMA PAKKUMISE VASTUVÕTMINE,
Infoklient: jah	NÕUSTUV JAH
Arvuti: üks hetk, palun!	PALVE OODATA
Arvuti: Buss Paidest väljub kell 10.35	AVATUD VASTUS: INFO ANDMINE
Arvuti: Buss Tallinnasse saabub kell 12.00	AVATUD VASTUS: INFO ANDMINE
Arvuti: Kas Te soovite lähimat bussi Tallinnast edasi liikumiseks?	SULETUD KAS-KÜSIMUS
Infoklient: jah	NÕUSTUV JAH
Arvuti: Buss Tallinnast väljub kell 13.00	AVATUD VASTUS: INFO ANDMINE
Arvuti: Buss Narvasse saabub kell 16.25	AVATUD VASTUS: INFO ANDMINE

Enamik kirjeldatud eksperimendis kogutud dialoogidest sisaldab rohkem kui üht teemat, sageli eelnevad vastusele täpsustavad küsimused või paranduste algatamised/läbiviimised. Samuti esineb situatsioone, kus "arvuti" annab rohkem informatsiooni, kui temalt küsiti (info andmine küsimata). Dialoogi kulgemise muutmiseks ning katseisiku reageerimise kontrollimiseks olid mõned "arvuti" vastused tahtlikult valed. Kahjuks enamikel sellistel juhtudel katseisikud jätkasid oma küsimusi või muutsid teemat. Polnud märgata soovi "arvutiga" vaidlema hakata: milleks – see on ju masin, mis ei saa niikuinii aru!

Töödeldud vestluste alusel koostati dialoogide kulgemist iseloomustav graaf, mis erineb suuliste dialoogide mudelist eelkõige selle poolest, et paljud aktid jäävad kirjalike vestluste puhul kasutamata. Näiteks pole meie eksperimendis kogutud dialoogides eneseesitlemist, tutvustust, kontakti loomist jms.

Arvutiklaviatuuri kaudu suhtlemise korral on täheldatud, et suhtlusvorm erineb tavalisest suulisest väljendusviisist. Kõige olulisemaks erinevuseks interaktiivse kirjaliku suhtlemise puhul peetakse keele kasutamisele seatud piiranguid: informatsioon püütakse optimaalselt "pakkida" võimalikult vähestesse ja lihtsatesse sõna-

desse. See aga muudab dialoogi struktuuri. Lisaks tekivad inimeste tüssamisel (st "arvuti" jäljendamisel) ka järgmised probleemid:

- inimesed on oma pöördumistes paindlikud, arvuti aga jäik;
- inimesed trükivad aeglaselt, arvuti aga väljastab kiiresti;
- arvutil ei teki iialgi väikesi eksimusi (nt juhuslikke õigekirjavigu), aga inimesed teevad neid pidevalt.

Teadmised selliste nn altminekukohtade üle sundisid arvutit jäljendaval inimesel oma sõnavara ning õigekirja eksperimendi käigus pidevalt kontrollima, mis ka õnnestus, nagu näitasid tulemused.

Suhtlemisel katseisikutega pärast katsete läbiviimist ilmnes, et see katse oli ka nende jaoks huvitav kogemus: mitmed said aru, et midagi on valesti, kuid mis, seda ei osatud öelda. Keegi ei kahtlustanud, et vastuseid otsib ja annab tavaline inimene. Arvati küll, et inimkäsi on dialoogide läbiviimisega seotud (nt aitab süsteemil vastuse leidmisel valikuid teha), kuid ei midagi enamat. Eksperimendi läbiviimine ja ankeedi vastuste analüüs kinnitasid, et sellised dialoogsüsteemid on inimestele tõesti vajalikud.

Kokkuvõte

Kõik siin kirjeldatud probleemid on seotud dialoogikorpuse loomise ja märgendamisega. Meie lähemaks eesmärgiks on testida, ühelt poolt, dialoogiaktide loendit, teiselt poolt aga märgendusprogrammi ja -juhendit. Sel otstarbel on kavas märgendada uusi dialooge, keskendudes reisiinfodialoogidele.

Kirjandus

Allwood, Jens; Ahlsen, Elisabeth; Björnberg, Maria; Nivre, Joakim 2001. Social activity and communication act-related coding. – Cothenburg Papers in Theoretical Linguistics 85. Dialog Coding – Function and Grammar. Göteborg Coding Schemas. Ed by Jens Allwood. Goteburg. 1–28.

Cole, R. A.; Mariani, J.; Uzskoreit, H.; Varile, G. B.; Zaenen, A.; Zampolli, A.; Zue V. (Eds.) 1997. Survey of the State of Art in Human Language Technology. Giardiani: Editori.

Dahlbäck N.; Jönsson, L.; Ahrenberg, L. 1993. Wizard of Oz studies – Why and how. – Knowledge-Based Systems 6:4, 258–266.

Dybkjær, L. 2000. MATE Deliverable D6.2. Final Report. – http://mate.nis.sdu.dk/about/deliverables.html (kasut. 19. 02. 2001).

Hakulinen, Auli 1989. Keskustelun luonnehtimisesta konteksti- ja funktionaalisten tekijöiden nojalla. – Kieli 4. Suomalaisen keskustelun keinoja 1. Toim. Auli Hakulinen. Helsingin yliopiston suomen kielen laitos. Helsinki. 41– 72

Hennoste, Tiit 2000–2001. Sissejuhatus suulisesse eesti keelde I–IX. Akadeemia 2000: 5, 1117–1150; 6, 1343–1374; 7, 1553–1582; 8, 1773–1806; 9, 2011–2038; 10, 2223–2254; 11, 2465–2486; 12, 2689–2710; 2001: 1, 179–206.

Hennoste, Tiit; Lindström, Liina; Rääbis, Andriela; Toomet, Piret; Vellerind, Riina 2000. Eesti suulise kõne korpus ja mõne allkeele võrdlemise katse. – Arvutuslingvistikalt inimesele. Tartu ülikooli üldkeeleteaduse õppetooli toimetised 1. Toim. T. Hennoste. Tartu. 245–284.

Jokinen, Kristiina; Hurtig, Topi; Hynnä, Kevin; Kanto, Kari; Kaipainen, Mauri; Kermanen, Antti 2001. Self-Organizing Dialogue Management. Käsikiri

Klein, M.; Soria, C. 1998. MATE Deliverable D1.1. Supported Coding Schemes. Dialogue Acts. – http://www.dfki.de/mate/d11/chap4.html (kasut. 7. 02. 2001).

Kullasaar, Maret 2001. Eestikeelse dialoogikorpuse arendamine "võlur Ozi" tehnikaga. Magistritöö. TÜ arvutiteaduse instituut.

Mengel, A.; Dybkjær, L.; Garrido, J. M.; Heid, U.; Klein, M.; Pirrelli, V.; Poesio, M.; Quazza, S.; Schiffrin, A.; Soria, C. 2000. MATE Dialogue Annotation Guidelines. Dialogue acts. – http://www.ims.uni-stuttgart.de/projekte/mate/mdag/da/da_1.html (kasut. 9. 02. 2001).

Nurmsalu, Evely 2001. Eestikeelse dialoogikorpuse märgendustarkvara. Magistritöö. TÜ arvutiteaduse instituut.

Rääbis, Andriela 2000. Telefonivestluste sissejuhatus. – Keel ja Kirjandus 6, 409–424.

Searle, John 1969. Speech Acts. Cambridge: Cambridge UP.

Sinclair, J.M.; Coulthard, R.M. 1975. Towards of Analysis of Discourse: The English used by Teachers and Pupils. London: Oxford UP.

Stenström, Anna-Brita 1994. An Introduction to Spoken Interaction. London and New York: Longman.

Stolcke, Andreas; Coccaro, Noah; Bates, Rebecca; Taylor, Paul; Van Ess-Dykema, Carol; Ries, Klaus; Shriberg, Elizabeth; Jurafsky, Daniel; Martin, Rachel; Meteer, Marie 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. – Computaional Linguistics 26:3, 339–373.

Strandson, Krista 2001. Kuidas vestluskaaslane parandusprotsessi algatab. – Keel ja Kirjandus 6, 394–409.

Suuline kõne ja morfoloogiaanalüsaator

Tiit Hennoste, Liina Lindström, Olga Gerassimenko, Airi Jansons, Andriela Rääbis, Krista Strandson, Piret Toomet, Riina Vellerind
Tartu ülikool

Käesolev artikkel annab esmase ülevaate kirjakeele morfoloogia-analüsaatori kasutamisest eesti keele suuliste dialoogide märgendamiseks. Morfoloogiaanalüsaator on arvutiprogramm, mis teeb sõnade automaatset morfoloogilist analüüsi, mille käigus sõnad saavad sõnaliigi nime ja andmed muude grammatiliste kategooriate kohta, mis antud sõnavormiga seostuvad (kääne, arv, aeg jms).

Eesti keele jaoks on olemas morfoloogiaanalüsaator, mis on tehtud kirjalike tekstide, kitsamalt kirjakeelsete tekstide tarvis. Selle autorid on Heiki-Jaan Kaalep ja Tarmo Vaino (vt Kaalep 1998; Kaalep, Vaino 2000). Selles on kaks osa: esmalt tehakse üksiksõnade morfoloogiline analüüs, mis annab paljudele sõnadele mitu morfoloogilist tõlgendust ja seejärel ühestamine, mille abil valitakse neist tõlgendustest välja üks, mis sobib kõige paremini antud konteksti. Analüsaator annab kirjalike tekstide korral korrektsed analüüsid ca 98% sõnadele (Kaalep, Vaino 2000: 91).

Meid huvitab, et antud morfoloogiaanalüsaatoriga saaks automaatselt analüüsida ka suulise kõne tekste. On selge, et suuline kõne erineb kirjalikust paljude parameetrite poolest, mida praegune analüsaator arvesse ei võta ega peagi võtma. Seega on vaja analüsaatorit muuta, et ta tuleks toime ka suulise kõne erijoontega. Selleks tuleb esmalt kindlaks teha, millega analüsaator praegu hakkama ei saa ja miks. Selle kindlakstegemiseks tegime lihtsa katse.[1]

Valisime TÜ suulise kõne korpusest 5804 sõnalise tekstikogumi, mis koosnes tavalistest argivestlustest (korpuse kohta vt http://psych.ut.ee/~linds; Hennoste *et al* 2000). Eemaldasime transkriptsioonimärgid ning jätsime alles vaid intonatsioonilisi piire markeerivad punktid ja komad (vt intonatsiooniliste üksuste ehk lausungite kohta Hennoste 2000–2001: 2226–2236). Samuti jätsime välja

[1] Katse arvutitehniline töö on tehtud Liina Lindströmi poolt.

transkriptsioonides esinevad kommentaarid. Lasime selliselt puhastatud tekstid morfoloogiaanalüsaatorist ilma ühestamata läbi. Seejärel kontrollisime tulemused käsitsi üle. Alguses vaatas iga inimene üle ühe katkendi analüüsitud tekstidest. Seejärel arutasime läbi probleeme tekitanud kohad ja ühtlustasime vastuolulised määrangud. Tulemused on erinevad olenevalt sellest, kuidas tõlgendada partikleid (vt nende kohta allpool). Kui leppida sellega, et partiklid on määratud nii, nagu praeguses sõnaliigituses tavaks, siis on vääraid märgendeid vaid 7%. Kui aga arvestada ka partikleid, siis on vääraid lahendeid 25% sõnadest. Märgendamisel esile tulnud probleemid saame jagada 4 põhirühma:

- sõnaliikide probleemid;
- suulise kõne varieerumise probleemid;
- suulise kõne erisõnade äratundmine;
- tehnilistel põhjustel analüüsimatud üksused.

1. Sõnaliikide probleemid

Kõige olulisema probleemirühma moodustab sõnade jagamine sõnaliikidesse. Sellel on omakorda kaks alaprobleemi: sõnaliikide valik ja liikide piiridele jäävate nähtuste märgendamine.

Esimene küsimus on, millistesse liikidesse üldse sõnu jagada. Eesti keele morfoloogiaanalüsaator kasutab 9 traditsioonilist sõnaliiki. Eesti keele teaduslik grammatika jagab sõnad 12 liiki (EKG I 1995: 18). Samas on uurimused näidanud, et kogu traditsiooniline sõnaliigitus mõnes kohas suulise teksti jaoks ei sobi. Siin on praeguse analüüsi põhjal kolm probleemi: partiklite, küsisõnade ja artiklite probleem.

Suulises kõnes on väga laialt levinud sõnarühm, mida nimetatakse pragmaatilisteks partikliteks, diskursusepartikliteks, insertideks jms (vt Stenström 1990; Sorjonen 1999; Longman 1999: 93–94, 1082–1099; Hakulinen 2000; eesti keele kohta vt Hennoste 2000–2001: 1773–1806; Strandson 2001). Need on muutumatud ja tavaliselt semantiliselt iseseisva tähenduseta sõnad, mis aga ei kuulu lausesse iseseisvate ega abisõnadena, vaid on süntaktiliselt sõltumatud (partiklid on nt *jah, ahah, noh, no, mhmh, vä, ah, oi, okei, eks, ee, õõ*). Sellist sõnaliiki või sõnaliikide rühma eesti sõnaliikide taksonoomias praegu ei ole. Osa sellesse rühma kuuluvatest sõnadest on jagatud interjektsioonide ja rõhumäärsõnade alla, osa aga puuduvad üldse, kuna neid kirjalikes tekstides praktiliselt ei esine. Samas on

selge, et suulise kõne analüüsi ilma seda sõnarühma välja toomata mõtet teha ei ole. Nt uued inglise ja soome grammatikad on toonud partiklid (Longmanil inserdid) juba sõnaliikidesse sisse (vt Longman 1999: 93–94, 1082–1099; Hakulinen 2000; partiklite kui sõnaliigi probleemi kohta vt Hennoste ilmumas). Partiklid analüüsis analüsaator kahel viisil:

- need partiklid, mis olid analüsaatorile tuttavad teiste sõnaliikide all (nt interjektsioonid, adverbid, sidesõnad), analüüsis ta vastavalt talle antud sõnaliigimääratlusele (jah+O//_D_//; noh+O //_D_//; et+O //_J_//; hehe+O // _I_//);
- need partiklid, mis olid analüsaatorile varasemast tundmatud, jättis ta analüüsimata või märkis nimisõnadeks (nooh+O //_S_sg n, //; tsau+O //_S_sg n, //; sis +O //_S_sg n, //).

Sealjuures tekkis kummalisi analüüse, mille alused pole päris selged. Nt partikli *noh* analüüsis arvuti adverbiks, *no* interjektsiooniks ja sõna *nooh* ei tundnud ta üldse. Sealjuures annab nt Kirjakeele seletussõnaraamat *noh* ja *nooh* esmaseks sõnaliigiks interjektsiooni ja *no* puhul adverbi (EKS III:4, lk 687, 689).

Praegu otsisime partiklid välja ja panime neil uue märgendi B. Selle märgendi sai 1045 sõna, ehk 18% kõigist tekstisõnadest. Kõigist meie poolt vigadeks analüüsitud sõnadest (kokku 1454) moodustasid partiklid 72%. Seega on partiklite rühm põhiline morfoloogilise analüüsi vigade andja. Ja kuna ilma partikli kategooriat välja toomata muutub suulise kõne edasine süntaktiline analüüs mõttetuks, siis on esimene vajadus viia analüsaatorisse partiklid kui iseseisev sõnaliik või -rühm. Partiklid moodustavad suures osas suletud sõnakogumi, st neid on võimalik esitada analüsaatoris loendina. Seejuures tuleb teha ringi mõne senise sõnaliigi loendid. Ja alles hiljem on võimalik lahendada partiklite, adverbide ja sidesõnade homonüümia probleem.

Teise sõnaliigilise probleemi moodustavad küsisõnad. Need on praeguses eesti sõnaliigituses paigutatud adverbide ja asesõnade hulka. See võib olla piisavalt hea lahendus kirjaliku teksti jaoks, kus küsimine ja seega küsisõna on küllalt harv nähtus. Kuid suuline kõne on enamasti dialoog ja seetõttu on küsimine seal palju tavalisem. Küsisõnade väljaeraldamine on ka dialoogi edasise analüüsi jaoks väga vajalik. Eeskujuks saab sealjuures võtta inglise keele, kus küsisõnad on toodud välja funktsionaalsete sõnade alarühmama (Longman 1999: 87–88).

Lisaks tekib suulise kõne küsisõnade puhul praeguste sõnaliikide juures rida probleeme. Nt on suhtluses tavaline küsisõna *vä*. Seda oleme praegu käsitlenud partiklina. Kuid temaga koos ja samas funktsioonis töötab teatud kontekstides ka sõna *kas*, mida on liigitatud määrsõnaks, ja sõna *mis*, mis kuulub asesõnade hulka (vt Hennoste 2000–2001: 1797–1798; Lindström 2001).

Kolmas sõnaliigiprobleem on artikli olemasolu või puudumine eesti keeles. Normikeele seisukohast eesti keeles määravat ja määramatut artiklit ei ole. Kuid suulises kõnes tegutsevad sõnad *see*, *mingi* ja *üks* mõnikord üsna sarnaselt nt inglise keele artiklitega (vt Pajusalu 1997: 2000). Praegu on nad liigitatud asesõnadeks, kuid süntaktilise analüüsi jaoks ei saa selliseid sõnu panna asesõna ega ka partikli kategooria alla.

Teise suure probleemi moodustavad sõnad, mis paiknevad sõnaliikide piiril.

Kas lingvistiliste kategooriate piirid peavad olema aredad või mitte, on igiammune vaidlusprobleem. Ühe lähenemise järgi on nad teravad, kuid meil puuduvad lihtsalt teadmised, et piiri korrektselt paigale panna. Teise järgi on piirialadele omane põhimõtteline hämarus ja üleminekuvormid, st keel on ehitatud tsentri–perifeeria põhimõttel. Praktilisest seisukohast ei ole neil kahel väitel vahet. Kummalgi juhul pole meil praegu võimalik määratleda mingi sõna X kuuluvust ainuliselt. Kui me seda teeme, siis me otsustame tegelikult ilma detailiuuringuteta asju, mida hiljem tuleb kardetavasti ringi teha. Siin tekib kolm eraldi probleemi.

Esimene probleem tekib seoses sellega, et suulise kõne uurijad ei pruugi alati olla teoreetiliselt sama meelt uurijatega, kes paigutasid kirjakeele korpuse põhjal mingi sõna mingisse rühma. Ühe probleemse rühma moodustavad nt sõnad *siuke*, *selline*, *sihuke* jms, mida võib käsitleda nii asesõnadena nagu on teinud morfoloogiaanalüsaatori loojad, kui ka (ase)omadusõnadena, nagu teeksime meie. Teise rühma moodustab nt *kätte, käes, käest* jms sõnade liigitamine, mida kirjakeele analüsaatori loojad on pannud kohati nimisõnadeks seal, kus meie seda teha ei sooviks (vt Kaalep *et al* 2000: 628).

Teiseks, kirjalike tekstide põhjal leitud sõnaliigipiirid ei pruugi lihtsalt kehtida suulistes tekstides.

Ja kolmandaks, suulise kõne piirialad on veel udusemad kui kirjalikes tekstides. See tuleb hästi välja nt partiklite, sidesõnade ja adverbide piiril. Osa sõnu töötab mõnel juhul partiklitena, mõnel

juhul aga adverbide või sidesõnadena ja piir nende vahel on väga udune (vt nt Keevallik 2001 *o(o)ta* kohta). Formaalseid tunnuseid valiku tegemiseks on vähe ja osa juhtumeid jääbki praeguse analüüsitaseme juures mitme interpretatsiooni vahele.

Praegu oleme määratlenud nt adverbi ja partikli piirjuhtumeid lihtsalt konkreetsete tekstikohtade ühise analüüsi põhjal. Sidesõnade ja piiripartiklite (ehk siduvate partiklite) vahel aga tegime vahet selle järgi, kas nad eraldavad üksusi sama lause/lausungi sees või paiknevad lausungite või voorude piiridel (tavaliselt alguses). Viimasel juhul töötavad nad partiklitena. Samasugust loogikat on järginud ka nt valmiv soome akadeemiline grammatika.

Omaette probleemiks on see, et osa partikleid on tegelikult mõne teise sõnaliigi tavalise sõna vormid. Nii tegutsevad ka partiklitena nt *kuule, vaata, oota, tähendab* ja mitte ainult nende lühenenud vormid.

Aga see pole ainult muutumatute sõnade probleem, kus piirid sõnaliikide vahel on ka normikeeles üsna udused. Suulises kõnes liiguvad ka muutuvad täistähenduslikud sõnad sõnaliikide vahel palju vabamalt kui normikeeles. Seetõttu kasutatakse seal nt omadus- või määrsõnadena sõnu, mida normikeeles ei kasutata (*kirves hind* 'kallis hind'). Omaette suure rühma moodustavad nt sõnad, mille tähendus on laias laastus *väga/eriti* (*hirmus, jõle, jube, ilgelt* jms).

Sõnaliikide piiril paiknevad sõnad ei ole loomulikult suulise kõne eriprobleem. Samade küsimuste ees seisid ka morfoloogia-analüsaatori tegijad kirjalike tekstide analüüsil. Üks konkreetne näide on *aga*, mida sõnaliigiliselt on traditsiooniliselt tõlgendatud side- või määrsõnana. Suulises kõnes kasutatakse teda lisaks ka partiklina (Hennoste 2000–2001: 1804, 2469–2472). Võib aga näha, et see pole ainult suulise kõne omapära. Kaalepi *et al* toodud näited *aga* kohta, kus see on lause või osalause keskel (*Hetke pärast andis kõrvetus maos aga järele* jm) on just juhtumid, kus *aga* on tegelikult partikkel (Kaalep *et al* 2000: 630–631).

Siit tuleb oluline küsimus, kas morfoloogiaanalüsaator peab alati otsustama, millisesse sõnaliiki sõna kuulub. Nii on morfoloogia-analüsaatori tegijad leidnud, et küsimus kas *nud*-partitsiip on verb või omadussõna, on vaja lahendada ilmtingimata morfoloogilises analüüsis, mitte lükata seda süntaktilisse analüüsi nagu on tehtud inglise keele kitsenduste grammatikas. Meie arusaam on vastupidine: katsed morfoloogiaanalüsaatori abil maksimaalselt kõiki probleeme

lahendada on kahjulikud. Keerukate üleminekualade tõttu sõnaliigituses oleks ilmselt kasulik, kui analüsaator jätakski alles kaks või enamgi analüüsivarianti (vähemalt mingiks ajaks) ega püüaks ilmtingimata neist ainult ühte välja valida.

Keskne põhjus sellise seisukoha eelistamiseks on asjaolu, et sõnade liigitus ei ole pelgalt morfoloogiline probleem. Sõnu jagatakse liikidesse morfoloogia, süntaksi ja semantika alusel. Ühte liiki pannakse sõnad, mis muutuvad ühtmoodi (pöördsõnad, käändsõnad, muutumatud sõnad), omavad samatüübilisi tähendusi (täistähenduslikud ja mittetäistähenduslikud sõnad) ja kannavad lauses samu funktsioone (iseseisvad sõnad ja abisõnad) (vt EKG I 1995: 14–17). Ja kuna morfoloogiline analüüs on esimene, mida arvuti teeb, siis määrab saadud märgend ette ära ka väga palju sellest, mismoodi saab sõnu edaspidi analüüsida nt süntaksianalüsaatoris. Eriti hulluks võib asi minna suulise kõne puhul, kus piirid sõnaliikide ja sõna funktsioonde vahel on hajusamad kui normkirjakeeles ja sõnade kasutust on vähe uuritud.

2. Varieerumine

Teise suure probleemirühma moodustab sõnade vormi varieerumine. Kui kirjakeeles on sõnade häälikulised ja morfoloogilised variandid viidud miinimumini ja üldjuhul küllalt rangelt reeglistatud, siis suulises kõnes kasutatakse väga palju erinevaid hääldusliikke ja morfoloogilisi variante, mille puhul saame öelda, et tegu on sama sõnaliigi, sama tähenduse ja sama morfoloogilise vormiga, mis aga realiseeruvad erinevatel (kon)situatiivsetel tingimustel. Siin on välja toodavad mitu alarühma.

Suulises kõnes hääldatakse sõnu lahku ja kokku üldjuhul nagu normikeeleski, kuid see pole reegel. Normikeele liitsõna võidakse hääldada kaheks sõnaks ja normikeeles mitu sõna moodustavaid ridu võib hääldada kokku (*ma=i=tea*). Siin on probleemiks see, kas tõlgendada neid ridu ühe sõnana või mitme sõna ühendina. Praegu on sellised kogumid tõlgendatud eraldi sõnadeks ja kokkuhäälduse märgid asendatud enne analüüsi tühikutega (nt *i* 'ei' võetud eraldi sõnaks, mida arvuti ei osanud loomulikult analüüsida).

Sõnad võivad suulises kõnes lüheneda (*kakskend, vaatsin*), vahetada häälikuid (*präegu, tian*), olla tugevamalt hääldatud (*tulep*). Lühenenuna esineb sealjuures rida tavalisi verbe (*ütsin, mõtsin* jms, mida kasutatakse küllalt kindlas kontekstis, otsese kõne saatelauses).

Aga kõnes võivad olla kasutusel ka sõnade murdevariandid (*latse-latselatsed, umale*), mõne inimese isiklikud idiolektierisused jms (vt ka Hennoste 2000–2001: 1555–1565).

Üldjuhul analüsaator selliseid variante ära ei tunne, analüüsides need nimisõnaks (pial+O //_S_ sg n,//) või jättes ilma analüüsita juhul, kui sõna on lühenenud selliseks häälikukombinatsiooniks, mida tema arvates olemas olla ei saa (*nd>nüüd*). Mõnikord suudab ta küll ära tunda lühenenud sõna, kuid ei suuda määrata selle morfoloogilist vormi. Nii on juhtunud mõne verbiga, mille puhul analüsaator tunneb ära, et tegu on verbiga, kuid määrab lühenenud minevikuvormi olevikuks (nt sõna *rääksin* pakub ta olevikuvormiks). Siin on tegu ilmselt sõnaliigi juhusliku äratundmisega, sest nt analoogilise sõnavormi *vaatsme* analüüsib arvuti nimisõnaks.

Omaette probleemi moodustab *h* hääldamine sõna algul. Nimelt ei ole kirjas esineva sõnaalgulise *h* hääldamises eesti keeles mingit ranget süsteemi. Mõnikord teda häältdatakse ja mõnikord mitte. Siin on välja toodavad ainult sotsiolingvistilised mõjurid, mis näitavad statistilisi erinevusi (vt Cui 1999). Praegu ei saa analüsaator enamasti ilma *h*-ta häältdatud sõnadest aru, mõned aga analüüsib korrektselt (nt *oiduda*). See probleem on lahendatav ilmselt lisades analüsaatorisse võimaluse iga *h*-algulist sõna ka ilma *h*-ta hääldada, kusjuures sõna vorm ega tähendus sellest ei muutu.

Eraldi probleemirühma moodustab morfoloogiliste tunnuste ja lõppude varieerumine.

Praeguses korpuses on kõige tavalisem ja ainus tõeliselt sage erijoon *nd*-partitsiip (*läind, näind, armastand, õppindki, tuld*), mille kasutamisel on omad piirangud ja statistilised erinevused (vt Keevallik 1996). Seda vormi morfoloogiaanalüsaator ära ei tunne. Selle probleemi lahendus on ilmselt lihtne: on vaja lihtsalt lisada analüsaatorile see võimalus.

Analüsaator ei tunne ära ka ebatavalisi tunnuseid ja lõppe, nagu *si*-mitmuse partitiivi seal, kus see pole normingukohane (*nastoikasi*). Ka ei tunne ta ära selliseid vorme, mis on normikeeles kasutusel, kuigi halvaks stiiliks peetuna. Siia kuuluvad nt vormid *kellegile* ja *kuhugile*.

Veel keerukam on lugu selliste suulise kõne morfoloogiliste vormidega, mis langevad kokku mõne teise kategooria vormiga. Üks selline probleemne rühm on kaudne kõneviis, mis normikeeles kannab tunnust *-vat*. Suulises kõnes aga kasutatakse samas tähenduses

kuut erinevat võimalust (vt Toomet 2000). Ja kuigi ka siin saab tuua välja mõned tingimused, millal millist vormi kasutatakse, ei ole homonüümsete vormide eraldamine siiski morfanalüüsis ilmselt võimalik. Nt kui kaudse kõneviisina on kasutatud enneminevikku, siis ei piisa abiks ka süntaktilisest infost, vaid vajalik on abistav info terve lõigu kohta.

Ja lõpuks, lisaks regulaarsetele variantidele on suulises kõnes, eriti argivestlustes olemas ühekordsed nähtused, mille kohta ei saa midagi üldistavat öelda (nt keelevääratused, nalja pärast valesti rääkimine, idiolektides esinevad kinnistunud häälduslikud vead võõrsõnades jms). Need jäävadki ilmselt määramata.

Üldiselt on meie korpuse keel morfoloogiliselt vägagi normikeelepärane, kuid see johtub informantidest, kelle hulgas on ülekaalus nooremad kesk- ja kõrgharidusega naised, kes on sotsiolingvistiliselt väga normikeelsed (vt selle kohta Hudson 1996: 193–199). On ilmne, et morfoloogiliste variantide hulk korpuse kasvades tõuseb. Morfoloogilise analüüsi jaoks on siin kaks suurt probleemi.

Esiteks, kuigi on leitav teatud süsteem varieerumises (nt on teada mõned positsioonid, kus lühenemine toimub statistiliselt sagedamini kui mujal, on teada, millistes tekstides eelistatakse *nd*- ja millistes *nud*-partitsiipi jms), ei ole tegu reeglitega vaid statistiliste erinevustega. Teiseks ei ole ennustatav, millised variandid korpusesse sisse tulevad. Nt traditsiooniliste murrete morfoloogia kasutamist ühiskeelses tekstis ei ole võimalik prognoosida.

Teiseks, on selge, et analüsaatorisse ei ole võimalik paigutada kõiki eesti keele võimalikke morfoloogilisi variante, sest nende hulk läheb liiga suureks. Seetõttu vajab morfoloogiaanalüsaator selliste kaldeliste variantide äratundmiseks ilmselt mingit eraldi oletajat.

3. Suulise kõne erisõnad

Kolmanda probleemirühma moodustavad sõnad, mida kirjalik keel ja kitsamalt normikeel ei tunne.

Suulises kõnes esineb lisaks klassikalistele sõnadele ka selliseid foneetilisi järjendeid, mida normikeele uurijad ei soovi sõnadeks pidada, sest nende foneetiline ja /või fonoloogiline struktuur ei vasta eesti keele normidele. Siia rühma kuulub rida interjektsioone, aga ka üneemid (*ee, õõ* jms, vt Hennoste 2000–2001: 1565–1568). Suulises kõnes on neil suhteliselt kindlad funktsioonid ja nad ei erine millegi

poolest sõnadest. Oma funkstioonidelt kuuluvad nad partiklite rühma. Ka nende esitamine on ilmselt lahendatav loendina.

Suulises kõnes, eriti argivestluses kasutatakse sõnu, mille moodustusmallid kirjakeeles puuduvad. Sõnu lühendatakse raide abil ja mõnikord lisatakse lühendatud tüvedele ka liiteid, kusjuures sõnade tähendust ei muudeta (*radiaator>radikas, ema>emps, mobiiltelefon>mobla* jms). Selliseid sõnu on avatud hulk ja neid võib lõputult juurde toota (vt Hennoste 2000–2001: 1369–1374). Analüsaator märkis sellised sõnad küll substantiivideks, kuid siin on raske otsustada, kas tegu oli korrektse analüüsiga või sellega, et analüsaator märgib endale tundmatud sõnad üldiselt substantiivideks.

Omaette probeemi moodustavad mitmesugused argisõnad, mida analüsaator ei oska ära tunda (*kammoon, kööme* jms). Ka siin on tegu avatud rühmaga, mida pole võimalik ennustada. Lisaks, kuna põhiosa neist on nimisõnad, on siin lisaks ka morfoloogilise analüüsi probleemid kitsamas mõttes (millise vormiga on tegu).

4. Analüüsimatud üksused

Neljanda probleemirühma moodustavad üksused, mis on tehnilistel põhjustel analüüsimatud.

Esimeseks rühmaks on pooleli jäävad sõnad, mida mõnikord küll sõnaliigiliselt saab analüüsida (nt kui sõna lõpust on jäänud ära üksnes käändelõpp: *nõrkushetk-*), kuid mis oleks mõttekas esitada siiski kui analüüsimatud, sest poolelijäämisel ei paista olevat mingeid reegleid. Lisaks võib sõna poolelijäänud osa olla samane mõne teise sõnaga, tekitades nii lisasegadust. Nii pani analüsaator poolelijäänud sõna *kah-* adverbide alla, kuna adverb *kah* on tal mälus ilmselt olemas.

Teise rühma moodustavad poolelijäänud laused või lauseosad, mille puhul ei ole mõnikord selge, millises tähenduses ja seega millise sõnaliigina on abisõnu kasutatud.

Väga keerukas on ka sõnade probleem, mille kuuluvuse kohta on raske midagi öelda, sest nad paiknevad lausungis imelikus kohas isegi suulise kõne jaoks (*ma ja juba....*). Siin pole *ja* puhul tegu sidendiga ja seega mitte ka sidesõnaga, kuid ka partikliks on seda raske analüüsida.

Lisaks võivad suulises kõnes, kuigi harva esineda ka takerduspausid või isegi partiklid keset sõna (*arva* (.) *musavaldusest*). Sellisel juhul ei suuda arvuti sõna kahte poolt kokku ühendada.

Kirjandus

Cui, Kaili 1999. Sõnaalguline *h* eesti keeles. Bakalaureusetöö. Käsikiri. Tartu ülikool.

EKG I 1995 = Erelt, M.; Kasik, R.; Metslang, H.; Rajandi, H.; Ross, K.; Saari, H.; Tael, K.; Vare, S. Eesti keele grammatika I: Morfoloogia. Sõnamoodustus. Tallinn: ETA Eesti Keele Instituut.

Hakulinen, Auli 2000. Partikkelit ja konjuktiot. Peatükk ilmuvast soome keele deskriptiivsest grammatikast. Käsikiri.

Hennoste, Tiit. 2000–2001. Sissejuhatus suulisesse eesti keelde I–IX. Akadeemia 2000: 5, 1117–1150; 6, 1343–1374; 7, 1553–1582; 8, 1773–1806; 9, 2011–2038; 10, 2223–2254; 11, 2465–2486; 12, 2689–2710; 2001: 1, 179–206.

Hennoste, Tiit ilmumas. Suulise kõne uurimine ja sõnaliigi probleemid.

Hennoste, Tiit; Lindström, Liin; Rääbis, Andriela; Toomet, Piret; Vellerind, Riina 2000. Eesti suulise kõne korpus ja mõne allkeele võrdlemise katse. – Arvutuslingvistikalt inimesele. Tartu ülikooli üldkeeleteaduse õppetooli toimetised 1. Toim. T. Hennoste. Tartu. 245–284.

Hudson, R. A. 1996. Sociolinguistics. Cambridge University Press.

Kaalep, Heiki-Jaan 1998. Tekstikorpuse abil loodud eesti keele morfoloogiaanalüsaator. – Keel ja Kirjandus 1, 22–29.

Kaalep, Heiki-Jaan; Muischnek, Kadri; Müürisep, Kaili; Rääbis, Andriela; Habicht, Külli 2000. Kas tegelik tekst allub eesti keele morfoloogilistele kirjeldustele? – Keel ja Kirjandus 9, 923–933.

Kaalep, Heiki-Jaan; Vaino, Tarmo 2000. Teksti täielik morfoloogiline analüüs lingvisti töövahendite komplektis. – Arvutuslingvistikalt inimesele. Tartu ülikooli üldkeeleteaduse õppetooli toimetised 1. Toim. T. Hennoste. Tartu. 87–100.

Keevallik, Leelo 1996. Maintenance of structured variability. – Estonian in the Changing World. Ed. by H. Õim. University of Tartu, Department of General Linguistics. 123–132.

Keevallik, Leelo 2001. Tracing grammaticalization of oota '*wait*' in Estonian conversation. – Papers in Estonina Cognitive Linguistics. Publications of the Department of General Linguistics 2. Ed by Ilona Tragel. Tartu. 119–144.

Lindström, Liina 2001. Grammaticalization of *või/vä* questions in Estonian. – Papers in Estonian Cognitive Linguistics. Publications of the Department of General Linguistics 2. Ed by Ilona Tragel. Tartu. 90–118.

Longman 1999. = Biber, Douglas; Johansson, Stig; Leech, Geoffrey; Conrad, Susan; Finegan, Edward. Longman Grammar of Spoken and Written English. Longman.

Pajusalu, Renate 1997. Is there an article in (Spoken) Estonian? – Estonian Typological Studies II. Publications of the Department of Estonian of the University of Tartu 8. Ed. by M. Erelt. Tartu. 146–177.

Pajusalu, Renate 2000. Indefinite determiners *mingi* and *üks* in Estonian. – Estonian Typological Studies IV Publications of the Department of Estonian of the University of Tartu 14. Ed. by M. Erelt. Tartu. 87–117.

Sorjonen, M.-L. 1999. Dialogipartikkelien tehtävistä. – Virittäjä 2, 170–194.

Stenström, A.-B. 1990. Lexical items peculiar to spoken discourse. – The London-Lund Corpus of Spoken English. Description and Research. Lund Studies in English 82. Ed. by J. Svartvik. Lund: Lund UP. 137–176.

Strandson, Krista 2001. Kuidas vestluskaaslane parandusprotsessi algatab. – Keel ja Kirjandus 6, 394–409.

Toomet, Piret 2000. Mõnest kaudsuse väljendamise võimalusest tänapäeva eesti keeles. – Keel ja Kirjandus 4, 251–259.

Püsiühendite leidmine teksti abil[1]

Heiki-Jaan Kaalep, Kadri Muischnek
Tartu ülikool

1. Sissejuhatus

Tekstis esinevate lausete edukaks automaatanalüüsiks ei piisa ainult morfoloogia- ja süntaksireeglite tundmisest ja kasutamisest. Hea tulemuse saamiseks peab tingimata arvestama ka selles keeles esinevate püsiühenditega. Lisaks on nende tundmine vajalik ka leksikograafias, keeleõppes jm. Kjellmer (1991) on näiteks väitnud, et meie mentaalne leksikon ei koosne mitte ainult üksikutest sõnadest vaid ka pikematest üksustest.

Momendil ei teata eesti keele püsiühenditest kuigi palju. Õigemini – teatakse küll, kuid see teave on suunatud peamiselt inimesele, keda huvitavad keelekasutuse nüansid ja väljendusrikkus. Nii sisaldavad sellealased publikatsioonid (nt EKSS; Hasselblatt 1990; Õim 1993, 1998) küll suurel hulgal fraseologisme ja idioome, kuid pole keele automaattöötluses kuigi lihtsalt kasutatavad. Esiteks seetõttu, et nende viimine automaatselt töödeldavale kujule pole lihtne, ja teiseks seepärast, et paljusid nendes sõnaraamatutes toodud väljendeid kasutatakse tegelikes tekstides küllaltki harva. Näiteks ei leidu tänapäeva eesti kirjakeele korpuse 1990. aastate allkorpuses (umbes 1 miljon sõna) kordagi selliseid väljendeid nagu *peenike peos* või *astla vastu üles lööma.* Nii et meie ees on kaks küsimust:

1) kas tekstides leidub veel (ja kui palju) püsiühendeid, mida sõnaraamatutes (ka fraseoloogiale orienteeritutes) ei esitata?
2) kui laialt kasutatavad on erinevad püsiühendid?

Alustasime vastuse otsimisest esimesele küsimusele, sest alles siis, kui meil on olemas loend võimalikest püsiühenditest, saame vastata küsimusele, kui sageli neist igaühte kasutatakse. Seejuures otsustasime töövahendina kasutada statistilist meetodit kasutavat arvutiprogrammi, mis tekstikorpusele rakendatuna lihtsustab lingvisti tööd.

Kuna kirjeldatavas eksperimendis kasutatakse püsiühendite leidmiseks tekstist statistilisi meetodeid, on püsiühend siin võrdsustatud kollokatsiooniga. Kollokatsioon on sõnaühend, mis on definee-

[1] Tööd on osaliselt finantseerinud ETF (grant nr 4352).

ritud selle järgi, et teda moodustavad sõnad esinevad tekstides koos sagedamini, kui võiks eeldada nende eraldi esinemise sagedustest. Kollokatsioonid võivad olla väga erinevad nii neid moodustavate sõnade arvu poolest kui ka nende sõnade süntaktiliste funktsioonide ja omavaheliste seoste poolest. Nendeks võivad olla nii idioomid (nt *hambasse puhuma*), mida sõnaraamatud esitavad põhjalikult, kuid mida tekstides harva esineb; ühend- ja väljendverbid, mida samuti sõnaraamatutes tüüpiliselt esitatakse (*üle saama, õppust võtma*); mitmesugused nimisõnafraasid (nt *rohelised mehikesed*). Lisaks nimetatutele on kollokatsioonid näiteks veel kindla verbi ja nimisõna seosed (nt puid lõhutakse, mitte ei tehta katki), mis võõrkeeleõppijatele suurt peavalu valmistavad. Kollokatsioone moodustavad sõnad ei pruugi paikneda lauses vahetult üksteise järel. See kõik teeb nende automaatse tuvastamise tekstis keeruliseks.

Sõnaühendite või kollokatsioonide leidmiseks tekstis kasutatakse sõnade omavahelise seotuse mõõte. Enamus sõnade omavahelise seotuse leidmise meetodeid põhineb ideel lükata ümber hüpotees, et sõnad A ja B on üksteisest sõltumatud. Statistilisi meetodeid kasutataksegi, et mõõta kõrvalekallet sellest hüpoteesist. Mida suurem on kõrvalekalle, seda tugevam on seos sõnade A ja B vahel. Seega mõõdab sõnade omavahelist seotust valem, mis võrdleb sõnade A ja B tegelikke sagedusi mingis etteantud suurusega naabruses ja nende eeldatavaid sagedusi üksteise naabruses, mida saab tuletada nende sõnade sagedusest kogu tekstis. Väga hea lühikese kokkuvõtte enamkasutatavatest leksikaalsete seoste (ja seega ka kollokatsioonide) leidmise meetoditest esitab Evert 2001. Evert ja Krenn (Evert, Krenn 2001) on võrrelnud erinevaid meetodeid ja leidnud, et ükski nende poolt vaadeldud meetod polnud oluliselt parem kui teised. Nad väidavad ka, et madala sagedusega sõnapaaride leidmiseks polegi sobivat meetodit. Nagu siin artiklis edaspidi näidatakse, on tekstis kord-paar esinevaid sõnaühendeid võimalik tuvastada küll, aga selle hinnaks on suur käsitsitöö maht kõigi võimalike kandidaatide käsitsi kontrollimisel.

Mitmesõnaliste üksuste tuvastamine elektroonilisest tekstikorpusest ei ole nii lihtne kui esmapilgul paistab: programm võib "leida" tekstist väljendeid, mis koosnevad sõnadest, mis küll võivad selles tekstis sageli koos esineda, aga mingit mõttelist tervikut ei moodusta. Samas võib programm jätta tuvastamata selliseid sõnaühendeid, mida teksti käsitsi läbi vaatav lingvist peaks omavahel kokku kuuluvateks.

Selleks, et statistikal põhinev programm võimalikult häid tulemusi annaks, peab ta arvestama ka analüüsitavate tekstide ja otsitavate väljendite lingvistiliste omadustega.

Selles artiklis kirjeldataksegi katset kombineerida lingvistilisi ja statistilisi meetodeid ühend- ja väljendverbide tuvastamiseks eestikeelses tekstikorpuses. Lühiduse mõttes ja ka analoogia põhjal ingliskeelse väljendiga *phrasal verb* on neid ühend- ja väljendverbe koos nimetatud siin fraasiverbideks. Miks ei ole kasutatud väljendit *perifrastiline verb*? Mõiste *perifrastiline verb* hõlmab ka modaalverbi ja infiniidi ühendeid (EKG II 1993: 19), mille tekstis tuvastamine ei olnud kirjeldatava töö eesmärgiks.

Programmi töö kontrollimiseks võrreldi korpusest leitud fraasiverbe püsiühendite andmebaasiga (http://www.cl.ut.ee/ee/ressursid/pysiyhendid.html), see võimaldas anda hinnanguid nii programmi tööle kui ka püsiühendite andmebaasile.

2. Tarkvara

Eesti fraasiverbide otsimiseks kohandasime G. Diasi loodud tarkvarapaketti SENTA (*Software for Extracting N-ary Textual Associations* – *n*-kohaliste sõnaühendite ekstraheerimise tarkvara) (Dias *et al* 2000). SENTA kasutab keerulist matemaatilist valemit ja lokaalse maksimumi leidmise algoritmi, et hinnata tekstis esinevate sõnade kokkukuuluvust. Allpool kirjeldame neid lühidalt; põhjalikuma ülevaate annab Dias *et al* 2000.

2.1. Ühise oodatavuse (ÜO, *Mutual Expectation*) mõõt

Mitmesõnalised üksused on definitsiooni kohaselt sõnajadad, mis esinevad üksteise läheduses liiga sageli, et see saaks olla juhuslik. Sellest eeldusest lähtudes defineeritaksegi sõnajadasse kuuluvate sõnade kokkukuuluvuse määra kirjeldav matemaatiline mudel. Seda mudelit kasutatakse, et arvutada ühist oodatavust, mis omakorda tugineb normaliseeritud oodatavusel.

2.1.1. Normaliseeritud oodatavus NO (*Normalised Expectation*)

N sõna vahelist normaliseeritud oodatavust defineeritakse kui keskmist ootust, et teatud positsioonis esineb mingi kindel sõna, kui (*n-1*) positsioonis juba esinevad sõnad on teada. Nt. kolmiku "*vahi alla võtma*" [*vahi +1 alla +2 võtma*] keskmine ootus peab arvesse võtma,

et *võtma* tuleb pärast *vahi alla*, aga ka seda, et *alla* esineb *vahi* ja *võtma* vahel ning et *vahi* esineb enne kui *alla võtma*. Olukorda kirjeldab tabel 1, kus iga rida tähistab üht võimalikku ootust.

Tabel 1. Näide ootustest, mida tuleb arvestada normaliseeritud oodatavuse hindamisel

Sõna oodatavus	Teades lünklikku kolmikut
vahi	[_____ *+1 alla +2 võtma*]
alla	[*vahi +1* _____ *+2 võtma*]
võtma	[*vahi +1 alla +2* _____]

Normaliseeritud oodatavuse põhiideeks on hinnata ühe sõna jadast väljajätmise maksumust (kokkukuuluvuse mõttes). Mida tihedamalt on jada sõnad omavahel seotud, st mida vähem lubavad nad endi hulgast mõne eemaldamist, seda suurem on normaliseeritud oodatavus. Normaliseeritud oodatavus defineeritakse kui *n* liikmega sõnajada esinemise tõenäosus, mis on jagatud kõigi selliste (*n-1*) liikmeliste sõnajadade tõenäosuste keskmisega, mis erinevad *n*-liikmelisest sõnajadast 1 sõna eemaldamise poolest.

$$NO = \frac{p(n - pikkusega\_jada)}{\frac{1}{n}\sum p(n-1-pikkusega\_jada)}$$

Seega, mida rohkem on tekstis selliseid (*n-1*) liikmelisi jadasid, mis esinevad kuskil mujal kui meid huvitava *n*-liikmelise jada koosseisus, seda suurem on nende tõenäosuste aritmeetiline keskmine ja seega seda väiksem on NO.

2.1.2. Ühine oodatavus (ÜO, *Mutual Expectation*)

Daille (1995) on näidanud, et üheks tõhusaks kriteeriumiks mitmesõnaliste üksuste leidmisel on lihtne sagedus. Sellest tulenevalt väidetakse, et kahest ühesuuruse NO-ga sõnajadast on see, kumb on sagedasem, ka tõenäolisem mitmesõnalise üksuse kandidaat:

$$ÜO = p(n - pikkusega\_jada) \times NO(n - pikkusega\_jada)$$

2.2. GenLocalMaxs algoritm

Kui oleme välja arvutanud ühe sõnajada ÜO ja temas sisalduva, ühe sõna võrra lühema sõnajada ÜO, siis kasutame GenLocalMaxs algoritmi otsustamaks, kumb neist on 'see õige'. See algoritm eeldab, et üks sõnajada on mitmesõnaline üksus või, antud juhul, fraasiverb, kui kokkukuuluvus seda moodustavate sõnade vahel pole väiksem tema alaosade kokkukuuluvusest ja kui see kokkukuuluvus ise on suurem pikema sõnajada osade kokkukuuluvusest, so kui see sõnajada ise ei ole mõne suurema püsiväljendi osa. Teiste sõnadega, üks sõnajada, ütleme *W*, on mitmesõnaline üksus või meie juhul fraasiverb, kui tema ühise oodatavuse väärtus, *ÜO(W)* on lokaalne maksimum. Olgu *n*-sõnalises jadas *W* sisalduvate *(n-1)*-sõnaliste jadade hulk Ω_{n-1} ja kogu *(n+1)*-sõnaliste jadade hulk, milles sisaldub *W*, Ω_{n+1}. Siis

$$\forall x \in \Omega_{n-1}\ ,\ \forall y \in \Omega_{n+1}$$

kui n=2 siis

kui *ÜO(W)* > *ÜO(y)*, siis *W* on mitmesõnaline üksus

kui N>2, siis

kui *ÜO(x)* ≤ ÜO*(W)* ja ÜO*(W)* > ÜO*(y)*, siis on *W* mitmesõnaline üksus

Seega, juhul kui:

- sõnajada pikkus on 2, siis tema hinne peab olema suurem kui teda sisaldavatel pikematel sõnajadadel;
- sõnajada pikkus on üle 2, siis saab tema hinnet lisaks võrrelda ka nende sõnajadade hinnetega, mida tema sisaldab; seejuures piisab, kui alamjadade hinded ei ületa tema oma.

3. Tekstikorpus

Kirjeldatavas katses kasutati üht osa tänapäeva eesti kirjakeele korpusest (http://www.cl.ut.ee/ee/corpusb/), nimelt 500 000-sõnalist väljavõtet 1990. aastate ilukirjanduse allkorpusest. 1990. aastate ilukirjanduskorpuses on kokku 611 000 sõna ja ta sisaldab ilukirjandustekste aastatest 1991–1998. Kuna eesti algupärase proosa kogutoodang on nii väike, on korpusse valitud üks katke igast eesti keeles ilmunud proosaraamatust, lisaks ilukirjandust ajakirjast *Looming*. Korpusse viidud tekstikatkete maht on ca 2000 sõna (või ka vähem, kui nt novell juhtus lühem olema), valikud arvestavad lõigupiire ja

pole seepärast täpselt 2000 sõna pikkused. Korpuse koostamispõhimõtete kohta vt lähemalt Hennoste, Muischnek 2000.

Tahtsime oma eksperimendis kasutada võimalikult tänapäevast teksti, valida oli ilukirjanduse ja ajakirjanduse korpuste vahel. Ilukirjandust eelistasime ajakirjandusele seepärast, et vastandina ajalehekeelele või suulisele keelekasutusele on see alati olnud traditsioonilise leksikograafia ja leksikoloogia põhiline "inspiratsiooniallikas". Eeldasime, et kasutades sama tüüpi teksti kui sõnaraamatute koostajad, õnnestub paremini võrrelda eelnevalt sõnaraamatute põhjal koostatud andmebaasi ja SENTA tulemusi.

4. Püsiühendite andmebaas

Püsiühendite andmebaas (http://www.cl.ut.ee/ee/ressursid/pysiyhendid.html) koosneb momendil väljendite andmebaasist ja fraasiverbide andmebaasist. Fraasiverbide andmebaasis, millega kirjeldatav eksperiment tehti, oli 2001. a. lõpu seisuga u 12 200 kirjet. Andmebaas on koostatud järgmiste inimkasutajale mõeldud sõnaraamatute baasil:

- "Fraseoloogiasõnaraamat" (Õim 1993);
- "Eesti kirjakeele seletussõnaraamat" (EKSS);
- Filosofti tesaurus (http://www.filosoft.ee);
- Partikkelverbide loend "Das Estnische Partikelverb als Lehnübersetzung aus dem Deutschen" (Hasselblatt 1990);
- "Eesti keele mõistelise sõnaraamatu" indeks (Saareste 1979);
- "Sünonüümisõnastik" (Õim 1991).

Fraasiverbid on meie andmebaasis jagatud ühend- ja väljendverbideks.

Ühendverb koosneb traditsioonilise grammatika järgi verbist ja selle tähendust muutvast abimäärsõnast. Käesoleva andmebaasi koostamisel on ühendverbideks nimetatud kõiki selliseid määrsõna ja verbi ühendeid, mida on sõnastikes esitatud omaette (ala)märksõnadena. Ka SENTA leitud määrsõna ja verbi ühenditest on andmebaasi võetud mitte ainult need sõnaühendid, kus (abi)määrsõna muudab verbi tähendust, vaid igasugused ühendid, mis tunduvad moodustavat omaette mõiste või mida nt inglise keelde ei saa tõlkida sõna-sõnalt (nt *istuli kukkuma*, *juurde tellima*). Seega võib öelda, et ühendverbide hulka on siinses andmebaasis tinglikult loetud igasugused määrsõna ja verbi ühendid v.a verbi *olema* ja määrsõna ühendid. Verbi *olema* ühendid on fraasiverbide andmebaasist välja jäetud.

Ühendverbide andmebaasis leidub ka selliseid muutumatu sõna ja verbi ühendeid, kus muutumatu sõna ei ole tõenäoliselt mitte määrsõnaks, vaid kaassõnaks; selline ühend nõuab kindlas vormis käändsõna. Nt sellistes sõnaühendites, nagu *üle naerma* või *küüsi jätma* funktsioneerib muutumatu sõna alati kaassõnana, mitte määrsõnana. Muidugi on selline tõlgendus ainult oletuslik, see, kuidas sellised sõnaühendid esinevad tegelikus keelekasutuses, vajab korpusel testimist.

Eesti keele grammatika (EKG II 1993) nimetab väljendverbideks neid perifrastilisi verbe, mille sisuliseks tuumaks on noomen. Kõik väljendverbid on ainukordsed ühendid, mis moodustavad idiomaatilise tähendusterviku (EKG II 1993: 20). Andmebaasi koostamisel on ka väljendverbi mõistet käsitletud võimalikult laialt: väljendverbidena on andmebaasi sisse võetud kõik nimisõna(de) ja verbi ühendid, mis on sõnastikes esitatud omaette (ala)märksõnadena. Samuti nagu ühendverbide puhul, on ka SENTA poolt tekstist tuvastatud nimisõna(de) ja verbi ühendite puhul andmebaasi võetud kõik ühendid, mis tundusid moodustavat omaette mõiste (nt *aega kulutama, imet tegema, bussi ootama*). Väljendverbide alla on andmebaasis tinglikult viidud ka kahe verbi ühendid (finiitverb+infiniit), nt *ajama panema, nahutada saama*. Kuid põhjustel, mida edaspidi täpsemalt selgitatakse, ei ole võimalik finiitverbi ja infiniidi ühendeid korpusest SENTA abil leida.

Enne katset võis oletada, et selline statistikal põhinev programm nagu SENTA ei anna eestikeelsele tekstile rakendatuna kuigi häid tulemusi, sest kokkukuuluvad sõnad võivad tekstis asuda üksteisest kaugel ja alati mitte samas järjekorras (*tahtis aru saada, sai aru*). Lisaks sellele ei esine kokkukuuluvad sõnad alati samal kujul (vrd nt *sai aru, ei saanud aru; rohelised mehikesed, rohelistele mehikestele)*. Võis eeldada, et programm leiab tekstist palju mõttetuid väljendeid, s.o sõnu, mis esinevad küll sageli koos samas osalauses, kuid ei moodusta ühte mõttelist tervikut. Statistikal põhinevad kollokatsioonide leidmise programmid on koostatud leidmaks korduvaid ja tõenäolisi seoseid sõnavormide vahel, nad ei arvesta erinevate keelte spetsiifikat.

Püsiühendite leidmisel eestikeelses tekstis tuleb kindlasti arvestada lemmatiseerimise probleemiga. Kas tekstis olevad sõnavormid on otstarbekam viia lemma kujule või jätta nad sellisteks, nagu nad tekstis esinevad? See sõltub sellest, milliseid püsiühendeid soovi-

takse leida. Näiteks nimisõnafraase tuvastades tuleb arvestada nii ühilduvate kui genitiivsete ja ka järeltäiendite ga. Kirjeldatava ülesande – ühend- ja väljendverbide leidmise – puhul tuli teksti ettevalmistamisel viia tekstis esinevad verbivormid lemma kujule ja ülejäänud tekstisõned jätta tekstis kasutatud kujule. Siit leiame vastuse küsimusele, miks ei ole SENTAga tuvastatud fraasiverbide seas finiitse verbivormi ja *da*-infinitiivi ühendeid (nt *tunda saama*), aga on *ma*-infinitiivi ja finiitse verbivormi ühendeid (nt *magama heitma*).

4. Eksperiment

Eksperimendis kasutati eelpoolkirjeldatud 500 000-sõnalist tekstikorpust ja fraasiverbide andmebaasi.

Ühend- ja väljendverbide korpusest leidmiseks töödeldi teksti järgnevalt:

1. Tekstid analüüsiti morfoloogiliselt ja ühestati. Iga sõnavormi juures oli seejärel tema lemma ja info tema sõnaliigi, arvu, käände või pöörde jms kohta.

2. Verbidel säilitati lemma ja eemaldati tekstis esinenud sõnavorm ja muu morfoloogiline info, teistel sõnaliikidel hoiti alles sõnavorm ja eemaldati morfoloogiaanalüsaatori poolt lisatud info.

3. Leiti kõik võimalikud kollokatsioonid.

4. Eemaldati käesoleva ülesande seisukohalt mitteolulised kollokatsioonid, s.o kõik kollokatsioonid, mis ei sisalda verbi; asesõna sisaldavad kollokatsioonid v.a mõned erandid (muidu oleks kõige sagedasem kollokatsioon *tema olema*), samuti eemaldati kirjavahemärke sisaldavad kollokatsioonid (SENTA jaoks on kirjavahemärk samasugune sümbol nagu täht, number vms), mõningaid adverbe sisaldavad kollokatsioonid. Nende eemaldatavate adverbide nimekiri kujunes töö käigus ja sinna kuuluvad näiteks sõnad *alati, muidugi, ikka* – s.o sellised adverbid, mis ei saa kuuluda ühendverbi koosseisu, kuid on tekstis küllalt sagedased.

5. Arvutati ÜO ja GenLocalMaxs; nende põhjal tehti leitud kollokatsioonide hulgast lõplik valik, mis läks programmi väljundisse.

Kirjeldatavas eksperimendis töödeldi tekstikorpust SENTAga 4 korda, iga kord erineva distantsiga (0 kuni 3), so kollokatsiooni moodustavate sõnade vahel võis olla 0 kuni 3 sõna. Selliselt saadud ühend- ja väljendverbi kandidaatide nimekirja võrreldi püsiühendite andmebaasiga, need väljendid, mida andmebaasis ei olnud, vaadati käsitsi üle ja otsustati igaühe puhul eraldi, kas tegemist on fraasi-

verbiga või mitte. See osa tööst – SENTA väljundist "mõistlike" verbiühendite väljavalimine – nõudis kõige rohkem inimtööd, sest programmi töö täpsus on vaid 19%.

5. Tulemused

SENTA leidis 500 000-sõnalisest tekstikorpusest 13 100 fraasiverbikandidaati. Neist 2500, s.o. 19% olid n.ö mõistlikud kandidaadid, sellised, mida võiks meie kriteeriumite järgi nimetada ühend- või väljendverbideks. Nendest 2500-st 1630 olid püsiühendite andmebaasis juba olemas ja 870 oli selliseid, mis püsiühendite andmebaasis puudusid. Tabelis 2 on esitatud mõned neist fraasiverbidest, mis olid püsiühendite andmebaasis ja/või mida SENTA leidis tekstikorpusest. Nagu näeme, leidub sõnastikes ühelt poolt väljendeid, mida korpuses ei esine, teiselt poolt on aga korpuses küllalt igapäevaseid väljendeid, mida sõnastikud ei esita.

Tabel 2. Ühend- ja väljendverbe fraasiverbide andmebaasis ja SENTA väljundis

Püsiühend	Andmebaas	SENTA väljundis
abiellu astuma	+	–
abiellu heitma	+	–
abielu rikkuma	+	–
abielu sõlmima	+	–
abielu lahutama	–	+
andeks andma	+	+
andeks paluma	+	–
andeks saama	–	+
allkirja andma	–	+
hulluks minema	+	+
hulluks ajama	–	+
külla minema	+	–
külla tulema	+	+
külla kutsuma	–	+

SENTA leitud 2500-st ühend- ja väljendverbist olid andmebaasis olemas ainult 2/3. Pealegi on 500 000-sõnaline tekstikorpus selliste ülesannete lahendamiseks väikesevõitu, suuremate tekstihulkade

kasutamisel võib loota veel olulisemat lisa olemasolevale andmebaasile.

6. SENTA töö kontroll

Töötades SENTAga nagu musta kastiga, kuhu pannakse tekstikorpus sisse ja saadakse väljundiks potentsiaalsete püsiühendite loend, ei saa me teada, kuivõrd on tulemused usaldatavad, s.o. kui palju tekstis tegelikult esinevaid püsiühendeid SENTA leiab ja kui palju SENTA leitud kollokatsioonidest tegelikult tekstis olemas on. Selle väljaselgitamiseks tehti järgmine katse. Püsiühendite andmebaasist valiti juhuslikult välja 500 ühend- ja väljendverbi. Nende esinemist korpuses kontrolliti käsitsi. Selgus, et nendest 500-st püsiühendist esines tekstikorpuses 131. Põhimõtteliselt peaks SENTA olema võimeline leidma korpusest neid kollokatsioone, mis esinevad seal vähemalt 2 korda. Selliseid ühend- ja väljendverbe oli 500-st 71.

Tekstikorpust SENTAga töödeldes tehti 4 katset, iga kord lubati erinevat distantsi (0 kuni 3 sõna) kollokatsiooni moodustavate sõnade vahel, lisaks veel kombineeritud distants so vahemik 0 kuni 3 sõna. Kontrolliti, mitu fraasi 71 hulgast SENTA leidis. Nende katsete tulemused on esitatud tabelis 3.

Tabel 3. SENTA leitud ühend- ja väljendverbide hulga sõltuvus lubatud distantsist püsiühendit moodustavate sõnade vahel

Distants	0	1	2	3	kombineeritud
Püsiühendeid	45	46	50	52	57

Tabelist 3 näeme, et mida pikem on lubatud distants, seda rohkem õigeid väljendeid suudab SENTA korpusest leida. Kuid väärib märkimist, et distantsi pikendamisel ei leia SENTA enam mõningaid fraase, mida ta leidis lühema distantsi puhul. Kui suurendada distantsi kahelt sõnalt kolmele, toimub järsk muutus: SENTA ei leia enam mõningaid püsiühendeid, mis on korpuses sagedased. Nende 19 püsiühendi hulgas, mida SENTA distantsi suurenedes enam üles ei leidnud, esinesid 12 korpuses kaks korda, kuid viis püsiühendit olid üsna sagedased (vt tabel 4). "Koosesinemisi" antud tabelis tähistab juhtumeid, kus mõlemad sõnad esinevad samas lauses arvestamata seda, kas nad moodustavad ühend- või väljendverbi või mitte.

Tabel 4. Korpuses sageli esinevad püsiühendid, mida SENTA 3-sõnalise distantsi puhul enam ei tuvastanud

Sõnaühend	Koosesinemisi	Püsiühendeid
ette näitama	10	9
hakkama saama	95	58
suitsu tegema	11	9
ära kasutama	21	19
ära maksma	12	9

Distantsidega 0, 1 ja 2 ei leidnud SENTA tabelis 4 toodud fraasidest ainult väljendit *ära maksma* (kuigi distantsi 3 puhul leidis ta rohkem väikese sagedusega püsiühendeid kui distantside 0, 1 ja 2 puhul).

SENTA andis välja ka selliseid püsiühendeid, mida sõnaraamatute põhjal koostatud andmebaasis ei olnud, aga mis sisaldasid endas andmebaasis leiduvaid fraase. Näiteks leidus andmebaasis väljend *ära maksma*, mida SENTA tekstikorpusest ei leidnud. Küll aga leidis ta väljendid *arve ära maksma* ja *võlga ära maksma*. See näib viitavat asjaolule, et ühendverbi *ära maksma* kasutataksegi põhiliselt nendes kontekstides. Kas ongi siis tegemist ühendverbiga *ära maksma* või hoopis kahe väljendverbiga?

Kontrollides käsitsi seda, kuidas suudab SENTA leida andmebaasist juhuslikult valitud viitsadat väljendit, selgus, et SENTA leidis tekstikorpusest ka mõned fraasiverbid, mida seal tegelikult ei olnud. Kuidas see võimalik on? Nimelt võivad ühend- või väljendverbi osad esineda samas osalauses ka ilma mõistelist tervikut moodustamata, s.o ilma kokku kuulumata. Nii leidis SENTA näiteks ühendverbi *tagasi tegema* lausest *Tagasi jõudes teeme sotid selgeks*, mis loeti veaks SENTA tulemuse hindamisel. Nagu arvata võib, kasvab selliste vigade hulk distantsi pikenedes.

Milliseid järeldusi saab teha nende 500 väljendi käsitsi kontrollimisest? Oletagem, et need 131 väljendit 500-st, mis korpuses esinesid, moodustavad juhusliku valiku kõigist korpuses esinevatest fraasidest. SENTA peaks suutma leida need väljendid, mis esinevad korpuses vähemalt 2 korda, antud juhul siis 71 131-st. Kasutades kombineeritud distantsi, võime me eeldada, et SENTA leiab korpuses olevatest väljenditest 57/71 = 80% neid, mis esinevad seal vähemalt kaks korda ja peaaegu 99% neist, mis esinevad kolm ja rohkem korda. Lisaks veel 8/60=12% neist, mis esinevad korpuses ühe korra.

Seega on saak (*recall*) väga hea, aga täpsus kehv – mäletatavasti ainult 19%.

7. Kokkuvõte

Tekstikorpusest fraasiverbide leidmine pole kerge ülesanne. Siin näidati, kuidas lahendada seda ülesannet kombineerides lingvistilisi ja statistilisi meetodeid, kusjuures väljund vajab käsitsi toimetamist.

Artiklis kirjeldati statistilisel tõenäosusel põhinevat sõnade omavahelise seotuse mõõtude leidmise programmi SENTA (*Software for Extracting N-ary Textual Associations*) ja selle eestikeelsele tekstile rakendamise katset. Paremate tulemuste saamiseks tuli teksti eelnevalt töödelda: verbid viia tekstis lemma kujule, muudel sõnaliikidel säilitada tekstis kasutatud sõnavorm. Eestikeelseks tekstiks valiti 500 000-sõnaline osa tänapäeva eesti keele korpuse 1990. aastate ilukirjanduse allkorpusest.

SENTA väljundit võrreldi mitmesuguste sõnaraamatute baasil koostatud eesti keele fraasiverbide andmebaasiga. SENTA töö kontrollimiseks teostati eksperiment, kus fraasiverbide andmebaasist juhuslikult valitud 500 ühendi esinemist kasutatud tekstikorpuses kontrolliti käsitsi. Kuigi kasutatud meetodi täpsus oli madal (19%), kompenseeris selle suur saak (*recall*) – SENTA leidis 99% ühenditest, mis esinesid korpuses kolm ja rohkem korda, 80% neist, mis esinesid kaks korda ning 12% ühenditest, mis esinesid üks kord.

Kuigi võis arvata, et eesti keele vaba sõnajärje ja keeruka morfoloogia tõttu ei tööta selline statistikal põhinev programm nagu SENTA eesti keelele rakendatuna eriti hästi, võib tulemusi vaadates järeldada, et see arvamus oli ekslik. See tähendab, et SENTA on leksikograafile hea abivahend: vaadata käsitsi läbi 13 100 fraasiverbi kandidaati on hoopis lihtsam kui otsida fraasiverbe 500 000-sõnalisest korpusest.

Kirjandus

Daille B. 1995. Study and implementation of combined techniques for automatic extraction of terminology. – The Balancing Act: Combining Symbolic and Statistical Approaches to Language. Ed. by J. Klavans, P. Resnik. Cambridge, MA; London, England: MIT Press. 49–66.

Dias, G.; Guilloré, S.; Bassano, J. C.; Lopes, J. G. P. 2000. Extraction automatique d'unités lexicales complexes: Un enjeu fondamental pour la recherche documentaire. – Journal Traitement Automatique des Langues 41:2, 447–473.

EKG II 1993 = Erelt, Mati; Kasik, Reet; Metslang, Helle; Rajandi, Henno; Ross, Kristiina; Saari, Henn; Tael, Kaja; Vare, Silvi. Eesti keele grammatika II: Süntaks. Lisa: Kiri. Tallinn: ETA Keele ja Kirjanduse Instituut.

EKSS = Eesti kirjakeele seletussõnaraamat (A–Žüriivaba). Tallinn: ETA KKI. 1988–2000.

Evert, S. On lexical association measures. http://www.collocations.de/EK/am-html/index.html.

Evert, S.; Krenn, B. 2001. Methods for the qualitative evaluation of lexical association measures. – Proceedings of the 39th Annual Meeting and 10th Conference of the European Chapter of ACL. CNRS, Toulouse, France. 188–196.

Filosoft = Tesaurus. http://ee.www.ee/Tesa.

Hasselblatt, C. 1990. Das Estnische Partikelverb als lehnübersetzung aus dem Deutschen. Wiesbaden.

Hennoste, Tiit; Muischnek Kadri 2000. Eesti kirjakeele korpuse tekstide valiku ja märgendamise põhimõtted ning kahe allkeele võrdluse katse. – Arvutuslingvistikalt inimesele. Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 1. Toim. T. Hennoste. Tartu. 183–218.

Kjellmer, G. 1991. A mint of phrases. – English Corpus Linguistics. Ed. by K. Aijmer, B. Altenberg. Longman. 111–127.

Rätsep, H. 1978. Eesti keele lihtlausete tüübid. Tallinn: Valgus.

Saareste, A. 1979. Eesti keele mõistelise sõnaraamatu indeks. Finsk-ugriska institutionen. Uppsala.

Õim, A. 1991. Sünonüümisõnastik. Tallinn.

Õim, A. 1993. Fraseoloogiasõnaraamat. Tallinn: ETA KKI.

Õim, A. 1998. Väljendiraamat. Tallinn: Eesti Keele Sihtasutus.

Semyhe tulemusi: kas tasub *naise* pärast WordNet ümber teha?

Neeme Kahusk, Kaarel Kaljurand
Tartu ülikool

1. Sissejuhatus

Automaatne semantiline ühestamine muutub tänapäeval üha aktuaalsemaks. Ühe olulise osa sellest moodustab sõnatähenduste ühestamine (ingl. k. *word sense disambiguation*). Tartu ülikooli arvutilingvistika uurimisrühmas on professor Haldur Õimu juhtimisel astutud esimesi samme sõnatähenduste ühestamiseks eestikeelsetes tekstides. Kuna tänavu toimunud sõnatähenduste ühestajate võistlusel SENSEVAL-2 võimaldati osaleda ka teistel keeltel peale inglise, siis sai sinna ülesanne püstitatud ja seda ka lahendada püütud.

Sõnatähenduste ühestamiseks on mitmeid meetodeid. Kuna TÜ arvutilingvistika uurimisrühmas tegeldakse ka eesti keele leksikaalsemantilise andmebaasi loomise ja täiustamisega, siis valiti ühestaja tööprintsiip selline, mis võimaldaks kasutada eesti WordNetis tehtut.

Käesolevas artiklis käsitleme ühe juhtumi näitel seda, kuidas sõnatähenduste ühestamine ja hierarhiliselt organiseeritud tesaurus üksteist mõjutavad.

2. Automaatne sõnatähenduste ühestamine

Automaatse sõnatähenduste ühestamise eesmärk on käsitleda tekstis leiduvaid sõnu kui erinevate tähenduste hulki (mis polüseemsete sõnade puhul sisaldavad rohkem kui ühe tähenduse) ning fikseerida iga hulga puhul üks tähendus, on mis antud kontekstis õige. Sellisena koosneb ühestamine kahest alamülesandest: (1) tuleb kindlaks teha vaadeldava sõna kõik tähendused ja (2) valida neist välja üks, mis antud juhul on kõige õigem (Ide, Véronis 1998).

Esimese alamülesande lahendamisel kasutatakse elektroonilist sõnastikku, mis mingil kujul sõnade kõikvõimalikud tähendused ära toob. Sellise sõnastiku valik sõltub mõnevõrra ka konkreetsest ülesandest, mille tarvis sõnatähenduste ühestamist läbi viiakse. Vahel on vaja väga detailset tähenduste eristamist (nt masintõlke puhul),

mõnikord on liiga detailne eristamine isegi segav, kuna võib analüüsi käiku liigselt keeruliseks ajada.

Teise alamülesande lahendamisel lähtutakse eeldusest, et vaatluse all oleva sõna tähendus sõltub tema kontekstist. Ühestamismeetodi väljatöötamisel ongi ülesandeks leida mudel, mis piisavalt adekvaatselt kirjeldaks vastavust sõna võimalike kontekstide ja tema võimalike tähenduse vahel.

Klassikaliseks ühestamismeetodiks on elektroonilise sõnastiku kasutamine, mis lisaks tähenduste eristamisele viitab näitelausete abil ka sõna kasutusolukordadele. Nõnda on esitatud polüseemne sõna kindlas kontekstis, kusjuures meil on teada, mis tähenduses ta vastaval juhul esineb. Sõna tähenduse määramiseks analüüsitavas tekstis võrreldakse tema konteksti ja tema erinevatele tähendustele vastavaid näitelauseid. See tähendus, millele vastav näitelause on kõige sarnasem sõna kontekstiga analüüsitavas tekstis, valitaksegi välja. Sarnasuse aluseks võib kõige lihtsamal juhul olla nt kahe sõnahulga ühisosa suurus.

Samuti saab ühestamisel ära kasutada semantiliselt käsitsi ühestatud tekstikorpust. See on mingis mõttes sarnane sõnastiku kasutamisega. Näitelauseid on korpuse puhul rohkem, kuid selle eest ei kata korpus kõiki tähendusi, vaid ainult neid, mis sagedamini esinevad. Piisavalt suur korpus võimaldab anda ka tähenduste esinemissagedusele adekvaatse hinnangu, mis on automaatse ühestamise puhul samuti oluliseks infoallikaks, kuna mingi sõna tähendused jaotuvad sageduse mõttes reeglina väga ebaühtlaselt.

3. WordNet ja kontseptuaalne kaugus

Meie oleme valinud automaatset ühestamist abistavaks sõnastikuks WordNet-tüüpi tesauruse, mis lisaks sõnade tähenduste eristamisele kirjeldab ka tähenduste vahelisi semantilisi suhteid. WordNet on oma olemuselt puuhierarhia, milles sõnatähendused (mõisted) on sünonüümihulkadesse grupeeritult puu tippudeks, ning nendevahelised suhted, nt hüperonüümia/hüponüümia on puu kaarteks.

Iga erinev suhtetüüp määrab tegelikult erineva puustruktuuri, mis eksisteerib teistega paralleelselt. Meie oleme automaatse sõnatähenduste ühestamise raames keskendunud esialgu vaid eelnimetatud hüperonüümia/hüponüümia struktuurile.

Seega eristab WordNet sõnatähendusi (üks sõna võib kuuluda mitmesse erinevasse sünonüümihulka), aga sõnatähenduste kirjelda-

misel ei keskenduta mitte definitsioonidele ja näitelausetele, vaid tähenduste vahelistele suhetele.

Selleks, et kirjeldada sõnatähenduste omavahelist sarnasust/erinevust on kasutusele võetud *kontseptuaalse kauguse* mõiste, mida on loomulik defineerida kui kaarte arvu puus, mida mööda liikudes on võimalik jõuda ühest sõnatähendusest teiseni (Rada, Mili, Bicknell 1989). Kuna struktuuri näol on tegemist puuga, siis leidub alati täpselt üks selline tee.

Sellisena on kontseptuaalne kaugus eelkõige defineeritud kahe mõiste jaoks. Kui me tahame korraga seostada omavahel rohkem kui kahte mõistet, siis võib kontseptuaalse kauguse definitsiooni üldistada, näiteks summeerides kaugused kõikide tähenduste paaride jaoks.

Tähistagu s sõna tähendust ning $d(s_1, s_2)$ kahe tähenduse vahelist kaugust, siis üldistatud kontseptuaalne kaugus on defineeritud kui

$$\mathrm{dis}(s_1, s_2, \ldots, s_n) = \sum_{i=1}^{n} \sum_{j=i+1}^{n} d(s_i, s_j)$$

Selleks, et siduda analüüsitava sõna kontekst tema võimalike tähendustega võibki kasutada eeldefineeritud kontseptuaalse kauguse mõistet. Kuidas võiks töötada kontseptuaalset kaugust arvestav sõnatähenduste ühestaja, seda vaatleme järgneva lause näitel:

Jüri võttis pangast raha.

Oletame, et kasutatava sõnastiku (WordNet) põhjal on sõnal *Jüri* 1 tähendus, sõnal *võtma* 7 tähendust, sõnal *pank* 2 tähendust ja sõnal *raha* 1 tähendus, siis kokku on võimalikke tähenduste järjendeid 1 * 7 * 2 * 1 = 14. Automaatse ühestamise ülesandeks on neist välja valida üks (või üldisemalt, järjestada kogu tähenduste järjendite hulk, nt sidudes iga järjend tõenäosusega, mis näitab järjendi adekvaatsust). Sellise valiku aluseks võibki võtta kontseptuaalse kauguse, mida on eeltoodud valemi abil lihtne leida: iga järjendi jaoks arvutatakse sellele vastav kontseptuaalne kaugus ning valituks osutub järjend, mille puhul see kaugus on vähim.

Juhul kui 'pank' kui finantsasutus asub WordNeti puustruktuuris 'rahale' lähemal kui 'pank' rannikutüübina, mis on intuitiivselt loomulik, siis figureerib väljavalitud järjendis just esimene.

4. Sõnatähenduste ühestamise süsteem *semyhe* ja esimesed tulemused

Süsteem *semyhe* tuginebki oma analüüsis eelkirjeldatud kontseptuaalse kauguse mõistele, kasutades oma analüüsis eesti keele jaoks loodud ja pidevalt täiendatavat WordNet-tüüpi tesaurust (EstWN).[1] Lisaks on kasutatud ka ideid, mis on pärit Agirre ja Rigau (1996) tööst, kus kasutati kontseptuaalse tiheduse (*conceptual density*) meetodit inglise keele nimisõnade WordNeti-põhiseks ühestamiseks. See meetod täiendab kontseptuaalse kauguse meetodit, arvestades lisaks võrreldavate mõistete hüponüümide hulka (st alampuude sügavust ja laiust).

Käesoleval aastal osaleti ka sõnatähenduste ühestajate võistlusel SENSEVAL-2, kus püstitati ülesanne sõnatähenduste ühestamiseks eesti keeles (Kahusk, Orav, Õim, ilmumas) ja püüti seda *semyhe* abil lahendada (Vider, Kaljurand, ilmumas). *Semyhe* esinemine oli esimese korra kohta tubli – täpsus (*precision*) 0.66 oli peaaegu võrdne John Hopkinsi Ülikooli süsteemi JHU tulemusega (0.67), mis samuti eesti ülesannet lahendas.

Süsteemi parandamiseks on põhimõtteliselt kaks võimalust.

- Täiendada algoritmi, nt kaasates analüüsi rohkem informatsiooni, mis WordNetis olemas on, kuid mida seni on eiratud (nt teist tüüpi suhted sõnade vahel). Samuti saab suhtuda sisendteksti teisiti, käsitledes sõnade konteksti mitte pelgalt sõnahulgana, vaid struktureerituna, kus mõned konteksti sõnad mõjutavad ühestatavat sõna rohkem kui teised.
- Parandada ühestaja aluseks olevat WordNetti, eeldades, et ühestaja vead on tingitud just WordNeti vigadest

Järgmiseks keskendumegi viimati nimetatud võimalusele ja vaatame ühte juhtumit, mille lahendamisega *semyhe* toime ei tulnud.

5. *Naisega* ei saa hakkama

Meie käsitsi ühestatud tähendustega korpuses, mis koosneb põhiliselt ilukirjandusest, esineb sõna 'naine' 59 korral. Tabelist 1 on näha, et selle sõna tähendusi on *semyhe* leidnud väga halvasti.

[1] Eesti WordNetist ja selle tegemisest võib lugeda lähemalt Vider, Orav 1998; Vider, Kahusk, Orav, Õim, Paldre 2000.

Tabel 1. Sõna 'naine' tähenduse ühestamine käsitsi ja süsteemiga *semyhe*

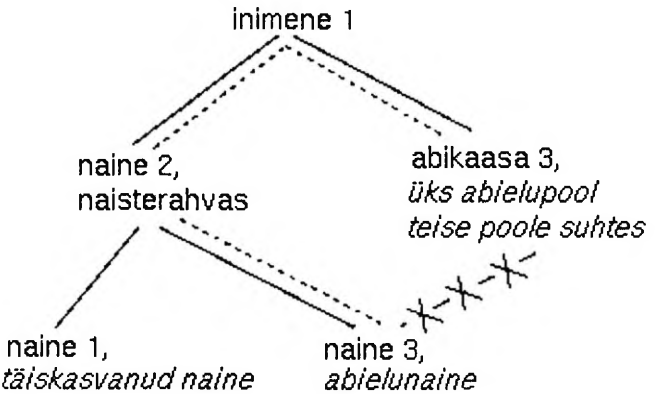
Viimases tulbas on komadega eraldatult kõik tähendused, kui süsteem pakkus neid rohkem kui ühe

Esinemisi	Tekst	Käsitsi ühestatud tähendus	*semyhe* leitud tähendus(ed)
1	tkt0020	1	2
4	tkt0020	3	1,2,3
15	tkt0020	3	2
1	tkt0031	3	2
3	tkt0031	1	2
1	tkt0038	1	2
1	tkt0038	2	2
1	tkt0086	1	2
3	tkt0086	3	2
9	tkt0087	3	2
2	tkt0088	1	2
1	tkt0088	2	1,2,3
1	tkt0090	1	2
1	tkt0112	2	2
2	tkt0119	1	2
1	tkt0119	3	1,2,3
12	tkt0119	3	

Eesti WordNetis on sõnal 'naine' kolm tähendust:
(1) täiskasvanud naine (vastandina mehele);
(2) naissoost inimene (olenemata vanusest);
(3) abielunaine.

Vaadates neid tähendusi võib tekkida küsimus, kas on oluline neid üldse tekstis eristada, kuna nad tunduvad olevat küllalt sarnased. Ette rutates võib märkida, et see sarnasus avaldub ka selles, kuidas nad praegu eesti WordNetis paiknevad. Siiski – kui meil on ülesanne tõlkida eestikeelne tekst näiteks inglise keelde, võib tekkida olukord, kus on vaja valida, millist sõna kasutada. Kui (2) tähenduse puhul võib vabalt kasutada asesõna *she*, siis (1) ja (3) puhul võib tekkida valik, kas *she/woman* või *she/wife*. Seega on tõlkeülesande puhul tegu valikuga vähemalt (1) ja (3) tähenduse vahel.

Et mõista, kuidas ja miks on 'naise' tähendusi *semyhe* praegu ühestanud nii, nagu on, vaatame, mis toimub sõna ühestamisel, ja



Joonis 1. Sõna 'naine' erinevate tähenduste paiknemine eesti WordNetis

Punktiiriga on tähistatud teekond mõistete 'abikaasa' ja 'abielunaine' (naine 3. tähenduses) vahel. Ristikestega pikitud joon tähistab kohta, kuhu saaks lisada teise hüperonüümiasuhte, mis rikuks puukujulise hierarhia.

millest tulemus oleneb. Lihtsuse mõttes võtame eelduseks, et sõna, mille suhtes 'naist' ühestatakse on 'abikaasa' ja et sellel on eesti WordNetis üks tähendus – üks abielupool teise poole suhtes[2]. Olgu antud kontekstis õige tähendus (3), nagu lausejupis:

Naine ütles oma abikaasale, et...

Eesti WordNetis on sõna 'naine' kõik kolm tähendust üsna lähestikku (joonis 1). Naine (1) ja naine (3) on mõlemad naine (2) hüponüümid. See tähendab seda, et kui me mõõdame suvalisest punktist[3] sõna 'naine' erinevate tähenduste kaugusi, siis jõuame alati tähenduseni (2) enne kui tähenduseni (1) või (3). Joonisel 1 on vastavad teed toodud punktiiriga. Mõõtes mõistest 'abikaasa' kaugust sõna 'naine' erinevate tähendusteni, on tulemus järgmine (loeme kaari ehk järgneval skeemil ühekordseid nooli):

[2] Tegelikult on olukord keerulisem: sõnal 'abikaasa' on eesti WordNetis kolm tähendust, käesolev on märgitud kolmandana. 'Abikaasa 2' on abielumehe sünonüüm ja 'abikaasa 1' on abielunaise – 'naine' (3) – sünonüüm

[3] Eeldusel, et see ei asu hierarhias *allpool*: näiteks 'noorik' on 'naine' (3) hüponüüm ja tema puhul järgnev ei kehti.

```
abikaasa 3 → inimene 1 → naine 2 → naine 3 ⇒3
abikaasa 3 → inimene 1 → naine 2           ⇒2
   3 > 2,        järelikult  on  vastus   naine 2
```

... ja saadaksegi mõistliku meetodiga vale vastus!

Asja lahendamiseks on kaks teed: kas muuta algoritmi või mõistete paiknemist tesauruses. Vaatame lähemalt neid naise erinevaid tähendusi.

Esiteks, võib väita, et abielunaine on täiskasvanud, seega peaks olema 'naine 3' (abielunaine) 'naine 1' (täiskasvanud naine) hüponüüm. Selline olukord on kujutatud joonisel 2. Kui hierarhia oleks lahendatud nõnda, siis kontseptuaalse kauguse seisukohalt oleks asi veel hullem: kaugus abikaasa ja abielunaise vahel suureneks 3-lt 4-le. Tegelikult ei pruugi abielunaine olla üldse täiskasvanud naine. Ajaloost on teada juhtumeid, kus kuninglikud paarid laulatati abiellu juba õige noorelt.

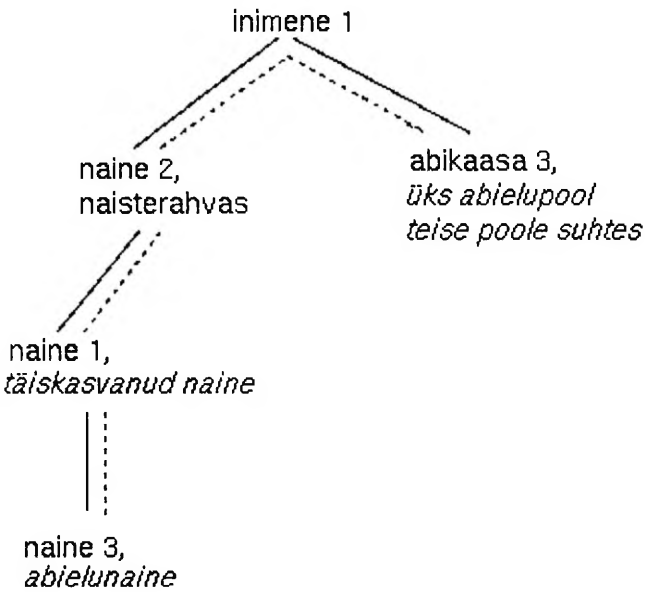
Jääb võimalus muuta hüperonüümia-hüponüümia hierarhiat nii, et 'naine 3' (abielunaine) oleks hoopis abikaasa hüponüüm, ega kuuluks üldse naisterahva hierarhiasse (joonis 3). Sel juhul lahendaks kontseptuaalsetel kaugustel põhinev sõnatähenduste ühestaja eeltoodud juhtumi alljärgnevalt:

```
abikaasa 3 →   naine 3                  ⇒    1
abikaasa 3 → inimene 1 → naine 2        ⇒    2
   1 < 2,       järelikult  on  vastus  naine 3
```

Samamoodi võiks 'naine 2' (täiskasvanud naine) olla hoopis 'täiskasvanu' hüponüüm ja üldse mitte kuuluda naisterahva hierarhiasse. Siis oleks 'naine' tähenduses (2) kergesti eristatav 'täiskasvanu' suhtes nii tähendusest (1) kui (3). Kuid kas me ei kaota olulist informatsiooni wordnetist, kui me hierarhiad sellisel viisil ümber teeme? Nii näiteks on 'noorik' 'naine 3' hüponüüm, ja kui meil esineb ühestatavas tekstis 'noorik', siis on olemas küll väike kaugus sõnast 'abikaasa', aga seda, et tegemist on naissoost inimesega, ei selgu kusagilt[4]. Mõnes rakenduses – näiteks intelligentses infootsingus – võib sellist seost vaja minna.

Siin oleks üks võimalik lahendus selline, kus 'naine 3' saab endale veel teise hüperonüümi, 'abikaasa'. Selline olukord on

[4] Täname Kadri Viderit selle tähelepaneku eest.



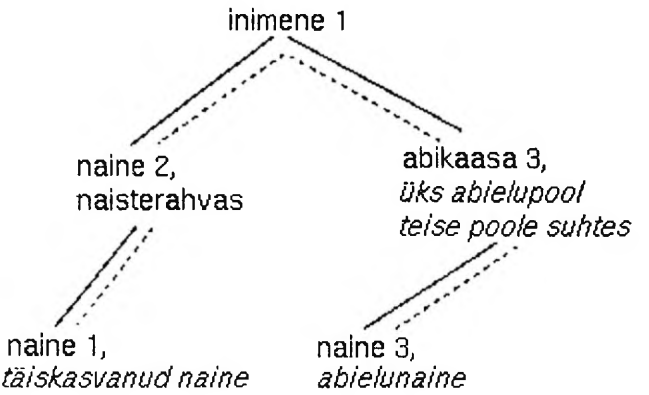
Joonis 2. Abielunaine on vahetanud kohta, paikneb hierarhias täiskasvanud naise all

Punktiiriga on tähistatud teekond mõistete 'abikaasa' ja 'abielunaine' (naine 3. tähenduses) vahel.

kujutatud joonisel 1, teist hüperonüümiasuhet tähistab ristikestega pikitud joon. Teiste hüperonüümide lubamine aga rikuks ära wordneti puustruktuuri. Ometi on Princetoni WordNetis kasutatud just sellist lahendust: sõnal 'wife' (abielunaine) on kaks hüperonüümi – 'woman, adult female' (täiskasvanud naine) ja 'spouse, partner' (abikaasa, partner).

6. Milline on naise kontekst tegelikult?

Sõna 'naine' esineb tähenduses (3) (abielunaine) 45 korral, tähenduses (1) (täiskasvanud naine) 11 korral ja tähenduses (2) (naisterahvas, olenemata vanusest) 3 korral. Tabelis 2 on toodud ülevaade erinevates tähendustes esinemisest.



Joonis 3. Kui abielunaine on abikaasa, aga naisterahva harusse ei kuulu, siis on kaugus abikaasa ja naise 3. tähenduse vahel väiksem kui teiste tähenduste vahel

Punktiiriga on tähistatud teekond mõistete 'abikaasa' ja 'abielunaine' (naine 3. tähenduses) vahel.

Tabel 2: Sõna 'naine' esinemine kontekstis

Tähendus	Esinemisi	Pärisnimi lauses	Mitmuse vormis
1 – täiskasvanud naine	11	1	5
2 – naine üldiselt	3	2	3
3 – abielunaine	45	19	0

Vaadates sõna 'naine' esinemist kontekstis, on nii väikese hulga pealt raske midagi põhjapanevat järeldada. Võiks oletada, et 'naine' (3) tähenduses (abielunaine) oleks eristatav selle järgi, et läheduses esineb sõna 'oma', 'tema' või midagi muud sellist. Paraku *semyhel* sellest suurt kasu ei oleks, kuna arvestatakse ainult läheduses olevaid samaliigilisi sõnu, antud juhul seega substantiive.

Siiski võib kontekstist üht-teist huvitavat leida: käesolevate korpusetekstide andmetel ei esine 'naine' kordagi (3) tähenduses (abielunaine) mitmuse vormis. Aga rohkem kui 42% juhtudest on samas lauses pärisnimi. Tähenduses (1) ja (2) esineb antud tekstides 'naine' kokku ligi pooltel juhtudest mitmuse vormis (tabel 2),

tähenduses (3) aga mitte kordagi. Ja ikkagi on see selline informatsioon, mida *semyhe* praegu arvesse ei võta.

7. Kokkuvõte ja järeldused

Sõnatähenduste ühestamisest on kasu ka tesauruse täiendamisele. Lisaks puuduvatele tähendustele võib leida kohti hierarhias, mis ülevaatamist vajavad. Siiski pole alati hierarhiate korrastamine nii lihtne, nagu esmapilgul paista võib. Kui tõepoolest hakata hierarhiaid selle järgi ümber tegema, kuidas paremini sõnatähendusi ühestada saaks, võib olemasolevatest hierarhiatest olulist infot kaduma minna. Seda saab vältida mitme hüperonüümi lisamisega, kuid siis pole hierarhia üheselt mõistetav – tegemist ei ole enam puustruktuuriga.

Teine võimalus oleks kontseptuaalsete kauguste arvutamisel arvestada ka teiste semantiliste suhetega. Agirre ja Rigau (1996) katsetasid ka meronüümia (osa–terviku) suhtega, aga vähemalt inglise keele puhul see erilisi paranemisi tulemustes ei andnud.

Olulist informatsiooni selle kohta, mis aitab sõna tähenduste ühestamisel, on võimalik leida uurides sõnade esinemist konkreetsetes vormides. Ülaltoodud naise näite puhul võimaldaks morfoloogilisest analüüsist saadavad andmed selle kohta, mis arvus sõna esineb, kallutada ühestamisotsust siia- või sinnapoole. Üks võimalus *semyhe* tulemusi parandada ongi ka detailsema morfoloogilise info arvesse võtmine.

Konteksti laiem analüüs võib samuti anda väga olulist informatsiooni selle kohta, mis tähenduses antud sõna esineb. *Semyhe* arvestab ainult samaliigiliste sõnade konteksti. Ülaltoodud *naise* näidet silmas pidades on *semyhet* juba kohandatud käituma pärisnimedega teisiti kui seni. Aga programm, mis arvestaks kõiki konteksti sõnu, töötab juba hoopis teistel põhimõtetel.

Kirjandus

Agirre, E.; Rigau, G. 1996. Word sense disambiguation using conceptual density. – Proceedings of the 16th International Conference on Computational Linguistics. Copenhagen. 16–22.

Ide, N.; Véronis, J. 1998. Introduction to the special issue on word sense disambiguation: The state of the art. – Computational Linguistics 24, 1–40.

Kahusk, N.; Orav, H.; Õim, H. ilmumas. Sensiting inflectionality: Estonian task for SENSEVAL-2. (Ilmub SENSEVAL-2 toimetistes)

Rada, R.; Mili, H.; Bicknell, E.; Blettner, M. 1989. Development and application of a metric on semantic nets. – IEEE Transactions on Systems, Man and Cybernetics, 19:1, 17–30.

Vider, K.; Kahusk, N.; Orav, H.; Õim, H.; Paldre, L. 2000. Eesti keele tesaurus. – Arvutuslingvistikalt inimesele. Tartu ülikooli üldkeeleteaduse õppetooli toimetised 1. Toim. T. Hennoste. Tartu. 127–152.

Vider, K.; Kaljurand, K. ilmumas. Automatic WSD: Does it make sense of Estonian? (Ilmub SENSEVAL-2 toimetistes)

Vider, K.; Orav, H. 1998. Sõna tasandilt mõiste ruumi. – Keel ja Kirjandus 1, 57–64.

Is there an upper limit to right-branching embedding of clauses?

Fred Karlsson
University of Helsinki

A common belief especially in generative linguistics is that a sentence is a sequence of clauses with no essential upper bounds. Surprisingly little empirical work has been done in 20th-century linguistics to check the truth of this assumption even if research on sentence composition has venerable traditions in philology and stylistics.

Clauses may be embedded in their matrix clauses in three different positions: initial, medial, and final. Consequently we may talk about initial, center-, and final embedding. Repeated initial embedding creates left-branching structures, repeated center-embedding self-embedded structures, and repeated final embedding right-branching structures. My topic in this paper, written to celebrate the 60th birthday of my old friend Haldur Õim, is whether there are restrictions on right-branching.

Linguistic intuition is fragile in other than so-called clear cases and cannot be expected to yield a reliable answer. I shall approach the problem in two other ways: by consulting available literature and existing large machine-readable corpora.

A terminological note is in order concerning the notion 'clause'. By clause I mean 'finite clause'. At the first stage of the analysis, it is important to keep finite and non-finite constructions apart because the claims will be different if all infinitival and participial constructions are included along with the finite ones. Thus, note the following example taken from the British part of the International Corpus of English (ICE-GB, one million words). Progressive levels of indentation indicate deeper levels of embedding:

	ICE-GB	FK
Selling is	depth 0	depth 0
what you do	depth 1	depth 1
to persuade people	depth 2	
to buy today	depth 3	
what you 've got	depth 4	depth 2
to offer to them today.	depth 5	

Given this restrictive finite-based definition of depth we now proceed to discuss some data. Alvar Ellegård (1978) used a sub-corpus of 128,000 words from the Brown Corpus in his analysis of the syntactic structure of English texts. The genres he studied were popular fiction, journalism, literary essays, and science. As clauses he accepted also non-finite verbs not governed by auxiliaries, which of course increases the syntactic depths reported. The total number of (finite and non-finite) clauses was 17,900. The maximum depth observed was 5 of which Ellegård (1978: 27–8) found around 50 instances, with the relative frequency 0.1–0.6% in the four registers being investigated. Depths 3 and 4 were collapsed in his statistics. Their relative frequencies ranged between 4% and 13%. As the 50 occurrences of depth 5 also contain non-finite clauses, whose share was almost 25% of all clauses, it can be safely concluded that pure sequences of five finite final embeddings must have been very rare in his sub-corpus if occurring at all.

Ikola *et al* (1989) investigated the syntax of spoken Finnish dialects and written standard Finnish using most of the machine-readable tagged corpus of the Finnish Syntax Archives (*Lauseopin arkisto*), University of Turku. The spoken corpus contains some 54,300 sentences (166,000 finite clauses; 885,000 words), the written part 15,600 sentences (27,300 finite clauses; 191,000 words). The material is not split up on initial, center- and final embedding, but even so it is equally relevant for assessing the maximal depth of embedding. Of course, the majority of the embeddings in this (and any) corpus are final.

Table 1. Frequency of depth of sub-clauses in spoken Finnish dialects and written standard Finnish (Ikola *et al* 1989: 18)

	Spoken		Written depth	
	N	%	N	%
1	42,864	84.5	5,863	85
2	6,884	13.6	877	13
3	858	1.7	94	1.4
4	109	0.2	12	0.2
5	11	(11)		
6	2	(5)		
7	3	(3)		
Total	50731	100	6865	100

The embedding depths of finite clauses in spoken and written Finnish are almost identical, which is interesting. Written language is not more complex in this regard, as many would expect. Embeddings occur down to depth 4, where a few instances may be encountered but they are extremely rare (0.2%). The corpus contains a couple of instances at depths 5–7 but the authors are hesitant about their interpretation and they state (p. 19) that these examples might be coding errors (therefore they are put in parentheses in table 1). All the data provided by Ellegård and Ikola et al. strongly indicate that 4 is the practical upper limit of final embedding of finite clauses, the 'usus maximum'. Instances of final embedding at 5–7 or even lower depths are extremely rare.

Similar results are easy to find in many other stylistic and philological studies of various languages. Final embeddings below depth 4 are extremely uncommon – although they might occasionally occur. An interesting example of depth 6 of final embedding was provided by Victor Yngve (1960: 460) in his classical paper on grammatical depth:

> The said rocker level is operated by means of a pair of opposed fingers
> which extend from a pitman
> that is oscillated by means of a crank stud
> which extends eccentrically from a shaft
> that is rotatably mounted in a bracket and has a worm gear
> thereon
> that is driven by a worm pinion
> which is mounted upon the drive shaft of the motor.
> (U. S. Patent)

The ICE-GB corpus is syntactically coded and therefore it is easy to spot structures according to desired search keys. In this corpus there is one instance of final embedding of depth 5, one of depth 6, and one of depth 8, which we quote here (capital 'F' indicates finite clauses, lower-case 'f' non-finite construction):

> The reason is
> that I think F-1
> this subject of symbolic representations of Christ has importance F-2
> because you know F-3
> that this was a problem F-4
> if that's the word F-5
> that assailed Byzantium in the twenties F-6
> when they convinced themselves F-7

that we shouldn't have portraits of Christ	F-8
shown in human form.	f-9

Note the reduced non-finite clause extending down to depth 9. Of coruse, instances like this are utterly rare, even if some more may be found in nursery rhymes and folklore. Still, they do not (in my mind) falsify the generalization established above, that the usus maximum of final embedding is 4.

References

Ellegård, Alvar. 1978. The Syntactic Structure of English Texts. A Computer-based Study of Four Kinds of Text in the Brown University Corpus. Göteborg, Sweden: Acta Universitatis Gothoburgensis. Gothenburg Studies in English, 43.

Ikola, Osmo; Palomäki, Ulla; Koitto, Anna-Kaisa 1989. Suomen murteiden lauseoppia ja tekstikielioppia. Suomalaisen Kirjallisuuden Seuran Toimituksia 511. Helsinki: SKS.

Yngve, Victor 1960. A model and an hypothesis for language structure. – Proceedings of the Americian Philosophical Society 104. 444–466.

Uudiste süntaks

Reet Kasik
Tartu ülikool

Käesolevas artiklis vaatlen Eesti ajalehtede uudiste lauseehitust. Vaatluse all on 21. augusti 2001. aasta Eesti Päevalehe ja Postimehe uudistelehekülјed, mis hõlmavad Eesti uudiseid, välisuudiseid ja majandusuudiseid. Kõrvale on jäetud arvamused, intervjuud, reportaažid, kultuuriarvustused jms, mis esindavad muid tekstiliike ja millel on seetõttu erinevad tekstilised omadused. Artikli alguses analüüsin kvantitatiivselt uudisteksti lauselisust, st millistest propositsioonidest koosnevad pindstruktuuri laused ja kuidas need propositsioonid on omavahel ühendatud. Seejärel vaatlen iga lauselisuse tasandit eraldi funktsionaalsest vaatepunktist ja püüan selgitada, millised tunnused iseloomustavad meediauudiste tekstimoodustust. Lausetasandite funktsionaalne iseloomustus piirdub siinses artiklis paratamatult üksikute, tekstiliigi seisukohalt kesksete tunnuste väljatoomisega, üksikasjalikum analüüs on pikemaajalise uuringu teema. Lausenäidete järel on ajalehe lühend (PM või EPL) ja lehekülg, kust näide pärineb.

Tekstilingvistilisest vaatepunktist koosneb iga pindstruktuuri lause ühest või mitmest süvastruktuuri elementaarlausest, nn propositsioonist. Propositsiooni moodustab süvastruktuuri predikaat koos oma argumentidega. Propositsioonid on tekstiaatomid, mida võib omavahel ühendada paljudel eri viisidel, nii et tekib erineva keerukusastmega pindstruktuuri lauseid. Tekstilingvistikas eristatakse kolmesuguseid seoseid: rinnastust, alistust ja sisestust. Algupärane elementaarlause esineb pindstruktuuris iseseisva lausena suhteliselt harva: enamasti on ta oma lauselisuse – s.t finiitverbi – kaotanud ja muutunud mingit tüüpi fraasiks. Lähtudes sellest, millise vormi algupärane süvalause pindstruktuuris on saanud, eristan järgnevas analüüsis kuus erinevat lauselisuse tasandit: pealaused, kõrvallaused, infiniittarindid, partitsiiptarindid, nominalisatsioonid ja predikaadita tarindid. Pealaused ja kõrvallaused on ka vormiliselt laused, infiniittarindid, partitsiiptarindid, nominalisatsioonid ja predikaadita noomenifraasid esindavad eri tasandi sisestusi, kus elementaarlause predikaat on infinitiivi, partitsiibi või verbaalnoomeni vormis või on

hoopis kadunud. Predikaadita sisestused kujutavad endast eri tüüpi täiendi- ja lisandifraase, mille taustal võib enamasti näha *olema*-verbiga predikatiiv- või omajalauset (*ajalehe peatoimetaja NN* < 1. *NN on peatoimetaja*; 2. *Ajalehel on peatoimetaja*). Eesti keelele iseloomulikud obliikvakäändeis substantiivtäiendid võivad seostuda ka muud tüüpi lausetega (*laused paberil* < *Laused on paberil*). Järgnevas analüüsitud näitelause koosneb 17 propositsioonist: 1 pealause, 1 kõrvallause, 2 infiniittarindit, 3 nominalisatsiooni, 10 predikaadita tarindit (peamiselt atribuutkonstruktsiooni). Iga propositsioon on noolest vasakul esitatud süvastruktuuri elementaarlausena ja noolest paremal pindstruktuuri fraasina, kusjuures on tehtud järgmised mööndused: mitmesõnalisi mõisteid (pärisnimed, arvud jms) ei ole lahutatud; rinnastatud argumendid on loetud üheks argumendiks, rinnastatud predikaadid on loetud eri lauseteks; lause- ja rõhulaiendeid, millel pole süntaktilist funktsiooni, ei ole arvesse võetud.

(1) Soov teenida välisvaluutat ning mitme kolmanda maailma riigi huvi Hiina sõjatehnika, ennekõike rakettide vastu, võivad maailma rahvarohkeima riigi muuta oluliseks tõkkeks teel ründerelvastuse leviku piiramise suunas, märgitakse dokumendis (PM, 9)

dokumendis märgitakse (PL)	
soov ja huvi võivad muuta (KL)	
(riik) soovib teenida	> *soov* teenida (NOM)
(riik) teenib välisvaluutat	> *teenida* välisvaluutat (INF)
riigil on huvi	> riigi huvi (ATR)
riike on mitu	> mitme riigi (ATR)
riigid on kolmandast maailmast	> kolmanda maailma riigi (ATR)
huvi on sõjatehnika vastu	> huvi sõjatehnika vastu (ATR)
sõjatehnika on ennekõike raketid	> sõjatehnika, ennekõike rakettide vastu (ATR)
(soov ja huvi) muudavad riigi tõkkeks	> *muuta* riigi tõkkeks (INF)
tõke on teel	> tõkkeks teel (ATR)
riik on rahvarohkeim	> rahvarohkeima riigi (ATR)
riik on maailmas	> maailma riigi (ATR)
tõke on oluline	> oluliseks tõkkeks(ATR)
tee on piiramise suunas	> teel piiramise suunas (ATR)
piiratakse levikut	> leviku *piiramise* (NOM)
ründerelvastus levib	> ründerelvastuse *leviku* (NOM)

Pindteksti moodustumine propositsioonidest on keeliti erinev. Tekstide lauselisust eri keeltes on võrrelnud kontrastiivsete uuringutega tegelevad tõlketeoreetikud (vt nt Ingo 1986, 1994). Ka eri tekstiliikide lausete komplekssus ja propositsioonide struktuur pindstruktuuris on erinev. Eesti keeles on eri tekstiliikide lausemoodustust

sellest aspektist võrrelnud Raili Remmelg oma seminaritöös (1999). Kuna Remmelg pole arvesse võtnud pindstruktuuris predikaadita esinevaid propositsioone, siis pole need andmed siinse analüüsiga kõrvutatavad. Et seostada meediateksti analüüsi andmeid usutavamalt uudisteksti kui tekstiliigiga, olen võrdluseks analüüsinud umbes sama lausehulga ilukirjandusteksti, 404 lauset August Gailiti romaani "Toomas Nipernaadi" algusest.[1]

Kui vaadata lauselisuse tasandite protsentuaalset jagunemist Postimehe ja Eesti Päevalehe uudistekstides, on võimalik võrdlevalt näha, mis algupärasest elementaarlausest on pindstruktuuris saanud.

	Postimees		**Eesti Päevaleht**		**Gailit**	
Lauseid kokku	496		337		404	
Propositsioone kokku	3652		2385		1695	
Propositsioone lauses	7,3		7,1		4,2	
	N	**%**	**N**	**%**	**N**	**%**
Pealaused	601	16,4	396	16,6	744	43,9
Kõrvallaused	469	12,8	363	15,2	342	20,2
Infiniittarindid	300	8,2	194	8,1	177	10,4
Partitsiiptarindid	210	5,8	125	5,2	51	3,0
Nominalisatsioonid	608	16,7	320	13,4	43	2,5
Predikaadita tarindid	1370	37,8	991	41,6	339	19,9

N – lausete arv

Mõlema analüüsitud ajalehe tulemused on omavahel sedavõrd sarnased ja niivõrd erinevad võrdluseks analüüsitud ilukirjandustekstist, et neid võib pidada tunnuslikuks uudistekstile kui tekstiliigile. Analüüsitud ilukirjandustekstis on pindstruktuuris lause vormis (pealauseid ja kõrvallauseid kokku) üle 64% lausetest ja pealauseid on üle kahe korra rohkem kui kõrvallauseid, veidi üle 10% on infinitiive ja ligi 20% predikaadita tarindeid (põhiliselt atribuute). Partitsiipe ja nomi-

[1] Muidugi ei ütle see midagi ilukirjanduse kui žanri lauseehituse kohta. Kasutan saadud andmeid vaid võrdlusmaterjalina meediateksti analüüsi taustaks.

nalisatsioone on väga vähe. Pealausete arv on ligi kaks korda suurem kui lausete arv, st palju on lausetevahelist rinnastust.

Uudisteksti suhtarvud on hoopis teistsugused. Umbes kolmandik propositsioonidest on pindstruktuuris lause vormis: Postimehes veidi vähem (29,8%) ja Eesti Päevalehes veidi rohkem (31, 8%). Pealausete suhtarv on sarnane, lausetevahelist rinnastust on vähe. Suurem erinevus puudutab kõrvallausete kasutamist: EPL-i lauselisus on suurem just kõrvallausete arvel. Kui Postimehe uudisteksti propositsioonidest esineb pindstruktuuris kõrvallause vormis alla 13%, siis EPL-i kõrvallausete hulk on üle 15%. Ka sisestatud tarindite kasutamise hierarhia on sarnane: kõige rohkem on predikaadita tarindeid, seejärel nominalisatsioone, infiniittarindeid ja kõige vähem partitsiiptarindeid. Eesti Päevaleht kasutab rohkem predikaadita fraase (EPL 41,6% ja PM 37,8%), Postimees aga rohkem nominalisatsioone (PM 16,7% ja EPL 13,4%). Ka infiniit- ja partitsiiptarindeid on Postimehes veidi rohkem, aga erinevus on alla protsendi.

Postimehe laused on komplekssemad: kui EPL-i lauses on keskmiselt 7,1 propositsiooni, siis PM-i lauses on propositsioone 7,3 (romaanitekstis keskmiselt 4,2 ja neistki pooled rinnastusseoses). Uudisteksti fraasid on suhteliselt pikad ja paljude laienditega, sest propositsioonides on rohkesti informatsiooni. See tähendab, et võimalikult palju potentsiaalsete argumentide kohti on täidetud. Järgmises lauses näiteks on predikaadil kuus argumenti: agent, vahend, lähtekoht, sihtkoht, tegevusaeg ja tegevuse sagedus, kusjuures kolmel argumendil on veel täiendid.

(2) *Oktoobris* **hakkab** *Eestist* 160-kohalise *lennukiga* **lendama** *iga nädal Gran Canariale* reisifirma *Horizon Travel.* (PM, 13)

Selline rohke informatsiooni kokkusurutus ühte fraasi ja võimalikult paljude fraaside põimumine üheks lauseks tundub olevat iseloomulik just uudistežanrile.

(3) Kaitsepolitsei (kapo) algatas Männili (81) väidetavate sõjakuritegude uurimiseks kriminaalmenetluse tänavu 29. märtsil seoses Simon Wiesenthali Keskuse direktori Efraim Zuroffi poolt peaminister Mart Laarile saadetud kirjas sisalduvate väidetega. (PM, 4)

(4) Kaitsepolitsei loodab rohkem kui kuus aastat kestnud uurimise järel jõuda kriminaalasja uurimisega lähikuudel lõpule seitsme endise miilitsa ja viie julgeolekutöötaja suhtes, keda süüdistatakse seotuses saarlaste küüditamisega Siberisse. (EPL, 5)

Kokkusurutusel on ka varjukülgi: peale raskesti mõistetavuse tekitab propositsioonide kuhjamine nii keerulisi lauseid, et autoreil–toimetajatel on kohati raskusi nende süntaktilise vormistamisega. Eriti paistab keeletoimetajast puudust tundvat Eesti Päevaleht. Mõned näited:

(5) Valitsus leidis, et Narva Elektrijaamade vähemusosaluse müük tagab elektrienergia varustuskindluse pikaks ajaks stabiilse hinnaga ja pannes AS-ile Narva Elektrijaamad kohustuse tarnida Eesti Energiale siseturu vajadusi rahuldav hulk elektrienergiat. (EPL, 2)

(6) Rahvaliidu esimees Villu Reiljan kinnitas, et Meri käitumise põhjuseks võib olla rahulolematus valitsusega ja eesotsas Mart Laariga. (EPL, 3)

(7) "Need, kes arvavad, et oleme näidanud end eriti usaldusväärse partnerina rikkudes ja tühistades saavutatud kokkulepet NATO juhtriigiga, siis võib-olla nad elavad mingis teises välispoliitilises maailmas kui kõik need lääne inimesed, alates USA välisministrist, kellega mina olen arutanud NATO laienemist," lausus Ilves. (EPL, 1)

(8) "Ümarlaua hinnagul tuleb poliitilistel jõududel teha kõik, et president valitaks ausas konkurentsis, ning toetas seisukohta, et kõik presidendi valimiseks sõlmitavad kokkulepped peaksid olema avalikud," teatasid ürituse korraldanud Mõõdukad. (PM, 3)

1. Pealaused

Klassikaliselt eeldatakse ajalehe uudissõnumilt vastuseid kuuele põhiküsimustele: kes, kus, millal, mida, kuidas ja miks (vt nt Peegel 1970: 125). Sellisele klassikalisele uudiste funktsionaalsele skeemile vastavad tänapäeva Eesti lehtedes veel üksnes sõnumid õnnetustest ja kuritegudest.

(9) **Neli purjus ajateenijat võtsid** ööl vastu pühapäeva Lääne-Virumaal Tapal asuvast sõjaväeosast omavoliliselt veoauto ja **rammisid** sellega Ambla asulas elumaja. (PM, 6)

(10) **Ukraina päritolu 27-aastane mees tappis** Californias oma raseda naise, **sõitis** seejärel teise Sacramento eeslinna ja **tappis** seal veel neli inimest ning **põgenes** siis koos oma kolmeaastase pojaga. (EPL, 8)

Sageli on ka sellistes sõnumites representeeritud mitte tegu (Kes mida tegi?), vaid sündmus (Mis juhtus? Milline on olukord?).

(11) Pärnu politseiprefektuuri arestimaja **valvur jõudis** üleeile õhtul purjuspäi tööle, mees **saadeti** ekspertiisi ja seejärel koju kainenema. (PM, 5)

(12) **Veretöö leidis aset** Sacramento Nord Islands'i eeslinnas esmaspäeva hommikul. (EPL, 8)

(13) Loomakaitsjate **ajanappus jätab** õnnetud loomad kaitseta. (PM, 5)

Valdav enamik tänapäeva Eesti uudissõnumitest on tekstid tekstidest: mida keegi kuskil ütles ja mida teised selle peale ütlesid. See tähendab, et uudissõnumit ei vormista ajakirjanik, vaid keegi teine. Eriti teravalt kirjutas sellest kommunikatsiooniuurija Elihu Katz seoses Pärsia lahe sõja valgustamisega ajakirjanduses (Katz 1992). Ta väitis, et ajakirjanduse lõpp on alanud, kui ajakirjaniku osaks jääb üksnes kanalit avava ja sulgeva väravavahi roll, aga sõnumi sisu formuleerivad asjaga seotud kindralid. Samast asjast on kirjutanud teisedki ajakirjandusuurijad: kuigi ajakirjaniku loomulik koostööpartner peaks olema lugev avalikkus, on selleks hoopis poliitika ja majanduse juhtisikud (Walton 1982). Ajakirjanik peidab ennast oma tootesse: sõnumid tulevad allikatelt ja ajatundjatelt. Ajakirjanik üksnes viib lugeja sündmuskohale, näitab sõrmega sündmuse poole, aga ei nõustu ütlema midagi oma nimel, vaid üksnes autoriteetide kaudu (Hallin 1992). Ajakirjanik on keeruliste asjade ees liiga sageli naiivselt usaldav ja muutub seetõttu meie ristuvate huvide maailmas kergesti, ise seda märkamata, propaganda vahendajaks (Hadenius 1992). Ametlike allikate positsioon ajakirjanduses on kõigutamatu. Allikaks võtab ajakirjanik kõige parema meelega sellise, keda on ennegi sageli kasutatud, ja allikaga kontakteerumise vahenditest on valitsev telefon, mitte trükitud sõna. Telefonisuhtlus on tekstidest selgelt äratuntav: üneemid, kokutamised ja sõnakordused võtab teksti üleskirjutaja välja, aga lauseehitus reedab suulise jutuajamise. Ähmastub ka uudiste ja arvamuskirjutiste piir: uudissõnumi jaoks polegi vaja sündmust, uudise teemaks piisab kommentaaridest. Sellest vaatepunktist sisaldab uudis arvamuskirjutise tekstuaalseid jooni ning žanride piirid on lahtised ja ähmased (Heikkinen 1999: 191).

Tekstianalüüs kinnitab, et praeguse Eesti ajakirjandusele on iseloomulik puhtalt tekstipõhiste uudiste koostamine. Enamik uudissõnumeid algab teate allikat sisaldava lausega. See on vormistatud iseseisvaks pealauseks, millele järgneb alistatud kõrvallause (14) või sisestatud fraas teate sisuga (15, 16). Teate allikas võib olla väljendatud ka pealausesse sisestatud lauselühendiga (17). Refereerivale esimesele lausele järgneb teine lause otsese tsitaadi (14) või täpsustava refereeringuga (17).

(14) **Valitsus teatas** eile üksmeelselt, et Narva Elektrijaamade müüki USA firmale NRG Energy ei katkestata, kuid pidas õigeks, et peaminister Mart Laar läheb homme ise riigikogu ette tehingut puuduta-

vatele küsimustele vastama. "Need kes arvavad, et me oleme näidanud end eriti usaldatava partnerina, rikkudes ja tühistades saavutatud kokkuleppeid NATO juhtriigiga, elavad võib-olla mingis teises välispoliitilises maailmas kui kõik lääne inimesed, kellega mina olen arutanud NATO laienemist," **lausus** välisminister **Toomas Henrik Ilves**. (EPL, 3).

(15) Concordia ülikooli meediateaduskonna dekaan ja endine Eesti televisiooni peadirektor **Hagi Šein peab** valitsuse eilset otsust kaotada ETVst reklaam praegustes oludes parimaks lahenduseks. (PM,5)

(16) Õiguskantler **Allar Jõks tegi** eile majandusminister Mihkel Pärnojale **ettepaneku** kaaluda Narva Elektrijaamade erastamise peatamist. (EPL, 3)

(17) Eesti Kohtuarstliku Ekspertiisibüroo **töötajate hinnangul** ei saa joobeekspertiisiks nimetada vereanalüüsi tegemist vananenud seadmetega haiglates ning roolijoodikute vere uurimine tuleb tuua kohtulaborisse. EKEB peaekspert **Marika Väli ütles** Postimehele, et purjus juhtide joobeekspertiiside tegemisel ei saa viimane sõna jääda ükskõik kui täpsele politsei mõõtevahendile, vaid inimesel peab olema alati õigus lasta teha vereanalüüs. (PM, 6)

Esimesele lausele järgnev uudistekst koosnebki valdavalt tsitaatidest ja refereeringutest. Kõigist analüüsitud lauseist nimetas teate allikat kas otse (*NN ütles et...*) või nominaliseeritult (*NN-i sõnul*) Postimehes 45,4% ja Eesti Päevalehes 36,2% lauseist. Kui võtta arvesse ka rinnastatud pealaused, siis väljendab verbaalset tegevust 37,6% Postimehe ja 33,3% Eesti Päevalehe pealausete predikaatidest.[2] Pealaused on valdavalt lühikesed, koosnedes subjektist ja predikaadist. Subjekt on enamasti individualiseeritud nimeliselt või pronoomeniga *ta*, harvem kollektiveeritud. Predikaadiks on ütlemisverb (*ütles, rääkis, teatas, kinnitas, märkis, möönis, sõnas, tõdes, lausus, lisas, leidis, kirjeldas, jutustas, toonitas, mainis, vastas...*). Suhteliselt harva (PM 15 korda, EPL 10 korda) on neutraalse ütlemisverbi asemel kasutatud mentaalset verbi, millega teksti autor annab hinnagu refereeritava suhtumisele. Saatelauses kasutatud mentaalsed verbid on *kardavad, loodab, tekitab ärevust, ei nõustunud, ei soovinud, muretses, soostus, ei osanud öelda, on veendunud, andis mõista, lubas, tunneb muret, süüdistas, võttis kokku, arvas*. Iseloomulik on, et mentaalset tegevust väljendavat predikaati ei kasutata otsese kõne saate-

[2] Tegelikult on laenatud teksti osa veel suurem, sest saatelauseta, vormiliselt sõltumatud tsitaadid on kvantitatiivses analüüsis loetud iseseisvateks (pea)lauseteks.

verbina, vaid üksnes refereeringutega seoses, kus piir mentaalse tegevuse (tundmine, tahtmine, arvamine) ja selle verbaalse väljendamise vahel pole kuigi selge.

(18) Otepääl Eesti Spordigümnaasiumis tulevasi tippsportlasi koolitanud **õpetajad kardavad**, et kooli ülevõtmine Audentese spordikooli poolt võib osale neist tähendada hüvitiseta töökoha kaotamist. (PM, 1)

(19) Eesti **poliitikategelased** eri leeridest **loodavad**, et Narva Elektrijaamade osaluse müügi asjus tekkinud pretsedenditu tipptasemel sõnavahetus Eesti ning Ameerika Ühendriikide vahel on mööduv nähtus ega halvenda riikide vahelisi suhteid. (PM, 2)

(20) Reformierakonna fraktsiooni esimees **Jürgen Ligi ei soovinud** spekuleerida peaministri ja presidendikoha jagamise teemadel. (EPL, 3)

Individualiseeritud agenti võidakse edaspidi samas tekstis nimetada ka sünonüümi või parafraasiga. Näiteks *Järvamaal Amblas õnnetuspaiga läheduses elav naine, kes ei soovinud oma nime avaldada* on järgnevas tekstis ka *pensionäripõlve pidav naisterahvas, vanadaam, pensionär, naisterahvas*. Ametiisikute puhul kasutatakse nime varieerimiseks ametikoha või staatuse nimetust (21) või piirdutaksegi ametikoha nimetamisega, nt *pressiesindaja* (22).

(21) "USA kui presidentaalse korraga riik oli hämmingus, sest kui neil teeks president sellise avalduse, siis see tähendaks ka valitsuse seisukohta, kuid meil kui parlamentaarses riigis pole presidendi seisukoht võetav üheselt valitsuse seisukohana, " **rääkis Ilves. Välisminister lisas**, et kui Eesti valitsus on jõudnud kokkuleppele mistahes teise riigi erafirmaga ja kui siis meie valitsus tühistaks selle kokkuleppe andmata adekvaatseid põhjuseid, siis teine riik vaatab, et meie valitsus pole usaldusväärne. (EPL, 2)

(22) **Ida-Viru politsei pressiesindaja sõnul** selgitasid eksperdid, et garaažis plahvatas 80–82 millimeetrise kaliibriga miinipilduja miin. "Kuidas miin päästekompanii garaaži sattus ja millistel asjaoludel ta plahvatas, peab selgitama edasine uurimine," **ütles politsei pressiesindaja**. (PM,4)

Alati ei ole tegijad ja rääkijad individualiseeritud, vaid nad on esitatud homogeense rühmana. Eesti Päevaleht on tunduvalt "mitmuslikum" kui Postimees, viimane toetub valdavalt individualiseeritud allikaile. Füüsilist või verbaalset tegevust väljendava predikaadiga seostub kollektiveeritud tegija (kas kollektiivsubstantiiv või mitmusevorm) 20,5% EPL-i pealausetest, neist pooltel juhtudel on tegemist verbaalse tegevuse agendiga. Postimehe uudiste pealausetes on vastav näitaja 12,8% ja neist vaevalt viiendik seostub verbaalse

tegevusega. Kõige sagedamini on tegijaks ja ütlejaks valitsus (*valitsus otsustas, kiitis heaks, avaldas lootust, teatas, leidis, tõdes...*), aga ka ministeeriumid, erakonnad, komisjonid, organisatsioonid jm institutsioonid (23, 24). Seevastu on kummaski lehes vaid üksikuid pealauseid, kus tegija on nimetatud kvantitatiivselt määratud hulgana (25). Nagu tekstiuurijad on täheldanud, on kvantitatiivsete andmete rohkus meediatekstile iseloomulik (Fowler 1991: 166–169 räägib kvantifitseerimise retoorikast), aga siin uuritavates tekstides on sellisedki andmed enamasti esitatud tsitaatide või refereeringute osana (26).

(23) "Strateegilist keskkonnamõju hindamist läbimata pole planeeritav erastamise kaudu realiseeritav põlevkivienergeetika ümberstruktureerimine kooskõlas Eesti seadustega," **kinnitasid keskkonnakaitsjad**. (EPL, 3)

(24) Põllumajandussaaduste eksport on tänavu võrreldes 2000. aasta esimese poolaastaga suurenenud 61,6 protsenti, **teatas põllumajandusministeerium**. (EPL, 7)

(25) Iseseisvuse taastamise aastapäeval **külastas** Tallinna rekordarv, **üle 5000 turisti**, kes tutvusid linna vaatamisväärsuste ja kaubanduskeskustega. (EPL, 6)

(26) Tallinna linna ekskursioonidel käis kokku 2038 turisti, keda veeti mööda linna ringi 96 bussiga, kus igas sõidukis jagas linna kohta teavet üks giid, **teatas Raepress**. (EPL, 6)

2. Kõrvallaused

Kõik tsitaadid, mille juurde kuulub saatelause, on siinses ülevaates analüüsitud kõrvallauseteks. Kui lisada ka allika ütlusi või mentaalseid mõtteid ja tundeid refereerivad kõrvallaused, tõuseb objektlausete osakaal ligikaudu 90%-ni kõigist esimese astme kõrvallausetest. Kui pidada meedia põhifunktsiooniks informatsiooni vahendamist lugejale, siis valdav osa informatsioonist jõuab lugejani kõrvallausete kaudu. Vastukaaluks lühikestele pealausetele on uudisteksti kõrvallaused väga pikad ja komplekssed. Enamasti on ühte lausesse ühendatud mitu eri liiki ja eri tasandi kõrvallauset (27). Iseloomulikud objektkõrvallause osad on tingimuslaused ja põhjuslaused, mõnel määral ka relatiivlaused. Kui kõrvallausega esitatud referaat algab omakorda kõrvallausega (nt tingimuslausega), siis on tavalised kaks eri astme alistavat sidesõna kõrvuti (28, 29). Ka enamik infiniitkonstruktsioone on sisestatud just kõrvallausetesse.

(27) "Arvestades seda, et surmanuhtlust käsitlev diskussioon on juba parlamendis kord läbi arutatud, ei saa pidada võimalikuks ega ka otstarbekaks surmanuhtluse diskussiooni uuesti alustamist, sest otsus inimõiguste konventsiooni kuuenda lisaprotokolliga ühinemise ehk surmanuhtluse keelustamise kohta on parlamendis juba langetatud," seisab justiitsminister Märt Raski (pildil) vastuses internetiportaalis Täna Otsustan Mina (TOM) esitatud ettepanekule. (PM, 3)

(28) Isamaaliitlasest riigikogu liige Peeter Olesk nentis, et Tulviste kandidatuuri ülesseadmine riigikogus pole vähetõenäoline, kuid seda ei tehta enne, kui tema läbiminek garanteeritakse. Olesk lisas, **et kui** parlamendivoorudes läheks presidendi valimine läbi väga paljude tehingute hinnaga, siis tulemuseks on nõrk president. (EPL, 3)

(29) Samas olevat Putini ootamatu Novgorodimaale saabumine tekitanud tõsist peavalu oblasti juhtkonnale, keda ehmatas Putini meeskonnalt tulnud teade, **et kuna** tegemist on igati mitteametliku visiidiga, siis ei soovi president mingit traditsioonilist soola-leivaga vastuvõtmist. (EPL, 8)

Kõrvallausetele lisab pikkust ka argumentide suur hulk propositsioonides.

(30) Eesti Humanitaarinstituudi rektor Jaan Tamm märkis, et kuna nende *õppeasutus* **muutub** *siis erakõrgkoolist avalik-õiguslikuks*, on kõige raskem osa liitumisel juriidilised terminid paika saada. (EPL, 6)

(31) Puhkuse esimene päev möödus siiski töiselt, sest kuni läinud kolmapäeva *õhtuni* **arutas** *Putin Peterburis* kohalike *ametnikega* Loode-Venemaa *probleeme*. (EPL, 8)

(32) Arhiividest saadud vastustest selgub, et *Männil* **töötas** *aastatel 1941 ja 1942* Tallinn Harju prefektuuri poliitilises *politseis assistendina ja vanemassistendina* ning Eesti *julgeolekupolitseis assistendina*. (PM, 4)

Peale objektlausetele alistatud mitut liiki teise astme kõrvallausete laiendatakse kõrvallausetega ka pealauses nimetatud osalisi (33). Selline lausemoodustus torkab silma rohkem Eesti Päevalehes, kus üks refereerimisviise tundub olevat refereeritava teksti jagamine tsitaadi ja relatiivlause vahel (34–36).

(33) Erinevalt oma eelkäijast **Boriss Jeltsinist, kes** puhkuse ajal enamasti napsitas ja jahti pidas ning viimastel aastatel end aina ravis, käis president Vladimir Putin ligi nädala kestnud puhkuse ajal ühest kloostrist sisse ja teisest kirikust välja. (EPL, 8)

(34) "Ma ei kujuta ette, mis talvel saab, kui kett külmub, " ütles **Kolman, kelle sõnul** võiks 11-aastase Robi küll varjupaika tuua, kuid peremeest nii vanale koerale enam ei leia. (EPL, 5)

(35) "Meil on siseministri käsk kontrollida töötajaid. Pealegi on politseil ranged reeglid, kuidas valvekorda alustada," ütles **Moora, kelle sõnul** annavad töökaastased lõhnadega tööle ilmunud ametniku välja. (EPL, 5)

(36) Siedentopi täiendavad oma artiklitega Policy Review's **Lee Casey ja David Rivkin, kes tõdevad**, et "Euroopa Projekti südameks on idee universaalsest võimust, mis ulatuks üle kogu Euroopa". (EPL, 8)

3. Infinitiivid

Kui suurem osa võrdluseks analüüsitud ilukirjandusteksti infinitiividest on *des*-lauselühendid, siis uudisteksti infinitiividest moodustavad valdava osa *ma-* ja eriti *da*-infinitiivid, mis esinevad koos modaalverbidega. *ma*-infinitiivi käändvorme ja *des*-gerundiivi esines tunduvalt vähem, kokku umbes 15%, ja neistki moodustasid märkimisväärse osa grammatikaliseerunud vormid *hoolimata, vaatamata, sõltumata, kahtlemata, alates, võrreldes, arvestades* (verbivormide grammatikaliseerumisest vt Uuspõld 2001).

(37) Kui inglased on palju kohanimesid Walesis vägisi enda keelele kohandanud, **nimetades** näiteks Caerdyffi Cardiffiks, siis küla Llanfairpwllgwyngyllgogerychwymdrobwllllantysiliogogogoch nimi tehti 1870. aastal selliseks just nimelt inglaste kiuste, **luues** koha, mille nime ei ole võimalik ei lühendada ega võõra päritoluga inimestel ka hääldada. (EPL, 9)

(38) Põllumajandussaaduste eksport on tänavu mullu esimese poolaastaga **võrreldes** suurenenud 61,6 protsenti, teatas põllumajandusministeerium. (PM, 11)

(39) Tallinna maa-ameti osakonnajuhataja asetäitja Mari Heinsoo ütles, et amet **on ette valmistamas** linnavalitsuse otsust kahe 600-ruutmeetrise krundi jätmiseks riigi omandisse. (PM, 6)

ma- ja *da*-nfinitiivi nõudvatest verbidest esinevad pealauses koos agendist subjektiga *otsustama*, kõige sagedamini valitsuse tegevuse väljendajana (40), aga ka tulevikule viitavad tahtmisverbid *soovima, tahtma, kavatsema, üritama, plaanima, lubama, püüdma, lootma* (41–43). Ka suur osa meediauudiste kõrvallausetes sisalduvatest refereeritud sõnumitest on tegelikult suunatud tulevikku, väljendades, mida kavatsetakse ja plaanitakse teha või hakatakse tegema (44, 45).

(40) Valitsus **otsustas** eile **eraldada** reservist 250 000 krooni, et kõik vangid saaksid läbida HIV-analüüsi.

(41) Eesti Energia **soovib** Paljassaare tippu **püsti panna** kaks ligi 60-meetrise masti ja 40-meetrise tiivikudiameetriga tuulikut, kuid taotleb selleks linnavalitsuselt kooskõlastust kahe 600-ruutmeetrise krundi jätmiseks riigi omandisse. (PM, 6)

(42) Tallinna linnakohus **üritab** täna **alustada** kohtuprotsessi Stockmanni kaubamajas plahvatusi korraldanud viie mehe ja ühe naise üle. (PM, 6)

(43) Piritale loodavalt koolilt **hakkab** linn **ostma** põhikooli teenust ehk siis kuni üheksanda klassini võivad õpilased Tamjärve koolis õppida tasuta. (EPL, 6)

(44) Rahandusminister Siim Kallas sõdib maksude tõstmise ähvardusega vastu võimuliidu partnerite plaanile, mille kohaselt **hakkaks** riik tulevast aastast **toetama** kõiki lapsi 300 krooniga kuus. (EPL, 1)

(45) "**Tahame teha** infotehnoloogia kättesaadavaks senisest enamatele inimestele," põhjendas Microsoft Baltikumi juht Bo Kruse Wordi ja Outlooki eestindamist. (PM, 3)

Suurem osa modaalverbi laiendavatest infiniittarinditest paikneb kõrvallauses või mingil muul viisil selgelt eristuvas refereeritavas tekstis. Nagu eespool nägime, väldivad neutraalsust taotlevad uudisteksti autorid isegi ütlemisverbi valikuga suhtumise väljendamist, seetõttu hoidutakse autoritekstis ka suhtumist väljendavatest modaalverbidest. Tsitaatides ja refereeringutes väljendavad küsitletud modaalse tarindiga kõige sagedamini võimalikkust: mis võib toimuda, mida on võimalik teha või oleks võimalik teha teatud tingimustel. Potentsiaalsuse mainimine ei seo ütlejat mingil viisil. *võima*-verb (kaasa arvatud *on võimalik, võimaldama* ja *võimalus*) koos *da*-infinitiiviga tuleb ette üle kahe korra rohkem kui sageduselt järgnevad modaalverbid *pidama, saama, hakkama* ja *tuleb*.

(46) "Suurem vaidlus alles läheb lahti ja seda, mis seal **saada võib**, on mul hetkel küll raske ennustada," lausus sotsiaalminister Eiki Nestor. (PM, 3)

(47) "Sellest **võib tekkida** riigieelarvele lisakoormus, kuna litsentsitasude ja reklaamist saadava raha vahe **võib ulatuda** kümnete miljonite kroonideni," rääkis Šein Postimehele. (PM, 5)

(48) "Kuna ekspertiisi otsus **võib minna** vaidlustamisele, on meil kokkulepe, et piirdume väljendiga keskmine joove," ütles Kukk. (EPL, 5)

(49) Variku sõnul **võidakse** Mõõdukate kandidaat Andres Tarand riigikogus **üles seada**, kui kolmikliit kokkuleppele ei jõua. (EPL, 3)

Kuna küsitletavad on meedia spetsiifikast lähtudes enamasti poliitikud või mingi ala spetsialistid, siis esineb referaattekstis üsna palju ka direktiivset modaalsust: *tuleb, peab, on vajadus*, eitavas lauses vastavalt *ei saa, ei tohi*.

(50) Rektorite ettepanekul **peab** uus kõrgkool koos allasutustega **olema** avalik-õigusliku ülikooli staatuses. (PM, 5)

(51) "Kui ühiskond tahab ühe käega toetusi jagada, siis teise käega **tuleb** ka makse **tõsta**, tuleb teha lihtsalt valik," ütles Kallas Postimehele. (PM, 1)

(52) Valitsuskabineti liikmed rõhutasid BNS-i teatel, et samas **tuleb tagada** soodustingimused alternatiivenergeetika arendamiseks. (EPL, 2)

(53) "Võõrkeelne sõnavara **ei tohi piirata** arvuti kasutamise võimalust ja väljendada töö efektiivsust." (PM, 3)

Uurijad on pannud tähele, et meediatekstile on iseloomulik kasutada modaalverbe tingivas kõneviisis (vt nt Heikkinen 1999). Tingiva kõneviisi vormidel *peaks, tuleks, võiks, ei tohiks* võib uuritavas ainestikus eristada kolm modaalset tähendust: ühelt poolt nad pehmendavad direktiivsust, muutes selle kindlast väitest soovituseks (54), teiselt poolt väljendavad (ebakindlat) lootust tuleviku suhtes (55, 56), kolmandaks sisaldavad implitsiitset kriitilist suhtumist, et asjad tegelikult ei ole nii, nagu on soovitav (57).

(54) Mõõdukad teatasid ümarlauda kokku kutsudes möödunud nädalal, et riigikogu **peaks püüdma täita** oma põhiseaduslikku kohust, mille järgi on presidendi valimine eelkõige riigikogu ülesanne. (EPL, 5)

(55) Samas loodab direktor, et erakool Audentes jõuab õpetajatele rohkem palka maksta, mis **peaks** hetkel herilasepesana sumisevat õpetajaskonda veidi **rahustama**. (PM, 1)

(56) Lõpliku heakskiidu NATO rahuvalveväe saatmiseks Makedooniasse **peaks andma** täna NATO liikmesriikide suursaadikud. (EPL, 1)

(57) Sepa hinnagul **tuleks** segaduste vältimiseks **sätestada** üheselt mõistetav kontsentratsiooniühik. (PM, 5)

4. Partitsiibid

Võrdluseks analüüsitud ilukirjandustekstis esines partitsiipe vähe ja peamiselt oli tegemist laienditeta täiendilise *v*-partitsiibiga (*haukuv koer, sahisevad puulehed*). Ajalehe partitsiiptarinditest võib uudistekstile iseloomulikuna esile tõsta isikunime täiendavat partitsiip-

tarindit, mille funktsiooniks on karakteriseerida põhisõnaks olevat isikut. Millist informatsiooni noomenifraasi põhjaks oleva isiku kohta pakutakse, sõltub uudise liigist ja osalise nimetamisest. Nimeliselt individualiseeritud ametiisikuid iseloomustatakse nende tegevuse või ametiseisundi järgi (58). Õnnetustest ja kuritegudest rääkivates uudistes esineb osaliste ja rääkijatena ka nn tavalisi inimesi ja juhuslikke pealtnägijaid, keda nimepidi ei nimetata, vaid karakteriseeritakse soo või isiklike suhete järgi (*mees, naine, naaber*). Selliseid osalisi laiendav partitsiiptarind lisab isiku kohta muidki isiklikke omadusi, nt elukoha või sotsiaalse staatuse (59–61).

(58) *Riigikogus Koonderakonna fraktsiooni juhtiv* **Siimann** ütles, et lähem eesmärk on osaleda kohalike omavalitsuste valimistel ja kui ühendus saab nendel piisava toetuse, siis võetakse suund erakonna loomisele. (EPL, 5)

(59) "Õnn, et poistel kaelad terveks jäid," ütles *Järvamaal Amblas õnnetuspaiga läheduses elav* **naine**, kes ei soovinud oma nime ajalehes avaldada. (PM, 6)

(60) *Urvaste vallas Lümatu külas elav* **naine** pöördus esmaspäeva hommikul politseisse teatega, et tema laudast on kadunud kaks noort lammast. (PM, 6)

(61) *Pensionäripõlve pidav* **naisterahvas** rääkis, et luges tol ööl kella kaheni raamatut. (PM, 6)

Isikut märkiva osalise juurde kuuluva partitsiiptarindi teine tavalisem funktsioon on anda informatsiooni isiku tegevuse või seisundi kohta konkreetses, just selle lausega edasiantavas situatsioonis (62–64).

(62) *Valitsuskabineti istungil viibinud* Eesti Energia juht **Gunnar Okk** tõdes, et NRG Energy pakutav tehnoloogia on keskkonnasõbralikum kui seni kasutatu, veelgi enam saastekoormust vähendav. (EPL, 3)

(63) *Peagi ametist lahkuva* president **Meri** viimase aja käitumist, mis jättis iseseisvuspäeva pidulikust õhtusöögist ilma peaminister Mart Laari, tema osalemist Keskerakonna korraldatud konverentsil ning avaldust NRG tehingu peatamise kohta, ei pidanud Olesk ootamatuks. (EPL, 3)

(64) *Eesti internetiriiki vedanud* **Viik** on kriitiline e-riigi projektide suhtes. (PM, 12)

Isikut laiendava partitsiiptarindi struktuuri ja funktsioone meediatekstis on lähemalt uurinud oma bakalaureusetöös Evelin Kivilo (2001). Tema andmetel on ligi kolmandik uuritavatest lausetest (500st lausest 155) sellise infojaotusega, kus predikaatverb väljendab osalise verbaalset tegevust ja muu selles situatsioonis oluline tegevus

või osalise rolli iseloomustav seisund on väljendatud partitsiipse sekundaartarindiga. Osalise tegevust (65) ja seisundit (66) väljendavaid partitsiiptarindeid on enam-vähem võrdselt.

(65) *Nädala vältel rahvusraamatukogu alla 16aastaseid külastajaid* **jälginud** lugemissaalide **töötajad kinnitasid**, et... (EPL, 3.11.00)

(66) Koos kolleegide Liisa Pakosta ja Heiki Kivimaaga sadamaala konkursi žüriis **olnud** Tallinna abilinnapea **Priit Vilba sõnul**... (PM 6.2.01)

Informatsiooni jaotumisel pealause ja sisestatud tarindite vahel ei ole kriteeriumiks uue ja tuntud info või kirjutaja meelest olulisema ja ebaolulisema info hierarhiline vastandamine. Suur osa partitsiiptarindis sisalduvast informatsioonist on samuti esmakordselt mainitav ja lugejale täiesti uus. Propositsioonide hirarhia aluseks on predikaatverbi funktsioon. Kui lauseks ühendatavate propositsioonide hulgas on sama osalise eri tegevuste seas ka verbaalset tegevust väljendav lause, siis see fokuseeritakse alati, muu esitatakse sisestatud tarindite vormis tausta või seletustena.

5. Nominalisatsioonid

Nominalisatsioonide poolest on eriti rikas Postimehe uudistekst, kus neid on rohkem kui pealauseid ja kõrvallauseid. Eesti Päevaleht kasutab nominalisatsioone mõnevõrra vähem, aga "tegevuste sooritamise läbiviimise organiseeriseerimise" tüüpi väljendus ei ole eesti ajakirjandusest kuhugi kadunud. Eriti hakkab see silma BNS-i uudistes ja ametiisikute juttu refereerivates tekstiosades. Nominalisatsioonide funktsioonist ja struktuurist võib esile tõsta kaks omadust, mida on põhjust pidada iseloomulikuks eeskätt meediauudistele kui tekstiliigile.

Verbaalset tegevust väljendav propositsioon on sageli transformeeritud nominalisatsiooniks. Postimehe uudiste pealausetes on nominaliseeritud vormis teate allikas 75 juhtumil, mis moodustab üle 12% kõigist nominalisatsioonidest. Kõige sagedamini on nominalisatsioon kujul *sõnul* (39 korda), järgneb *hinnangul* (19 korda). Veel kasutatakse väljendeid *andmetel, kinnitusel, ütlusel, ütlust mööda, ettepanekul, ettepaneku kohaselt, teatel, sõnutsi.* Eesti Päevalehes on teate allikat väljendav propositsioon nominaliseeritud harvem, kokku 20 juhtumit, neist 11 korda *sõnul*, järgnevad *hinnangul, teatel* ja

andmeil. Kõigist EPL-i nominalisatsioonidest moodustab see veidi üle 6%.

(67) *Välisminister Toomas Henrik Ilvese* **sõnul** ei näitaks Eesti end usaldusväärse partnerina, kui ta selle tehingu, milles NATO juhtriigi USA firmaga on kokku lepitud, nüüd katki jätaks. (EPL, 1)

(68) *Mõõdukate peasekretäri Tõnu Kõivu* **hinnangul** on samas Kreitzbergi konkurentsist eemaldamine Rüütlist märksa hõlpsam, sest ainsana suurtest erakondadest pole Keskerakond oma kandidaati kinnitanud mitte erakonna kongressil, vaid volikogus. (PM, 3)

(69) *Kasela* **ütlusel** mõjutab konkurents eelkõige neid reisijaid, kelle jaoks puhkuse sihtkoht pole määrav. (PM, 13)

Nominalisatsioonid, nagu muudki sisestatud tarindid, on uudistekstis suhteliselt pikad. Kui analüüsitud ilukirjandusteksti vähesed nominalisatsioonivormid olid enamasti üksiksõnad, erandjuhtudel ühe genitiivtäiendiga, siis uudistekstide propositsioonide nominaliseerimist ei takista ka üsna suur hulk argumente ja argumentide laiendeid.

(70) Mitmed opositsiooni ja kolmikliidu poliitikud peavad Peeter *Tulviste eduka presidendiks* **kandideerimise** hinnaks peaminister *Mart Laari kohalt* **tagandamist**. (EPL, 1)

(71) *Egiptusesse otsereiside* **alustamiseks** taotles luba ka Itaalia ärimehe Ernesto Preatoniga seotud Domina World Travel, kelle pakutavasse lennukisse mahub üle 200 inimese. (PM, 13)

(72) NATO lükkas *põhiväe Makedooniasse saatmise* **otsustamise** edasi tänasele, samal ajal kui **võitluses** *Makedoonia valitsuse ja albaanlastest mässuliste vahel* hävis õigeusu klooster. (PM, 10)

6. Predikaadita tarindid

Võrdluseks analüüsitud ilukirjandusteksti tüüpilisim predikaadita noomenifraas on genitiivtäiend. Veidi vähem oli adjektiivtäiendeid, veel vähem obliikvakäänetes ja postpositsioonilisi substantiivtäiendeid. Seda kõike esineb ka meediatekstis, aga uudistekstide kõige esindatum predikaadita fraasi tüüp on siiski lisandifraas. Eespool oli nimetatud, et uudisteksti prototüüpne pealause on ütlemislause, kus verbaalse tegevuse agent, st ütleja on individualiseeritud nimeliselt. Meediatekstile on iseloomulik, et üksikisikute nimetamine ja kategoriseerimine on põimunud, st osaline esitatakse nimepidi, aga lisatakse ka, kes ta on. Nagu eespool partitsiiptarinditega seoses viidatud, võidakse osalisi iseloomustada kahel viisil: kas nende tegevuse (ameti) või isiklike omaduste (rahvus, elukoht, vanus, sugu) alusel

(Leeuwen 1996). Uudistekstis on isik esimest korda nimetades alati karakteriseeritud. Nimeliselt on individualiseeritud ametiisikud, ja neid karakteriseeritakse funktsiooni (ametikoht, staatus, tiitel) järgi (vt nt laused 14–17). Nagu eespool öeldud, ei nimetata õnnetustest ja kuritegudest rääkivates uudistes osalisi nimeliselt, vaid neid inimesi nimetatakse ja ka iseloomustatakse üksnes isiklike omaduste, suhete või identiteedi järgi (vt laused 9 ja 10).

Lisandifraas võib olla erineva pikkusega. Presidenti, valitsusliikmeid ja veel mõnesid kõrgeid riigiametnikke iseloomustatakse ainult tiitliga: *president Lennart Meri, peaminister Mart Laar, rahvastikuminister Katrin Saks, õiguskantsler Allan Jõks.* Teiste riikide vastavate ametiisikute puhul lisatakse riigi nimi: *Venemaa president Vladimir Putin, Inglise kuninganna Elisabeth, USA suursaadik Melissa Wells.* Sama puudutab ka firmade nimetamist: Eesti firmade nime ees pole tavaliselt lisandeid ega täiendeid, kuna kohanimi sisaldub juba nimes (*Eesti Energia, AS Narva Elektrijaamad*), välismaa firmad esinevad eesti meediatekstis aga alati koos asukohamaale viitava täiendiga (*USA firma NRG Energy, USA Minnesota osariigi energiafirmale Northern States Power kuuluv NRG Energy*).

Kui tiitel või ametinimetus ei ole piisavalt eristav, vaid piirdub rolliga mingis institutsioonis, asutuses või organisatsioonis, siis lisatakse ka institutsiooni nimi (*Rahvaliidu esimees Villu Reiljan, Keskerakonna parlamendipoliitik Siiri Oviir, valitsuse pressiesindaja Priit Põiklik*). Täiendeid lisandub, kui institutsioon pole üleriigiline, vaid kohalik, või on tegemist institutsiooni osaga. Sel juhul lisandub institutsiooni nime ette kohanimi või järele allasutuse nimi (*Isamaaliidu fraktsiooni liige Kalle Jürgenson, Reformierakonna fraktsiooni esimees Jürgen Ligi, Saksa välisministeeriumi Euroopa osakonna juhataja Reinhard Schweppe, Pärnu politseiteabetalituse juhtivinspektor Kaja Kukk, Tallinna koerte varjupaiga juhatuse esinaine Liina Kolmann*). Mõnikord karakteriseeritakse isikut ka mitme eri funktsiooni järgi (*isamaaliitlasest riigikogu liige Peeter Olesk, Kapo pressiesindaja komissar Olari Valtin, Keskerakonna abiesimees ja presidendikandidaat Peeter Kreitzberg*). Isiku tiitlite ja ametite loetelu võib mõnikord osutuda üsnagi pikaks, eriti kui lisandub muidki täiendeid.

(73) Valitsus kutsus kabinetinõupidamisele ka AS-i Eesti Energia juhatuse esimehe **Gunnar Oki** koos Narva Elektrijaama AS-i vähemusosaluse müügi juriidilise nõustaja vandeadvokaat **Sven Papiga** advokaadibüroost Raidla ja Partnerid. (EPL, 6)

(74) *Concordia Ülikooli meediateaduskonna dekaan ja endine Eesti Televisiooni peadirektor* **Hagi Šein** peab valitsuse eilset otsust kaotada ETVst reklaam praegustes oludes parimaks lahenduseks. (PM, 5)

7. Kokkuvõtteks

Meediauudistele on iseloomulikud pikad ja komplekssed laused. Sel on kaks põhjust. Lausete aluseks olevad propositsioonid on ise pikad, sest iga predikaadiga seostub palju argumente. Teiseks ühendatakse rinnastus-, sisestus- ja alistustransformatsioonidega üheks lauseks kokku suhteliselt palju propositsioone. Propositsioonide arv ühes uudisteksti lauses on üle seitsme. Lauselisuse (pea- ja kõrvallaused kokku) on pindstruktuuris säilitanud umbes kolmandik propositsioone. Pealausena on nii Postimehes kui Eesti Päevalehes esindatud alla 17% propositsioonidest, kõrvallausete hulk on Eesti Päevalehes mõnevõrra suurem. Sisestatud konstruktsioonidest on Postimehes tunduvalt rohkem nominalisatsioone, Eesti Päevalehes aga rohkem predikaadita tarindeid.

Valdav osa kummagi ajalehe uudistekstidest esitatakse vahendatud kujul, kellegi sõnu tsiteerides või refereerides. Pealaused on lühikesed ja märkimisväärselt suur osa neist annab edasi verbaalset tegevust (*NN ütles, et...*). Ka nominalisatsioonide hulgast eristub omaette rühmana teate allikale viitav fraas (*NN-i sõnul*). Teate allikas on Postimehes enamasti nimeliselt individualiseeritud, Eesti Päevalehes sageli ka kollektiivne. Esmakordsel nimetamisel on isik funktsiooni alusel karakteriseeritud, mistõttu suure osa predikaadita propositsioonidest moodustavad lühemad või pikemad lisandifraasid. Ka partitsiiptarindid on sageli isiku karakteriseerimise vahendiks, mistõttu suur osa neist kuulub täiendina noomenifraasi koosseisu. Sisuline info sisaldub kõrvallauseis ja sisestatud tarindites. Kõrvallaused on pikad ja mitmeastmelised, süntaktiliselt funktsioonilt enamasti objektlaused, mis on vormistatud tsitaatide või refereeringutena.

Kirjandus

Fowler, Roger 1991. Language in the News. Discourse and Ideology in the Press. London and New York: Routledge.

Hadenius, Stig 1992. Sweden in an iron grip. – Journal of Communication 42:3.

Hallin, Dominick 1992. The passing of the "High Modernism" of american journalism. – Journal of Communication 42:3.

Heikkinen, Vesa 1999. Ideologinen merkitys kriittisen tekstintutkimuksen teoriassa ja käytännössa. Helsinki: SKS.

Ingo, Rune 1986. Kontrastiivisia havaintoja lausemaisuuden asteiden käytöstä. – Erikoiskielet ja käännösteoria. Vaasa: Vaasan korkeakoulu. 7–16.

Ingo, Rune 1994. Täysien lauseiden ja upotusten käyttö eri kielissä. – Emakeel ja teised keeled. Toim. R. Pool, J. Valge. Tartu: Tartu ülikool. 13–24.

Katz, Elihu 1992. The end of the journalism? Notes on watching the war. – Journal of Communication 42:3.

Kivilo, Evelin 2001. Isikuid laiendavad partitsiiptäiendid ajalehtedes. Bakalaureusetöö TÜ eesti keele õppetoolis.

Leeuwen, T van 1996. The representation of social actors. – Text and Practices. Eds. C.R. Caldas-Coulthard, M. Coulthard. London: Routledge. 32–70.

Peegel, Juhan 1970. Ajalehežanrid. – Ajaleht. Tartu: Tartu Riiklik Ülikool.

Remmelg, Raili 1999. Lausestruktuuri keerukus eri tekstiliikides. Seminaritöö TÜ eesti keele õppetoolis.

Uuspõld, Ellen 2001. *des-* ja *mata*-vormide kaassõnastumine ja eesti komareeglid. – Keele kannul. Pühendusteos Mati Erelti 60. sünnipäevaks. Koost. ja toim. R.Kasik. Tartu: Tartu ülikool. 306–321.

Walton, Dominick 1992. Journalists: The tarpeian rock is close to the capitol. – Journal of Communication Vol 42:3.

Semanttinen tyhjiö

Mauno Koski
Åbo Akademi

1. Yhtenä leksikaalisen merkityksenmuutoksen syynä on se, että kieliyhteisössä tunnetaan kyllä sana morfologisena yksikkönä ja tiedetään sen käsitekenttä ja joukko kollokaatioehtoja, mutta sen semantiikka ei ole yhtä selvä. Kaikki sanankäytön denotatiivinen tai referentiaalinen hajonta ei johdu merkityksen epäselvyydestä, ei esim. metafora, metonymia, kielellinen konservatismi (esim. saksan *Feder* 'sulka' > 'mustekynä') tai referenttiluokkien eron irrelevantiksi tulkitsemisesta johtuva uudelleenluokittelu (esim. sukupuolieron neutralisaatio eräissä sukulaisnimityksissä) eikä säännönmukainen moniviitteisyys (Регульярная многозначность, Apresjan 1971) niin kuin *Hän istuttaa tomaatin* ja *Hän syö tomaatin* kasvista ja kasvin funktionaalisesti relevantista osasta tai *Karsinassa on hyvin syötetty porsas* ja *Taulussa on muutamin vedoin piirretty porsas* entiteetistä ja sen kuvasta (ks. Koski 1991). Tarkoitan sanankäytön denotatiivisella hajonnalla varsinaista leksikaalista monimerkityksisyyttä, polysemiaa (pohtimatta nyt polysemian ja leksikaalisen homonymian rajankäyntiä), ja referentiaalisella hajonnalla tarkoitan sitä, että sanaa käytetään saman merkityksen puitteissa erilaista tarkoitelajeista. Denotatiivinen hajonta perustuu synkronisesti (missä tahansa synkronisessa vaiheessa) eri leksikaalisten merkitysten sulkeiseen joukkoon, ja referentiaalinen hajonta perustuu synkronisesti avoimeen joukkoon. Denotatiivisesta hajonnasta ei tyypillisesti voi osoittaa semanttista invarianssia ("semantic invariant"), joka taas on referentiaalisessa variaatiossa olennaista. Anna Wierzbicka (1996: 242) tavoittelee tämäntapaista asetelmaa englannin esimerkeillään *spring*, jolla on ainakin neljä eri merkitystä ('hyppy', 'vieteri', 'lähde', 'kevät'), ja *love*, jolla on vain yksi merkitys, sillä (1) romanttisella, (2) parentaalisella ja (3) veljellisellä rakkaudella ei ole länsimaiseen maailmankuvaan perustuvaan semanttiseen rakenteeseen vaikuttavaa eroa. Wierzbicka ei kuitenkaan osoita esimerkkiensä semanttista tyyppieroa. Nykykielen *spring*-lekseemin semantiikasta ei voi osoittaa käyttöön vaikuttavaa invarianssia, vaikka diakronisessa tarkastelussa löytyisikin jonkinlainen lähtöön

ponnahtamisen merkityspiirre (germaanisen juuren merkitys 'sich hastig bewegen'). Lekseemi *love* on ongelmattomampi kuin esim. *mouth* tai suomen *suu*. D. A. Cruse (1986: 71–74) on esittänyt englannin *mouth*-sanan avulla miellespektrin (sense-spectrum) käsitteen, sen miten sanan erilaisista konventionaalisista käytöistä muodostuu semanttinen jatkumo ilman selviä merkitysrajoja, vaikka jatkumossa etäämpänä toisistaan olevat käytöt voisivatkin edustaa eri merkitystä. Sana *mouth* käy saman merkityksen puitteissa ihmisen ja kalan suusta (John keeps opening and shutting his mouth like a fish) mutta ilmeisesti eri merkityksen puitteissa ihmisen ja joen suusta (?The poisoned chocolate entered the Contessa's mouth at the same instant that the yacht entered that of the river). Cruse osoittaa laboratoriotekoisilta tuntuvilla esimerkeillään, että saman lekseemin konventionaalisista tarkoiteluokista toiset ovat lähempänä ja toiset kauempana toisistaan, mikä on kattavammin sanottu kuin se, että toisten lekseemien merkitykset kaikkiaan ovat lähempänä ja toisten kauempana toisistaan, niin kuin esim. Wierzbicka esittää (tässä ei nyt ajatella metaforista tai metonyymista käyttöä, mitkä ovatkin eri asioita).

Voitaneen olla yhtä mieltä siitä, että englannin *spring* on polyseeminen ('hyppy', 'vieteri', 'lähde', 'kevät', jos ei ajatella erikseen homonyymisena referenssiä *spring* 'kevät', mikä ei ole tarpeellista). Ehkä englanninkielisten mielestä 'hyppy' olisi jotenkin muita merkityksiä perusluonteisempi, koska se korreloi johtamattomaan verbiinkin, mutta sellaista semanttista invarianssia tuskin voisi ajatella, että vieteri olisi eräänlainen hyppy, lähde eräänlainen hyppy (lähde on ensimmäisen luokan entiteetti, jollainen hyppy ei ole) tai kevät eräänlainen hyppy. Tällainen invarianssiajattelu ei sovi lekseemin käytön denotatiiviseen hajontaan, mutta se sopii yhden merkityksen puitteissa referentiaaliseen hajontaan. Suomen *suu* tarkoittaa paitsi ihmisen tai eläimen suuta (millä ei ole mitään semanttista eroa; referenttien erilaisuudesta huolimatta kyseessä ei ole edes semanttinen varianssisuhde) myös erilaisia muita väylän tai pitkulaisen säiliön tai maastonkohdan päässä olevia aukkoja (tässä VÄYLÄ ja SÄILIÖ mielikuvaskeemoina). Nykysuomen sanakirjassa on konneksiot *kohdun suu*, *astian suu*, *pullon suu*, *hylsyn suu*, *tykin suu*, *taskun suu*, *hihan suu*, *saappaanvarren suu*, *sukan suu*, *katiskan suu*, *rysän suu*, *viemärin suu*, *säkin suu*, *solan suu*, *laakson suu*, *luolan suu*, *sataman suu*, *salmen suu*, *käytävän suu*, *kujan suu*, *oven*

suu ja yhdyssana *joensuu*, Suomen kielen perussanakirjassa on lisäksi *tunnelin suu*. Jouduttaisiin loputtomalle tielle, jos ajateltaisiin, että kaikissa ilmauksissa olisi erimerkityksinen *suu*. Samoin ei ole syytä ajatella, että sana *pää* edustaisi eri merkityksiä konneksioissa *sormen pää*, *nenän pää*, *köyden pää* ja *tien pää* vaan kyseessä on pitkänomaisen entiteetin kumpaa tahansa loppupäätä tarkoittava *pää* 'Ende, Spitze', mutta eri merkitys on kyllä ilmauksissa *kaalinpää* 'Kohlkopf' ja *nuppineulan pää* 'Stecknadelkopf', joissa ei ole kyse pitkänomaisen entiteetin (PÖTKÖ) jommastakummasta päästä sinänsä vaan pyöreästä osaentiteetistä (PALLO), joka kyllä on kokonaisentiteetin päässä, kärjessä.

Kielentutkijat ovat taipuvaisia ajattelemaan niin, että sellaiset sanat kuin suomen *suu* ja *pää* ovat alkuaan ruumiinosan nimityksiä ja että niiden muu käyttö on metaforaa (ihmiskeskeinen maailmankuva). Eri kielissä onkin ruumiinosametaforia, mutta kaikki ruumiin ja ruumiinosien nimitykset eivät ole tässä funktiossa primaareja. Diakronisellakin ulottuvuudella metaforasuhdetta olennaisempi on monissa tapauksissa entiteetin muotoon (hahmoon) tai spatiaalisiin suhteisiin perustuva nimeäminen (ks. esim. Koski 1997 ja 2000). Ei ole syytä ajatella denotatiivisen hajonnan periaatteen mukaisesti, että sanalla *suu* on eri merkitys aseen piipun yhtä osaa ja säkin yhtä osaa varten, mutta ei myöskään ole ajateltava metaforakaavan mukaisesti, että pyssyllä ja säkillä on ikään kuin suu tai köydellä ikään kuin pää. Paremmin pullon, aseen piipun, säkin ja joen suut ovat eräänlaisia suita, jotka on nimetty referentiaalisen hajonnan mukaisesti. Toinen asia on, että lekseemin kaikki vakiintuneet käytöt eivät ole välttämättä saman jakoperusteen mukaisia. Jokin käyttö voi liittyä muita prototyyppisempään tarkoitteeseen, ja jokin käyttö voi lekseemin semanttisen rakenteen kannalta perustua erilliseen merkitykseen (semeemiin). Tällainen olisi suomen *pää* ruumiinosan nimityksenä, sillä sitä ei voi pitää metaforasuhteen kuva-jäsenenä sen paremmin 'nuppi'-merkityksiseen kuin 'loppupää'-merkityksiseenkään käyttöön. Toisaalta eivät saman lekseemin merkitykset tai merkitykset ylipäätäänkään ole läheskään aina yksioikoisesti vain kahdenvälisissä suhteissa (niin että *pää* ruumiinosaan viittaavana olisi suhteessa *pää*-sanan käyttöön vain nuppimaisesta osaentiteetistä tai vain spatiaalisesta kärjestä, loppupäästä), vaan referenssisuhteet muodostavat monisuuntaisen verkoston.

Lekseemin semanttiset hajontatyypit ovat siis (1) eri merkityksistä koostuva denotatiivinen hajonta ja (2) saman merkityksen puitteissa vain referentiaalinen hajonta. Denotatiivinenkin hajonta sisältää viime kädessä referentiaalisen hajonnan. Lekseemin semanttinen rakenne voi koostua jommastakummasta hajontatyypistä tai molemmista hajontatyypeistä. Denotatiivinen hajonta voi perustua metaforaan, metonymiaan tai uudelleenluokitteluun tai siihen, että on syntynyt semanttinen tyhjiö. Referentiaalinen hajonta perustuu sinänsä samaksi luokitteluun (x:n y on eräänlainen y), mutta siihen kuuluu myös "säännönmukainen moniviitteisyys" ja paikoin uudelleenluokittelukin. Semanttisen tyhjiön käsiteteellä ei ole tavallisesti operoitu. Esimerkkeinä semanttisen tyhjiön aiheuttamista muutoksista ovat (1) *kolkka* puintitermeinä, kalastustermeinä, metsästysterminä ja lasten piilosleikkiin liittyvänä terminä, (2) *hiisi* olentoa merkitsevänä ja (3) maannimien *Livland* ja *Marienland* appellatiivinen käyttö.

2.1. Itämerensuomalaisissa kielissä substantiivi *kolkka* variantteineen merkitsee paitsi nurkkaa tai kulmaa myös erilaisia puisia tarvekaluja. Variantit ovat suomen *kolkka ja* harvemmin viron *kolk* : *kolga*, derivoituna suomen *kolkko* [~ *kolko* < *kolkkoi*] tai *kolkku*, yleisesti viron *kol'k* : *kol'gi*, harvoin *kolk* : *kolgu*. Yleistävinä, kattavina selityksisinä ovat suomen sanalle *kolkka* "pertica varii usus" (Renvall 1826) ja viron sanalle *kol'k* : *kol'gi* "puutükk, klomp", "lühike ja jäme tarbepuu" (Saareste 1962 sub puutükk ja tarbepuu), "nööri(de) külge seotud lühike jäme puutükk (enamasti töö v. veoriista osa)", esimerkkeinä *koodi kolk* 'varstan iskuri' ja valjastuksessa aisojen etupäähän kiinnitettävää puukappaletta tarkoittava *kolk* (Eesti kirjakeele seletussõnaraamat); valjastusterminä *kolk* tarkoittaa myös vetokarttua (Rapla Märtna, Saareste 1958 sub ader, 1963 sub äke). Samoillakin murrealueilla sana esiintyy erilaisten entiteettien nimityksenä, mikä osoittaa, että se merkitsee yleisesti määräehdot täyttävää esinettä (niin kuin *kapula* suomessa tai *pulk* virossa).

Virossa *kol'k*-sanan tarkoitteet ovat aina jotenkin narulla tms. tai molemmista päistä naruilla kiinni jossakin: 'aisan etuosassa oleva puu', 'vetokarttu', länsiviron *kol'k* 'puinen [narun päässä riippuva] säppi' (samoin suomen eteläsatakuntalaismurteiden *kolkka* 'säppi'), 'varstan iskuri', pohjoisviron ja liivin *kol'k* 'eläimen jalkoihin juoksemisen estämiseksi sidottava puukapula' (EKI-M; Vestring 1700-luvun alussa "Der Klotz daran sas Vieh gebunden wird";

Kettunen 1938), viron saarten murteissa ja länsimurteissa *kol'k* 'kytkyen puinen levymäinen osa' ja Väike-Maarja *kolkmed* 'puukytkyt' (Saareste 1959: 460), pohjoisviron pääasiassa läntisten murteiden *niie kolgad ~ kol'gid* 'kangaspuun pyöräset' (EKI-M; Saareste 1958: 964–965), Helme 'nelikulmainen puinen karjankello' (Saareste 1958: 1156), Martna Võnnu '[naruilla poikkipuuhun sidottu puinen] kolkutin' (Saareste 1958: 1157), Saarenmaa '[seinään sidottu] puinen suolasalkkari' (*kol'k* : *kol'gi* suullisesti Ellen Niit, *kolk* : *kolgu* Saareste 1959: 1220). Sanan hyperonyymiseen merkitykseen 'naru(i)lla kiinni oleva puukappale' perustuvat hyponyymiset käytöt edustavat paremmin pragmaattista soveltamista kuin semanttista polysemiaa: varstan osana on eräänlainen "kolk" ja kytkyen osana eräänlainen "kolk", vaikka eri käytöt ovat konventiaalistuneet leksikaalistumiseen rinnastettavalla tavalla. Joskus tällaiseen pragmaattiseen termiytymiseen liittyy metonyyminen siirtymä, esim. *kolgid* [= *pandid* länsimurteissa] nuottaa vedettäessä käytettävät leveät nahkavyöt, jotka on yhdistetty rinnan kohdalta *kol'k* -puupalalla (Saareste 1959: 1141). Verbi *kolkima* 'loukuttaa pellavia' esiintyy koko eteläviron murrealueella, koillisrannikon murrealueella, Viljandimaalla ja Etelä-Pärnumaalla, ja siihen korreloivat pellavaloukun nimitykset *kolgispuu* ja *kolgits* (Saareste 1959: 602); sanaa *kolgispuu* on käytetty myös pesukartusta ja sanaa *kolgits* varstan iskurista (Saareste 1962: 126, 557).

Eräät lasten piilosillaololeikit ovat nimitykseltään *olla kolkkapiilosilla ~ kolkkasilla*, mikä perustuu eräänlaisen kepin keskeisyyteen leikissä (samoin kuin *karttu* ja *leikkiä karttua*, *olla karttusia*, *karttupiiloa* jne.). Sana *kolkka* esiintyy kuitenkin melko vähäisesti leikkivälineen nimityksenä: Alavus "Jos etsijä löysi jonkun, hän nopeasti juoksi kolkan luo ja kolautti sillä seinään", Soini *kolkko* ja Ilmajoki *kolkku* [*kolokku*] '[piilosleikkiin kuuluva] puukapula', hiukan toisenlaisessa piilosleikissä Vaala "yksi on kolkalla varustettu – kolkkaaja tulee ja heittää lattiaa pitkin kolkkansa (SMS). Useammin *kolkka*-sanalla on leikin terminologiassa muunlainen funktio, se tarkoittaa etsijää, mutta se voi samassa tekstissä tarkoittaa leikin keskuspaikkaakin, esim. Ikaalinen "Kolkka lähtee piiloittautuneita hakemaan. Keksittyään jonkun hän palaa kolkalleen ja sylkäisee sinne lausuen: minun kolkkani", Raahe "Yksi kolkalla olijoista heitti kepin niin kauas kuin jaksoi, kolkan piti hakea se ja lyödä sillä seinään t. puuhun – se joka ei ehtinyt piiloon,

oli kolkka" (SMS). Leikin nimi ja siihen kuuluva huudahdus on *kolkka*: Kolari *olhaan kolkala – yksi navetan perältä huusi: kolkka, kolkka! ja napautti seinään kepilä* (SMS). piilosleikeissä *kolkka* esiintyy hämäläis- ja pohjalaismurteissa, pohjoisissa savolaismurteissa ja Kainuun murteissa (SMS).

Nykysuomen sanakirjan *kolkka* 'nuijamainen iskuase, jolla pyydetään kaloja (mateita) läpinäkyvään jäähän lyömällä' esiintyy kansankielessäkin, mutta toisesta merkityksestä 'tylppäpäinen nuoli, vasama' ei ole tietoja SMS:n eikä Vanhan kirjasuomen sanakirjatyön kokoelmissa. Merkityksen 'vasama' on pannut Renvall sanakirjaansa (1826) ilmeisesti päättelyn tuloksena. Hän otti aikaisemmasta sanakirjaperinteestä yhdyssanan *kolkkapoika* (Juslenius 1745 "Colcapoika – puer sagittas reportans. then som igenhemtar pilar", samoin Ganander 1786), ja Jusleniuksen sanakirjan välilehdellä on Porthanin Sotkamosta saaduksi merkitsemä *kolkkamies* "Kolkamies en gosse som på ekorns-skytte följer med bogskytten och samlar bilarne", samoin Kajaanin läänin kuvauksessa: "Kolckamies, hwars syssla är, att noga gifwa ackt hwarest kolfwen eller pilen nedfaller" (Castrén 1754: 56), samoin Paltamo (SMS kaksi eriaikaista tietoa). Näiden lähteiden mukaan kolkkapojan ja -miehen tehtävä oravametsällä oli vasamien keruu, mutta missään ei sanota, että *kolkka* tarkoittaisi vasamaa. Toisena näiden oravanmetsästäjän apulaisten tehtävänä oli kolkalla puuhun lyömällä hätistää orava liikkeelle: Haapavesi "Oravan ampujalla oli ennen *kolokkamies*, aseenaan kolkka, jolla iski puuhun, jolloin orava säikähti ja tuli huomatuksi" (SMS); tässä tehtävässä olevan apulaisen nimitys *kolkkamies* on tunnettu savolaismurteissa ja Kainuun murteissa, samoin harvinaisempi *kolkkapoika* sporadisesti eri puolilla (Hollola Sotkamo Ylitornio SMS) ja samasta tehtävästä on kyse tiedossa *kolokka on oravametällä joka kopistaa puuta* (Puolanka SMS). Mitään rakenteellista semanttista estettä ei olisi ollut sille, että *kolkka*-sanaa olisi pohjoisissa murteissa voitu käyttää vasamasta yhtä hyvin kuin nuijastakin, vrt. NS *nuijapäät vasamat* ja *nuijapää* 'sammakontoukka', SMS Ylitornio *kolkkapää* 'sammakontoukka'. Haapavedeltä saadun tiedon lisäksi ei kolkkamiehen tai -pojan kolauttamisvälinettä sanota kolkaksi (SMS), mikä osoittaa, ettei tällainen korrelaatio enää ollut kovinkaan yleinen. Näin oli muodostunut semanttisen tyhjiön mahdollisuus.

Suomen partiopoikaliikkeessä vaihdettiin nuorimpien partiopoikien kansainvälinen nimitys *sudenpentu* (wolf cub) 1940-luvun

alussa kansallisromanttiseksi nimitykseksi *kolkkapoika* (josta sittemmin vähitellen luovuttiin ja v. 1982 käytettiin taas yksinomaan sudenpentu-nimitystä). Partiolaisillekaan ei aina ollut selvää, mitä *kolkka* tai *kolkkapoika* vanhastaan oli tarkoittanut. Nimityksen *kolkka*-osaan muodostui semanttinen tyhjiö, mutta siitä tuli institutionaalinen merkki ilman kytkentää aikaisempaan intensioon, ja näin tehtiin myös termit *kolkkaparvi* 'kolkkapoikien ryhmä' ja *kolkkajohtaja*; mielteiden suhde oli jotenkin sellainen kuin 'kolkka-instituutioon kuuluva määräjoukko' ja 'kolkka-instituutioon kuuluva johtaja'. Sporadisesti *kolkkapoika*-sanaa on käytetty häissä avustavasta pojasta (Mäntyharju SMS), tässäkin on täytetty semattista tyhjiötä.

Joissakin sanan käyttömalleissa *kolkka* on integroitunut lyömistä tai lyömisen ääntä merkitseviin verbeihin, erityisesti suomen *kolkata* ja viron *kolkida*. Lyömiseen liittyy verbin käyttö loukuttamisesta ja puimisesta ja substantiivin käyttö loukusta ja varstasta tai sen osasta, myös viron *kolki saama* ja *kolki andma* 'saada/antaa selkään'. NS:n *kolkka* (~ *kolkkapuu*) 'nuijamainen iskuase, jolla pyydetään kaloja (mateita) läpinäkyvään jäähän lyömällä' on merkitty kansankielestä vain Ylitorniolta ja Oulun läänin Pyhäjärveltä, Vesannolta on *kolkkakurikka* (yleensä iskuaseena mainitaan nuija tai kirveenhamara), mutta kyseisen kalastustavan nimitys *käydä kolkalla ~ olla kolkalla*, verbi *kolkata* tässä yhteydessä sekä *kolkkakala* tunnetaan paitsi pohjoisten murteiden alueella myös savolais-karjalaisella alueella Repolassa asti (SMS; Nirvi 1949: 153; KKS). Verbi *kolkata* 'lyödä', 'kolauttaa' esiintyy suomen pohjalaismurteissa ja itämurteissa ja karjalassa, kun taas *kolkka* pitkänomaisen puisen tarvekalun nimityksenä kuuluu suomen länsimurteisiin. Pohjoisissa lounaismurteissa (Raumalta Taivassaloon ja Nousiaisiin) on *kolkko* 'puunuija' (kirvesmiehen nuija, puusepän nuija, nuija jolla hakataan merilevää verkoista, nuija jolla pienitään kokkareet pellolla, perunanuija, "toppasokeri paloiteltiin iskemällä kolkalla [nuijalla] puukon hamaran päälle") ja *kolkka* 'hiirenpyydys, jossa hiiri jää painavan osan alle', ja eteläpohjalaisissa murteissa on *kolkku* 'nuija, survin' (SMS). Eteläpohjalaisissa ja läntisissä keskisuomalaisissa murteissa on *kolkka* 'kirnun mäntä, männän alaosa' (SMS; Ganander 1786; Renvall 1826), pohjoisissa hämäläis- ja satakuntalaismurteissa on *kolkka* 'myllyn karistin [keppi, joka säätää jyvien valumista myllynkiven silmään]' ja satakuntalaismurteissa 'oven säppi' (SMS).

Kalastustermit *käydä~olla kolkalla*, *kolkata* ja *kolkkakala* ovat paremmin verbilähtöisiä itämurteisuuksia kuin substantiiviin *kolkka* 'nuija' perustuvia länsimurteisuuksia, ja kalastuksessa käytettävän puunuijan nimitys *kolkka* tunnetaan vain peräpohjalaismurteissa, ilmeisesti *kolkata*-verbi tulkittiin instrumentatiiviseksi, 'iskeä kolkalla', mikä puolestaan edellyttää läntisen esinetarkoitteisen *kolkka*-sana tuntemista. Lampaita on kuohittu menetelmällä, jossa olennaisena osana on kivesten rikkominen nuijaniskulla, ja tästä käytetään Peräpohjolassa verbiä *kolkata*, ja apuvälineenä käytettyjen puristuspihtien nimitys on Pellossa *kolkinpuut* ja Rovaniemellä *kolkkapuu* (SMS); sama käyttö on merkitty myös viron *kol'kima*-verbistä, "castriren (durch Schlagen mit einem hölzernen Hammer)" (Wiedemann).

Suomen eteläisissä länsimurteissa on merkkejä siitä, että *kolkka* on puintiterminä tarkoittanut varstaa tai varstan nuijaosaa niin kuin virossakin *kol'k*. Vakiintuneissa sanonnoissa Suomusjärvi [puintiaikana] *kolkkakim paukku taas* ja Somero *Sitte vasta nauris kasvaa kun kolkan äänen kuulee* (SMS) *kolkka* ymmärettäneen paremmin metonyymiksi 'puinti' kuin välineen nimitykseksi, mikä se alkuaan on ollut. Välineen nimityksenä *kolkka* on menettänyt käyttönsä, mutta sana on säilynyt samassa käsitekentässä niin että semanttinen tyhjiö on täytetty merkityksellä 'se [ihminen] joka viimeisellä puintikerralla lyö viimeiseksi' konneksiossa *tulla kolkaksi*, ja samasta asiaintilasta sanotaan myös *saada kolkka* (lounaismurteet, alasatakuntalaismurteet, eteläpohjalaismurteet, keskiset hämäläismurteet, kaakkoishämäläiset murteet SMS), esim. Suomusjärvi *see sai sen kolkan nime*, Kalanti [joka viljalajille tuli oma kolkka] *rukkin kolk*, *ohran kolk*, *kaoran kolk*. Toinen tyhjiön täyttävä merkitys on 'viimeinen puintikerta' (etelä- ja kaakkoishämäläiset murteet SMS), esim. Vanaja *Meillä saatiin öylön kolkka*, Hollola *Kolkka on ku saatiir riihi puitua*. Verbiin *kolkata* on integroitunut ilmaus *riihenkolkka* 'kolaus, joka syntyy siitä, että viimeinen riihenparsipari lyödään tiukasti yhteen' (Artjärvi SMS).

Suomen *kolkka* esiintyy keskiajalta lähtien ruotsinkielisissä asiakirjoissa eräänlaisesta nuotasta puhuttaessa. Kokemäen Lammaistenkosken kalastuksesta sanotaan v. 1347 "met kolkwm, næthiom eller nothom" ja 1412 "ath the hafde fiskath met theris kolkom j Lammass viik", Tornionjoella *kolkka* on tarkoittanut apajapaikkaa, esim. 1482 "wp i kolkarne" , ja Oulujoelta mainitaan

v. 1558 kolkkaverkot, "kolkenäth" (Nirvi 1949). Kokemäenjoen kolkalla pyynnistä on kuvaus vuodelta 1751 (F. R. Brander, De regia piscatura Cumoënsi), olennaista tässä kuvauksessa on maininta siitä, että toinen siula jäi puoliksi uppeissa olevan puun varaan [a ligno], sillä juuri tämä puupölkky on *kolkka*, niin kuin Kustaa Vilkuna (1951) on esittänyt. Vilkuna mainitsee tällaisesta uittoverkkoa kuljettavasta puupölkystä käytetyn eri kielissä samaa nimitystä kuin nuijasta, esim. yläsaksan *keule*, englannin *hammer* ja samaa nimitystä kuin puupölkystä niin kuin Tenon saamen *čoska* ja Inarin saamen *čuuskaž*. Myös aunuksessa on Säämäjärvi *kol'kkɨ* [*kolkka*+] ~ *kol'kkoi* 'n. 30–50 cm pitkä nelkulmainen puu, halko, verkon tai pitkänsiiman kohona (KKS aineskok.) ja viron länsimurteissa on Audru *kolk* : *kolga* 'kolmiomainen tai soikea lauta verkon kohona' (EKI-M; Saareste 1963: 634–635). Sanan *kolkka* käyttö tässä funktiossa on jäänyt semanttisesti epäselväksi. Yhdessä Tornionjoen kalastuksen kuvauksessa v. 1789 kolkkaa sanotaan tarpomanuotaksi, "så kallade Kålkar eller Puls-notar", ja joidenkin tulkintojen mukaan jokien "kolkka" perustuisi lohien tappamiseen kolkkaamalla (Nirvi 1949).

2.2. Ims. kielissä, erityisesti virossa ja suomessa, on ollut esihistoriallisen ajan lopulta sana *hiisi* 'kulttipaikka, uhrilehto'. Sana esiintyy jossakin määrin virolaisessa ja hyvin runsaasti suomalaiskarjalaisessa folkloressa, alkuaan mytologisen, aina jotenkin vainajauskoon liittyvän paikan nimenä. Kristinuskoon käännyttämisen jälkeen Viron hiis-paikat säilyttivät ainakin osittain funktionsa hyvin kauan, paikoin aina 1800-luvulle asti ja sana *hiis* säilyi kielessä vanhassa funktiossaan kulttipaikan nimityksenä. Suomessa kirkon vaino kohdistui selvemmin vanhaan uskontoon. Alkuaan hiisi-kulttipaikat olivat kylän sisäpiirissä, mutta ne oli vietävä takamaille, ja lopulta ne menettivät kulttipaikan funktionsa. Näin *hiisi*-sana menetti aktuaalisuutensa, ja ihmiset eivät enää tienneet sen merkitystä, mutta se säilyi paikannimien ja folkloren elementtinä. Sakraalisille hiisi-paikoille oli ominaista jonkinlainen kivisyys, esim. kivilatomus, ja ilmeisesti myös jotkut siirtokivilohkareet, hiidenkivet, olivat jotenkin sakraaleja. Tuloksena oli se, että oli useita mäkiä tai vuoria, joiden nimessä oli *hiisi*-elementti, esim. *Hiidenmäki*, *Hiidenvuori*. Vaikka *hiisi*-sanan varsinainen merkitys oli unohtunut, sen sakraalisluonteinen tai ainakin myyttisluonteinen konnotaatio säilyi. Kun suomalaisessa tarinaperinteessä oli germaanisia der

riesische Teufel-aiheita, *hiisi*-elementtiset nimet saivat uuden tulkinnan. Kun *Hiidenmäki*-nimisellä paikalla oli kivilatomus, siihen liitettiin tarina, jonka mukaan jättiläinen oli koonnut kivet heittokivikseen. Semanttisesti tyhjäksi jäänyt *hiisi* tulkittiin jättiläistä tarkoittavaksi. Samalla tavalla uskottiin siirtokivilohkareista, että ne olivat jättiläisen heittämiä, *hiidenkivi* tulkittiin hiiden heittämäksi kiveksi. Kun der riesische Teufel sisälsi jättiläismotiivin ja paholaismotiivin, olennoksi tulkittuna *hiisi* tarkoitti toisin paikoin jättiläistä ja toisin paikoin paholaista. Paikoin uskottiin myös, että Hiidenmäellä asui vuorenhaltija tms. ja näin tulkittiin *hiisi*-elementti olentotarkoitteiseksi, ja samalla tavalla tulkittiin uudestaan muutamat folkloreaiheet (Koski 1967–1970, 1990). Se, että nykykielessä *hiisi*-sanalla tuntuu olevan ensisijaisesti olennon merkitys, johtuu semanttisen tyhjiön täyttämisestä.

2.3. Ruotsin. ja saksankielinen nimi *Livland* tarkoitti alkuaan liiviläisten heimomaata Riianlahden itäpuolella, sitten saksalaisten valtaamaa liiviläisten ja lättiläisten aluetta, ja 1200-luvun loppupuolella nimi tarkoitti koko saksalaisten alistamaa aluetta Liettuan rajalta Suomenlahteen. 1500-luvulla tämä konfederaatio hajosi, mutta olojen vakiinnuttua 1620-luvulla nimi tarkoitti Liivinmaan kuvernementtia, joka ensin kuului Ruotsin valtakuntaan ja 1700-luvulta alkaen Venäjän valtakuntaan. Viroksi nimi oli aina 1800-luvulle asti *Liivlandi maa* tai *Liivlant*. Vaikka eteläinen osa virolaisalueesta kuului tähän kuvernementtiin, sen nimeä virolainen väki ei täysin omaksunut kirjakielen ulkopuolelle, ei edes tämän kuvernementin väki. Kansan keskuudessa nimi sinänsä tunnettiin, mutta ei tiedetty, mitä se tarkoitti. Se oli täytettävässä semanttisessa tyhjiössä, ja se appellatiivistui idiomeihin yleensä seutua tarkoittavasti: *lihvlant on kaugel* (Häädemeeste), *üits igävene lihvland* 'pitkä matka' (Karksi), *ku lähe sinna livlanti tagasi* (Võnnu). Joissakin konneksioissa sanaa käytetään merkityksessä 'tässä ympäristössä': *sii lihvlantis pole seda nähtud* (Märjamaa), *sii lihplantis põle seda olnd* (Vändra), samoin *lihtlantis* (Vigala), *liislantis* (Kullamaa Mihkli Pärnu-Jaagupi), *Sel ajal teis söukest selles lihvlantis polnd* ja *randas kasub maailmatu lihvland roogu* (Kaarma); yhteisöä tarkoittavasti *kes sest nüid aru saab, se on lih'landi asi* (Halliste) (Koski 2001: 350–351).

Liivinmaan piispa Albert hankki 1200-luvun alussa sotaretkelleen ristiretken statuksen.

Niin kuin Palestiina oli pyhitetty "Jumalan pojalle", Albert sai paavin pyhittämään v. 1215 voittomaansa "Jumalan äidille" Marian maaksi: *Terra matris*, *Terra Mariae*, *Terra Mariana*, saksaksi *Marienland*, sittemmin latviaksi "Mahras seme" ja viroksi *Maarjamaa*. Kun Marian kulttia ja Vanhaa Liivinmaata ei enää uudella ajalla ollut, *Marienland* jäi paikallisiin kieliin vain sakraalisen konnotaationsa varassa. Virolaiset käyttivät ilmausta jotenkin pyhästä, Jumalalle kuuluvaksi käsitetystä maaperästä, mutta osaksi myös maaperästä ylipäätään ilman sakraalia ajatusta. Oli siis siirrytty merkityksestä 'Land' merkitykseen 'Erde'. *Maarjamaa* on 'Geheiligter Boden, Gottes Erdboden, *kõik on Maarja maa* 'über all ist Gottes Erde', *Maarja maa sisse panema* 'beerdigen', ja ilmeisesti manalaisten apua pyydetään huudolla *Maarja maa rahvas, tulge appi!* (Wiedemann 1869). On sanottu, että *inimene sureb ära, aga Maarjamaa jääb ikka alles* (Häädemeeste Saareste 1958: 272), sateesta puhuttaessa *Maa'ramaa pääl um jumalal pant kõikkõ sagat'sit surmõ* (Lutsi), *Juu taeva toat kastis jälle oma moarja maad* (Muhu), *kastab seda kuiva maarjamaa pinda jälle* (Pöide). *Timahava Vanaezä es õlogi me puul, nii et timä Maa'ra maad är tõrbut* (O. Kallas, Lutsi maarahvas s. 74). Tausta-ajatus Jumalan maasta voi olla laulun sanoissa *tükk sinis taevas piä piäl ja moarja moa mull jalge all* (Kodavere, samoin Martna). Samaan tapaan kuin Liivinmaan nimestä on ilmaus *sedä olĕ ei maariä maa pääl keäki nännu* (Setu) (Koski 2001: 42–43).

Lähteet

Apresjan 1971. О регульярной многозначности. – Серия литературы и языка. Известия Академи Наук СССР 30: 6. 509–523.

Castrén, Erik 1754. Historisk och Oeconomisk Beskrifning Öfwer Cajanaborgs-län. Åbo.

Cruse, D. A. 1986. Lexical Semantics. Cambridge: Cambridge University Press.

EKI-M = Eesti Keele Instituut, Murdesektor. Tallinn.

Ganander, Christfrid 1786–1787. Nytt Finskt Lexicon I–III (Näköispainos käsikirjoituksesta, WSOY. Porvoo–Helsinki 1937–1940).

Juslenius, Daniel 1745. Suomalaisen Sana-Lugun Coetus. Stockholm (Välilehditetty näköispainos, SKS 1968).

KKS = Karjalan kielen sanakirja 1–5. Helsinki: Suomalais-ugrilainen Seura, 1968–1997.

Kettunen, Lauri 1938. Livisches Wörterbuch. Helsinki: Suomalais-ugrilainen Seura.

Koski, Mauno 1967–1970. Itämerensuomalaisten kielten hiisi-sanue I–II. Turku.

Koski, Mauno 1990. A Finnic holy word and its subsequent history. – Old Norse and Finnish Religions and Cultic Place-Names. Ed. by Tore Ahlbäck. Scripta Instituti Donneriani Aboensis. Åbo. 404–440.

Koski, Mauno 1991. Polysemiaa vai ei? – Erikoiskielet ja käännösteoria. VAKKI-seminaari XI. Vöyri 9.–10.2.1991. Vaasa: Vaasan yliopisto. Kielten laitos. 130–138.

Koski, Mauno 1997. Selkä. – Sananjalka 39, 17–29.

Koski, Mauno 2000. Ruumis raamatussa. – Pipliakielestä kirjakieleksi. (Kotimaisten kielten tutkimuskeskuksen julkaisuja 105.) Helsinki. 26–72.

Koski, Mauno 2001. Liivinmaan nimi. – Virittäjä, 530–560.

Nirvi, R. E. 1949. Lohijokiemme kolkka. – Kalevalaseuran vuosikirja 29, 141–153.

NS = Nykysuomen sanakirja (3. painos). WSOY Porvoo–Helsinki 1970.

Renvall, Gustavus 1826. Suomalainen Sana-Kirja. Aboae.

Saareste, Andrus 1958–1979. Eesti keele mõisteline sõnaraamat I–IV. Stockholm: Vaba Eesti.

SMS = Suomen murteiden sanakirjatyön kokoelmat. Helsinki: Kotimaisten kielten tutkimuskeskus.

Wiedemann, F. J. 1869. Ehstnisch-deutsches Wörterbuch. St. Petersburg.

Wierzbicka, Anna 1996. Semantics. Primers and Universals. Oxford University Press.

Vilkuna, Kustaa 1951. Mikä on lohijokiemme kolkka ja saarua? – Suomen museo 58, 12–20.

Kas elu on konteiner?

Arvo Krikmann
Eesti Kirjandusmuuseum

Kõik, mis ma tean, on see, et ma mitte midagi ei tea.
Sokrates, Platon, Seneca jt targad

Pidulik preambulatoorne pöördumine

Austet prof. akad. Õim, armas Haldur!

Eestlase prototüüpne juubelikõne eesti soost klassikule on umbes järgmise temaatilise kompositsiooniga:
1) juubilari roll Eesti asjas (kultuuris ~ kunstis ~ teaduses ~ ...);
2) minu isiklikud kokkupuuted juubilariga;
3a) juubilari roll minu klassikuks kujunemise loos ja/või
3b) minu positiivne roll juubilari elus;
4) minu roll Eesti asjas (kultuuris ~ kunstis ~ teaduses ~ ...).

Sellest põhiskeemist esineb, tõsi küll, ka põikeid. Nii näiteks ütles Juhan Peegel 1978. a veebruaris midagi umbes sellist: "Kui ma mõtlen neile asjadele, mida Rudolf Põldmäe on oma elus suutnud ära teha, siis läheb mu meel nii kadedaks, et tahaks lausa hambaid kiristada. Aga mul on ainult kaks hammast ja need ei asu kohakuti." Mina olen Sind kadestama hakanud nii varakult, et mul on olnud tehniliselt võimalik Juhani tuline soov korduvalt ka teoks teha.

Kui mõelda, mis asjad meid kahte seovad, siis võiksin seda lühidalt määratleda: **hingeasjad**. Mina kohtasin Sind esmakordselt vist 1970. aastate vahetus alguses Huno Rätsepa avangardse GGG-rühmituse seminaridel. Sa olid seal nii noor, nii habemega (just noored kandsidki tollal habemeid) ja nii hiilgav, et võttis lausa **hingetuks**. Sinu esimesest raamatust, mida tuli nimetada artikliks, sest Partei soovis tollal niimoodi (Õim 1971: 207–208), sain kinnituse sellele, mida mu **hing hämaralt aimas** juba enne – et keele semantilist struktuuri ei saa kujutleda mingi ühtlaselt liigendatud seoste võrguna, vaid keel on teatud suundades või regioonides struktureeritud märksa tugevamini kui teistes, ühtedel teemadel on võimalik hoopis peenemalt nüansseeritud kommunikatsioon kui teistel. Sinu teist raamatut (Õim 1974) lugedes sai mul hirmsasti **hing täis**: kust

põrgust on ta saanud sellise põhjaliku eruditsiooni ahvide keeleõppe, arheoloogiliste "eelhomode" jpt asjade kohta, millest mul polnud õrna aimugi! Sinu kolmanda raamatu (Õim 1983a) juures elasin **kogu hingest** kaasa vaese ELIZA tragöödiale, kes püüdis näida intelligentsem, kui ta tegelikult oli, ning pidi lõpetama häbi ja alandusega. Sinu doktoritöö eesti direktiivleksika kohta (Õim 1983b) jättis mulle vähemalt kaks suurt **okast hinge**:

1) miks küll on ilmaelu nii ülekohtune, et inimesed on leiutanud head vahendid lausepikkuste ja romaanipikkuste tekstijuppide sisustruktuuri kirjeldamiseks, aga seal kuskil nende vahel asub kämblalaiune, kuid ületamatu kuristik?

2) kuidas küll saab olla nii, et eesti verbid *lubama* kui 'permit' ja kui 'promise' satuvad Sinu klassifikatsioonis külmalt eri kohtadesse ega kohtu enam kuskil ega kunagi?

Sina oled neli aastat tagasi kirjutanud **hingest**, meelest ja vaimust kena inspireeriva artikli, esimese taolise katse vaadelda seda eesti keele mõistevälja kognitivistlikus vaimus (Õim 1997). Mina olen veelgi neli aastat varem kirjutanud ja samuti neli aastat tagasi avaldanud silmapaistvalt kauni neandertali luuletuse mehe, **hinge** ja mure suhetest (Krikmann 1997), mis, tõsi küll, ei ole esmakatse vaadelda neid suhteid eesti luules üldse, kuid on kahtlemata esmakatseks vaadelda neid suhteid neandertalistlikus vaimus. Mina kui luuletav šarlatan püüdsin lugejat petta ja väita, et kui mees ise on hinges ja mehel on veel omakorda hing sees, siis tekib topoloogiline paradoks. Sina sellele liimile ei lähe ja näitad oma artiklis, et hing käitub sootuks teistmoodi kui teised "mentaalsubstantiivid", või õigemini: kui teised käituvad selles "üldtopoloogilises" plaanis ühtmoodi, siis hing käitub kahtmoodi. Hing võib tähendada inimese suhtes nii **välist** lokatsiooni, olemissfääri (ja on selles plaanis siis suuresti elu sünonüüm) kui ka mentaalset ainest inimese **sees**, ülejäänud saavad tähistada ainult viimast.

Käesolev kirjutis tahabki jätkata juttu hingest ja elust. Hing on ajaliselt-ruumiliselt (ja siis ka temaatiliselt) palju ulatuslikum (ja seega ammendamatum) substants kui elu, sest hinged võivad teatavasti jätkata oma rännakuid ka pärast elu lõppu ja rohkete reinkarnatsioonide kaudu konnata läbi suure hulga erinevaid elusid. Seetõttu räägin alul ainult pisut ja pinnapealselt hingest, seejärel veidi üksikasjalisemalt elust. Metafoorsete tekstide genereerimise ja arendamise metodoloogiliseks lähtekohaks on olnud klassikaline ajaproovi

läbinud LLP-tehnika.[1] Töö põhieesmärgiks on leida optimaalne lõikepunkt või selle rahuldava täpsusega lähendpiirkond olukordadele, kus

a) normaalmõõtmelise artikli pikkus on saavutatud;
b) on leitud mõistlik ettekääne viidata 30-le (eelistatavalt võõrkeelsele) tööle;
c) on tekitatud minimaalne vajalik hulk lüürilisi kõrvalepõikeid (ingl. *notes*, *endnotes*) ja
d) autor on end oma jutuga lõplikult nurka kinni sõitnud.

Elu ja hing kui väliskonteinerid

Kohakäänete semantika kohta märgib "Eesti keele grammatika" (EKG I: § 28–30), et nende esmatähenduseks on väljendada vastavalt latiivse, lokatiivse või separatiivse tähendusega kohta, peale selle veel ka samade tähendustega aega ja seisundit, lisaks *mutatis mutandis* muidki asju. Kõiki neid jt tähendusvariante käsitatakse põhimõtteliselt alternatiivsetena ja nad tuuakse jadaloeteluna. "Metafoorikognitivistika" jaoks oleks neist otsetähenduslik ilmselt ainult esimene, kohakohane tähendus. Ajakohane oleks hüpermetafoori (või baasmetafoori, või kontseptuaalse metafoori) TIME IS SPACE realiseering, seisundikohane oleks teise hüpermetafoori STATES ARE LOCATIONS realiseering. Ja koha-kategooria ning aja- ja seisundikategooria suhete tõlgendus oleks muidugi kardinaalselt teistsugune: koha-kategooria (või -domeen? või -väli? või -mõõde?) saaks neis realiseeringuis allika (*source*), aja- ja seisundi-kategooria aga sihi (*target*) kvaliteedi, ja muidugi postuleeritaks allika ja sihi vahele projektiivsed suhted.

Kui me niimoodi ühest ebaadekvaatsest lineaarsusest oleme lahti saanud, oleme jõudnud teise ebaadekvaatse lineaarsuseni, sest kognitivistlik metafooriteooria on seni tegelnud peamiselt mõistedomeenide ~ skeemide ~ mentaalsete ruumide ~ ... paaritamisega kontseptuaalseteks vm metafoorideks ja väga vähe huvi tundnud nende paarikute endi kontseptuaalsete või semantiliste vahekordade vastu (välja arvatud ehk John Grady ja mõnede teiste uurijate ponnistused – vt nt Grady, Taub, Morgan 1996; Grady 1997, 1999), nõnda et A IS B -vormile taandatud paaride kogum näib jälle alternatiivide nimistuna, kuigi kõikvõimalikud sünonüümia-, hüpo/-hüperonüümia-, eeldus/järeldus- jm suhted paaride ja nende komponentide vahel (kas või mõistete RUUM ja ASUKOHT vahel) on päev-

selged ja me põrkame neile igal sammul. Kognitivistid on neid ka ise ammu märganud, kuid ei tee neist endale probleemi, sest seda probleemi oleks raske lahendada.

Kui eesti keeles öeldakse, et keegi *on elus*, siis on see teoreetiliselt (olgu ehk konventsionaalne, aga ikkagi) metafoor, kuigi pole selge, mis vahekordades täpsemalt on (ilmselt "geneerilisemad") RUUMI-metafoorid (ilmselt "spetsiifilisemate") ASUKOHA- ja KONTEINERI-metafooridega. Kuid kuna viimane neist on ilmselt täpseim või sisurikkaim ja sihtalal on tegu sisekohakäändeliste väljenditega, siis mööngem, et ehk võib siin siis elu vaadelda konteinerina.

"Inessiivset olekut" kokkusiduvaist verbidest sobib elu ja hingega vägivallatult (st. keelevea või nalja muljet loomata) eelkõige ja ainult enam-vähem sisutühi *on*, ehk ka *püsib* ja *seisab*, kuid nende kohale kerkib siiski hoomatav konnotatsioon 'hädavaevu' või 'millestki hoolimata'. Kuid kui kujutleda kommunikatiivseid põhjusi öelda välja lauseid *Ta on elus* või *Ta on hinges*, siis võib öelda, et samad konnotatiivsed varjud tekivad tegelikult ka *on*-verbi korral. Kumbki viimastest lausetest ei loo assotsiatiivse KONTEINERI-metafoori tunnet ja hinge on eriti raske kujutleda konteinerina või inimese muu ruumilise asukohana üldse. Teisalt jätab lause *Ta on hinges* (mulle vähemalt) palju "stilistilisema" või "fraseoloogilisema" mulje kui täiesti tehnilis-neutraalne *Ta on elus*.

Latiivsete ja separatiivsete siirete märkimiseks on aga konnotatsioonivaba, mittemetafoorse taustaga "bukvaalverbi" peaaegu võimatu leida.

Ärkab (*ellu* ~ *hinge*) näib otsekohe toovat kaasa liite metafooriga SURM ON UNI / ELU ON VIRGEOLEK, õieti KONTEINERI-metafoori taandumise või tuhmumise kogunisti.

Tõuseb (*ellu* ~ *hinge*) toob kaasa liite metafooridega, mida võiks tituleerida umbes: POTENTNE ON ÜLES; ÄRKVELE ON ÜLES; HEA ~ PAREMAKS ON ÜLES, ning ühtaegu samuti KONTEINERI-metafoori taandumise või tuhmumise.

Sõnaühendite *ärkab ellu* ~ *hinge* ja *tõuseb ellu* ~ *hinge* tähendusliku laienemise võimalused väljapoole inimolendite füüsilist elluvirgumist on segaselt tajutavad, kuid nähtavasti üsna erinevad. *Hinge*-väljendid sobiksid ehk pigem loomsete või taimsete üksikorganismide, ka nt süte ja tule kohta, *elu*-väljendid aga mingite üldisemate liikumise, ettevõtete, protsesside, nt kevadise looduse kohta. Kui need asjad *ärkavad uuele elule*, märkame siin uut selget, kuid

taas raskesti kirjeldatavat tähenduslikku ja stiililist nihet. Ühendi *tõusis hinge* mõtestused nt lausetes *Krookus tõusis hinge* ja *Kahtlus tõusis hinge* on drastiliselt erinevad: esimesel juhul on ebaselge, kus see adverbiks grammatikaliseerunud hing asub, kui üldse asub; teisel juhul näib, et tegu on süntaktiliselt puuduliku lausega, kuna me eeldame, et kõnealune hing kuulub **kellelegi** ja tõenäoliselt asub tolle kellegi **sees**, aga siin viide hinge omanikule puudub.

Kaasaegses ruumikantseliidis neutraalsed *sisenema* ja *väljuma* assotsieeruvad otsetähenduses peamiselt füüsiliste tehisruumide, eelkõige hoonete või tubadega ning nende ustega ja sobivad metafooridesse mentaalsete ruumide ja alade kohta vaid üksikuil piiratud erijuhtudel (*püüame siseneda poeedi mõttemaailma*; *meeleavaldus väljus kontrolli alt* vms). Seoses *eluga* mõjuvad nad kõige pigemini kavatsuslike, kuid saamatute naljadena; sõnaühendeisse *ta siseneb hinge* ~ *väljus hingest* vmt pole võimalik sisse puhuda vist mingit mõistlikku tähendust.

Teise kantseliidinäolise antonüümipaari *saabuma* : *lahkuma* komponentidel on elu ja hingega seondumise võimalused veel kirjumalt erinevad. *Lahkus*, *lahkus meie hulgast* ja ka *lahkus elust* kuuluvad klišeedena teatavasse kindlasse "surmakuulutuste" allkeelde ja märgivad surma, *lahkus elust* eelkõige just enesetappu. *Saabus hinge* ja *lahkus hingest* mõjuvad keelevigade või absurdidena – viimane taas juhul, kui hing ei ole kellegi **oma** ega asu omaniku **sees**: vrd nt ?*Naine lahkus hingest* ja *Paine lahkus hingest*.

Hinge seostamine surma tähendava separatiivse operatsiooniga näib üldse võimatuna. Võib-olla siis on "välishinge" tõlgendamine koha või mahutina ülepea õigustamatu – seda ka lokatiivsetel ja latiivsetel normaaljuhtudel? Ehk on vormid *on hinges* ~ *tõusis hinge* jts üldse seostatavad *hing*-sõna nende alltähendustega, mis viitavad hingamisele, õhuvahetusele – n.ö "on hingamas" ~ "tõusis (üles) hingama", mis on semantiliselt lähedaste *elus*- ja *ellu*-vormidega sarnaseks assimileerunud? Või ehk on see pärit *hing*-sõna veel ühe alltähenduse 'inimindiviid, "-üksus"' (*seitse hinge peret*; *mitte elavat hingegi*; *surnud hinged*) inessiivsest **mitmus**vormist, mis võis ka eesti keeles algselt tähendada '(elusate) inimeste **hulgas**' (vrd. soome *on hengissä ja voi hyvin* vms), ning assimileerunud hiljem sünonüümse *elus*-paralleeliga analoogiliseks, ainsuslikuks.

Elu, kuhu *astutakse*, on kindlasti ka "topos" (pigem konteinerist ebamäärsema siluetiga), kuid *ellu astuma* ei tähenda 'taaselustuma',

vaid see käib peamiselt mitmesugustest inkubaatoritest väljuvate inimolendite, eriti vist abiturientide jm koolilõpetajate kohta, ja elu vastandubki sellele sotsiaalsele embrüo-olekule kui "tõeline elu", "iseseisev elu", isegi vast "senisest karmim elu".

Tulema ja *minema* on siirdumist märkivaist verbidest kõige tehnilis-neutraalsemad, ülipolüsemantilised ja igal viisil grammatikaliseerunud, kuid nende latiivsed ja separatiivsed mõtestused sõltuvad suuresti nn vaatepunktist ega ole üheselt kindlad. Ma ei tea, kui normaalseteks võib lugeda lauseid *Ta tuleb ellu* või *Ta tuleb hinge* tähenduses 'Ta ärkab ellu', minu idiolekti nad igatahes ei kuulu.

Kui aga elu poleks elavate sfäär üldse, vaid **kellegi** elu kui teda ümbritsev individuaalne mitmemõõtmeline sfääristik ja ühtlasi ajas kulgev protsess, siis paraneksid meie lausemoodustamise võimalused järsult.

Minu ellu saab tulla mitmel moel

Ta tuli mu ellu on normaalne, pateetilis-magedale stiilialale omane väljend, mis võib tähendada ka püsiva tutvuse või sõpruse teket, ennekõike vist siiski armastussuhete sõlmumist. Ühtlasi annab sõnaühend *minu elu* ajendi väita, et kognitivistlikus metafooriteoorias tekkivail hullustel ei näi olevat üldse mingit piiri. Vaevu oled end suutnud harjutada mõttega, et iga viimane kui kohakäändetarvitus, inglise keele prepositsioon *in(to)* vmt loob paratamatult ruumi- (st. asukoha-, konteineri- vm) metafoori, kui Kurt Feyaerts (2000: 66) kuulutab, et ka kõik ingliskeelset *have*-sõna sisaldavad laused konstitueerivad metafoori – ATTRIBUTES (PROPERTIES) ARE POSSESSIBLE OBJECTS: kui sul on nt köha, pohmelus või hemorroidid, siis pea neid kalliks, sest nad on su omand ja omand on vaikimisi üks hea asi.[2]

Siit pole raske jõuda mõtteni, et kõik eesti vm keele genitiivid lubavad või sunnivad tõlgendama endid samuti OMANDI-metafooridena. Ja ega olegi eriti loomuvastane kujutleda, et minu elu on ühest küljest sfäär või konteiner, kus ma asun, teisest küljest minu omand – on ju inimestel ka autod, majad, jahid jne. Kuid kuskil ei leidu õpetust, mis taoliste topeltmetafooridega peaks teoorias peale hakatama.

Edasi on rikkalikud võimalused kellegi ellu tuleku situatsiooni igas mõõtmes nüansseerida ja erinevaid metafoorseid assotsiatsioone mängu tuua.

Ta tormas mu ellu – see liikumine on kiire ja energiline ning etümoloogiliselt seotud veel naturaalsete tormidega. Taolised etümoloogilised kumad näivad tõesti tihtipeale kinnitavat, et "otsene" ja "figuratiivne" tähendus pole mingid kahevalentsed alternatiivid, vaid meie ees laiub poolkontinuatiivne kujutluste ja assotsiatsioonide kogum, millest mõned on vähem, teised rohkem salientsed ja vahel paistab kõigi ülejäänute hulgast välja mingi "kõige salientsem". Kaasaja eesti keeles on *tormamine* kindlasti juba mingi võimas, kiire (ehk veidi ülepeakaela kiirustav) kindla suunaga liikumine mingit pinda mööda (ja/või kuhugi sisse ja/või kuskilt välja) ning seondub peamiselt inimeste, rongide jt suurte sõidukite, piisonikarjade jms-ga. Tormavail objektidel peab peale kiiruse ja sirgevõitu trajektoori olema ka piisavalt massi: on imelik öelda nt *Kuul tormas püssist välja*, *Pall tormas mulle vastu pead* või *Hiir tormas auku*. Meteoroloogiline *tormamine* kui 'tuisk' (*Ilm tormab*; *Väljas tormab*) on tänapäeva eesti keelest ilmselt taandumas.

Nõtke kassina kargas ta mu ellu – siin on lisaks toimuva füüsilisustamisele mängus veel loomametafoor (või õieti võrdlus), *kass*, mida võiks kujutleda üliproduktiivse metafoori INIMENE ON LOOM esindajana. Selline määrang aga ei ütleks väga palju ilma iga konkreetse loomallika kultuurilis-semantilist spetsiifikat teadmata.[3]

Kassi-metafoori kontekstis hakkab meile eriti visalt tunduma, et KONTEINERI metafoor on tegelikult tõesti üliavar (loe: vähesisukas) ja paneb pigem paika kohalisuse kui mõtestusviisi üldse, kui et suudab midagi öelda just **elu** kui konteineri lähema loomuse kohta. Elu, kuhu kassitaolised olendid võivad sisse, peale või külge karata, hakkab tahes-tahtmata assotsiatiivselt täpsustuma — "etnograafilisel ajastul" tavatsesid nad karata eelkõige ehk villavakka või ahju peale või siis lauale paha tegema, tänapäeval ehk pigem voodisse kaissu, päris universaalselt aga kindlasti hiirte kallale. Nende lokatiivide keelelised väljendused ei esinda enam formaalset inessiivi, mis lubab küsida, kas nad esindavad enam KONTEINERI metafoori. Igatahes aga esindavad nad kassi liikumise füüsilist sihtkohta koos kõigi sinna juurde kujutletavate hedooniliste ja utilitaarsete taotlustega. Ja miskipärast tundub meile päris kindlasti, et elu omanik on siin meessoost, "kass" aga naissoost (vastasel korral oleks ilmselt valitud mõni suurem kaslane – tiiger, panter vmt). Või siis oleks kassi sooline kuuluvus selgesti markeeritud: *Küür kühmus, loivas see tokerdanud **kõuts** mu ellu*.

Unustame nüüd hetkeks *kassi*-kujundi ja elu kui KOHA metafoori liitmise probleemid: kuimitu allikdomeeni siin on mängus, kuidas neid peaks tituleeritama jne. Käesolev näide võiks aga inspireerida jätkama ka põlist ja lõputut arutelu metafoori- ja võrdlusvormi suhete üle (vt nt Ortony 1979; Indurkhya 1992: 26–28 jpt; Fass 1997: *passim*; Bredin 1998; Nogales 1999: eriti ptk 6; Aisenman 1999; Kennedy, Chiappe 1999 jpt).

Essiivset võrdlusvormi *Nõtke kassina kargas ta mu ellu* ei saa muuta metafoorivormiks sel teel, et võrdlusvahend lihtsalt pannakse metafoorse lause agendi (väliselt: metafoorse subjekti) positsiooni. Lauset ?*Nõtke kass kargas mu ellu* mõistaks normaalne eestlane kuuldelisel vastuvõtul vist ainult nii, et keegi indiaanlane räägib siin oma pruudist nimega Nõtke Kass. Mitte-sententsiaalmetafoorse lause subjekti positsioonis olevaid indiviide märkivaid (ja üldse eksistentsikvantori all seisvaid) sõnu saab metafoorideks signaliseerida ainult lisavahendite, eelkõige *see*-pronoomeni abil. Katsetus *See nõtke kass kargas mu ellu* on aga esiteks stiililiselt täiesti hirmuäratav, teiseks jätab mulje, et lause on lõpetamata: "kass" peaks pärast ellukargamist justkui veel midagi ette võtma. Ja kui me püüaksime teisendada oma lähtelauset võrdlusvormi säilitades, siis avastaksime, et me tohime üsna ohutult asendada essiivi tavalise *nagu*-võrdlusega, kuid ei tohi likvideerida inversiooni. *Nagu nõtke kass kargas ta mu ellu* kõlab palju normaalsemalt kui *Ta kargas mu ellu nagu nõtke kass*. Sõna *nõtke* ja inversioon kuuluvad lahutamatult kirjakeelde, järjestusviis, kus võrdlusvahend asub kõige lõpus, on aga iseloomulik just rahvapärastele võrdlustele. Nagu Katre Õim on näidanud, kipuvad rahvapärastes võrdlustes võrreldava ja võrdlusvahendi vahele sugenema (kui ka mitte eksplitsiitselt, siis juba mõttelise automatismina) erilised "švaa-ainesed", seega jätab viimane, inversioonitu näide mulje, et lause tegelik sisu on umbes: *Ta kargas mu ellu nagu üks va igavene kuradi nõtke kass*, mis tundub üheselt kavatsusliku paroodiana.

Ilmatu prahmakaga maandus ta mu ellu – siin kostab juba helisid ja lendab kujuteldavasti tolmu või kilde, miski täriseb ja vappub, talad värahtavad. Kuid meile ei assotsieeru siin võimsad lainerid ega reipad parašütistid, kuna prahmakad oleksid sel juhul letaalsed ja sobimatud. Ning elu ei ole siin kindlasti lennuväli. St kui me *prahmakat* usume (ja miskipärast me usume), siis *maandumine* ise on siin juba semantiliselt nihutatud, maksiime rikkuv[4], ja kvalifitseerub ehk

irooniana, kuna tegelikult viidatakse ontoloogilises plaanis sündmusele, mis mõtestub hädamaandumise või allakukkumisena. Meil on päris kindel tunne, et ütleja on seekord naissoost ja maanduja meessoost. Meie meel ei visanda siia ei selget lennukit ega langevarjurit, kuid kuskil teadvuse horisondil nad hõljuvad. Me võime nad sealt muidugi ka välja aktsenteerida: *See, kuidas ta mu ellu tuli, oli tõeline hädamaandumine* võib assotsieerida kujutelmi toimunu soovimatusest, ravi ja hoolitsuse vajalikkusest kukkunu suhtes vmt. Kuid me võime ka lennundusliku domeeni mängust täiesti välja lülitada:

Ilmatu prahmakaga sadas ta mu ellu – siin lennundust ei ole, aga me tajume, et tegu pole ka *sadama*-verbi kõige prototüüpsema (redundantsema, "salientsema") alltähendusega (ja võib viivuks imestada, kuidas ere salientsus ja tuhm redundantsus võivad teineteise sünonüümideks olla). See esmaassotsieeruv või automaatseim alltähendus, n.ö 'vihma või lume kukkumine' on konventsionaalselt laienenud tähendamaks 'ükskõik mille alla- või mahakukkumist' juba ammu enne ja kaugel väljaspool meie näitelauset. Kuhu täpselt ta on konventsialiseerunud – kas polüseemia ilminguna keelde või fraseoloogilise ilminguna retoorilisse folkloori, seda ma ei oska öelda. Igatahes lisab kukkumist tähendav *allasadamine* öeldavale humoristliku nüansi, mis olukorrast ja ütleja taotlusest sõltuvalt võib saada kas pehmendava või põlastav-parastava allnüansi.[5]

Elu kui muld ja käsn

Pikkamisi imbus ta mu ellu – pahade traditsionalistide jaoks moodustaks siin metafoori, semantilise "võõrkeha" ilmselt ainult sõna *imbus*, kõik ülejäänu tundub laitmatult otsesõnalisena – st selle sõna taga peituv mõiste tuleks asendada mõne muu, konteksti sobiva mõistestruktuuriga, nt 'laskis mul harjuda ta kohaloluga (mu elus)' vmt. See väldiks paljud edasised komplikatsioonid, kuid me ei tohiks küsida, kuidas meie meel sellised sobivusotsustused tuletab. Kognitivistid näeksid siin kindlasti põhimõttelist sihtdomeeni (mis on antud juhul visandatud 4 sõnaga) ja põhimõttelist allikdomeeni (mis on siin markeeritud üheainsa sõnaga *imbus*). Kui hakata mõtisklema, mis allikdomeen see selline võiks olla ja mis nime võiks kanda elule vastav allikmõiste, millele kohaldatakse metafoorne predikaat *imbus*, siis tundub, et konteiner see igal juhul enam ei ole. Elu tunduks siin kõige pigemini MULLANA, võib-olla ka KÄSNANA, kuid miskipärast pole võimalik neid oletatavaid metafooripõhju leksikaalselt eksplit-

seerida. Lausetele *Pikkamisi imbus ta mu elumulda* ja *Pikkamisi imbus ta mu elukäsna* on raske mingit selget sisu omistada ja stiili poolest on nad mõlemad grotesksed peletised. Selle vastikuse põhjused on kummalgi juhul vist küll erinevad.

Elul ja surmal on mõlemal tugevad metonüümilised püsiseosed mullaga ja mõlema kaudu annab arendada ka metafoore. Mida tuleks arvata nende metafooride allika domeenilise päritolu kohta, on taas raske küsimus. Igatahes kuuluvad nad mõlemad BIOLOOGILISE RINGKÄIGU stsenaariumi, kuid esindavad kindlasti selle eri faase: elu korral toitepinnast (*Idee sattus viljakasse mulda*), surma korral lisandeid, mis kanduvad mulda zooloogiliste või botaaniliste organismide lagunemise tulemusel. See stsenaarium on metonüümiliselt seotud (kuid ei kattu) inimolendite surma puhul assotsieeruva HAUDAPANEKU stsenaariumiga.

Elumuld saaks siis olla botaaniline metafoor, TOITEPINNASE metafoor – umbes 'miski, millest kellegi elu jõudu ~ energiat ~ ... ammutab' ja ta aksioloogiline märk peaks olema ilmselgelt positiivne.

Niipea aga, kui me *imbumise* tõttu oleme elu tõlgendanud MULLAKS (ükskõik, kas seda sõna otse kasutades või mitte), peaks miski sihtstsenaariumis justkui sattuma "ontoloogilisse vastavusse" TAIMEGA, mis sellest mullast toitub. Meil võib tekkida mõtisklusi selle üle, kas lauses, mis lõpeb sõnaga *ellu*, on taimeks kindlasti elu omanik ise (sest muld oli siin projitseeritud justkui elule), lauses aga, mis lõpeb sõnaga *elumulda*, esindab *muld* ise juba metafoorset allikmaailma ja taimele vastava elemendi määratlemiseks, olenevalt sellest kuidas genitiivset liitsõna tõlgendada, jääb rohkem võimalusi: kui *elumuld* panna tähendama 'elu **kui** mulda', saab taimeks olla vaid elu omanik; kui *elumuld* panna tähendama metafoorset mulda, mis on kuidagi **seotud** sihtmaailmas viidatava eluga, võiks taimeks ilmselt olla nii elu omanik kui ka elu ise või nad mõlemad kokku. Kui me oleme niimoodi arutlenud, tabab meid kuskil terav ebaadekvaatsustunne ja vaistlik tajumus, et niimoodi arutleda on skolastiline ja mõttetu, sest inimesed ei mõtesta kujundeid niimoodi ja *elumulla*-kujund, mida me arendada püüame, pole üldsegi eriti lootustandev kujund.

Kui lubame, et sihtmaailma elule vastas allikmaailmas MULD ja seda elu elavale inimolendile vastavalt PUU vm TAIM, siis metafoorse imbumise subjekt mõtestub kõige pigemini ehk VEDELIKUNA.

Sellega on "imbuja" mitte üksi depersonifitseeritud, vaid viidud elutute ainete klassi, mis tõepoolest võib ütleja kommunikatiivsetele taotlustele hästi vastata, kui ta nt soovib lähenejat esitada passiivse, tahtetu, "iseeneslikuna". Teisalt aga, kui allikskeem on viidud oma täiskujule, kus sihtskeemi *tema*-tegelasele vastab midagi vedelat või püdelat, siis hoolimata sellest, et *tema*-tegelase ja -tegevuse aksioloogiline märk sihtskeemis oli selgelt negatiivne, on see allikskeemis endas (kust ju kogu projitseeritav uus info peaks pärinema) esialgu täiesti umbmäärane. Teades, et mõistatuse 'vedelik, mis spontaanselt imbub mulda' tõenäolisim vastus on VESI, võib öelda, et spontaansus ja aksioloogiline markeeritus satuvad siin koguni vastuollu, sest keskmise eestlase jaoks on prototüüpne mulda imbuv vesi pigem kosutav vihm kui närvutav liigniiskus, kuigi esiisad on tõesti vormistanud ka ütluse *Põua lapsed naeravad, vihma lapsed nutavad.*

Ülal püüti väita, et mitte-sententsiaalmetafoorse lause subjekti positsioonis olevaid indiviide märkivaid (ja üldse eksistentsikvantori all seisvaid) sõnu saab metafoorideks signaliseerida ainult lisavahendite, eelkõige *see*-pronoomeni abil – vt kas või siin mujal toodud lausenäiteid ***See siug puges vargsi mu põue*** jt (N. Arutjunova (1979: 151) arvab oma lausenäidet ***Эта*** *ёлка* ***всегда*** *все* ***нумаем*** kommenteerides, et taolisi metafoore kasutataksegi just identifitseerimise eesmärgil, mis esmaselt on hoopis metonüümia funktsioon.) Sõnu, mis oma otsetähenduses märgivad mitteloenduvaid asju, näib olevat raske panna metafoorsete subjektidena inimindiviide tähendama: ?*Pikkamisi imbus* ***see vesi*** *(~* ***solk*** *~* ***mürk*** *~* ***väetis*** *~ ...)* ***mu elumulda.*** Üldisuskvantori all sententsiaalmetafoorides on see muidugi täiesti võimalik: *Vaga vesi – sügav põhi.*

Kui me säilitame *tema*-isiku esituse otsekeelsena, siis näib tekkivat võimalus neid sõnu ka "possessiividena" metafoorse skeemi esitustesse sisse tuua. Kuid antud leksikaalsetes konfiguratsioonides jääb seegi võimalus pelgalt teoreetiliseks. Lause *Pikkamisi imbus ta mürk mu elumulda* tagant võivad semantilised kavatsused olla äratuntavad, nagu ka see, et toimunut tahetakse esitada *mina*-tegelase jaoks halva, kahjulikuna. Kuid metafoorse lausena logiseb ta nii mitmest kandist, et raske loetleda. Aksioloogiliselt positiivse analoogi leidmine on veel lootusetum. Lünklause *Pikkamisi imbus ta* [x = hea, kasulik vedelik või lahustuv aine] *mu elumulda* esindab situatsiooni, mis näib VÄETAMISE või KASTMISENA. Variandid x = *väetis*, *sõn-*

nik, *guaano*, *superfosfaat* jne resulteeruvad aga groteskidena, mille juures on raske valida, kas naerda või nutta.

Kui tõlkemasin valiks heauskselt KASTMISE skeemi ja kirjutaks prototüüpseimal viisil *Pikkamisi imbus ta **vesi** mu elumulda*, ei oleks seegi vist ka mitte halva stiili näitena vastuvõetav, sest vesi ei hakka meile siin mõjuma kosutava vihmana, vaid millegi lahjana, laialivalguvana, sisutuna (see metafoorne voolusäng on *veel* tugevasti sisse uurdunud – vrd valvenäited, nagu *Sellest kirjutisest annab rohkesti **vett** välja pigistada* vmt), või siis parasiitveena, mis tungib inimese poolt loodud tehiskeskkonda. *Pikkamisi imbus ta reovesi mu elukeldrisse* vmt oleks iseendast ju tõesti võluv lause, kuigi *elukeldri* metafoorile ei leidu ei eesti ega muus keeles kindlasti mitte mingit "arhiveeritud" tähendust: kelder on liiga argine ja tehniline koht, et teda panna tähendama nt elu rõskeid ja pimedaid süvatasandeid, mida harva külastatakse ja puudulikult kontrollitakse.

Keegi võiks vastuvõetavamaks pidada aksioloogiliselt veel selgemini lahtikirjutatud lahendeid *x* = *kosutav vihm*, *viljastav mahl*, *virgistav neste*, *vägi* vmt. Kuid väe agregaatolekut kui ka esmakuuluvust allik- või sihttasandi leksikasse oleks päris raske määrata. Ja üsna üldiselt jääb õhku rippuma või evotseerib ebasoodsaid vastuseid küsimus, mida sisulisemat possessiivsuhe *tema*-tegelase ja *x*-aine vahel endast kujutab ja kui aktiivset rolli *tema*-tegelane mängib *x*-aine muldaimbumisel. Selle nurga alt vaadatuna oleks ilmselt soodsaim, kui ta suudaks *x*-ainet kuidagi loomupäraselt ja spontaanselt **eritada**. Kui selline *x*-aine oleks MÜRK, siis muutuks *tema*-tegelane maoks, skorpioniks, kuraareks jne, ja kõik on kena, kuni me ei taipa, et mullasse immutatud mürk on tegelikult tulutult äraraisatud mürk. Kui *x*-aine oleks MAHL, oleks võimalikke tõlgendusvariante kaks, mõlemad jälle halvad: a) *tema*-tegelast on kavatsetud esitada taimena, nt kasena, mille mahl iseeneslikult tilgub mullasse (ja kuulaja ei oska sellest mingeid järeldusi tuletada); b) *tema*-tegelane valab meelega mingist anumast talle kuuluvat mahla maha ja see imbub mullasse (ja kuulaja ei saa jälle aru, milleks ta seda teeb). Kui kellelegi kuulub VIHM, siis saab see olend olla ilmselt ainult Jumal. Mingeis kontekstides võiks lause, nagu *Jumala kosutav vihm imbus mu janunevasse elumulda* olla ehk isegi vastuvõetav metafoor; kuid meie juhul on selge, et jutt ei ole Jumalast.

Elu kui KÄSN on täiesti lootusetu arendussuund. Muidugi ei tõukaks lähtelause meid mõtestama *käsna* puupahana või soola-

tüükana, vaid kõige pigemini pesukäsnana, "svammina". Sellist käsna saab aksioloogilise plussmärgi all kõnesse tuua vist ainult võrdlusvahendina inimese kohta: *Nagu käsn imes ta endasse muljeid ja teadmisi.* Ning taas: me ei saa viimasest lausest kaotada inversiooni, sest siis tekiks väljund *Ta imes endasse muljeid ja teadmisi nagu* [lisandub implitsiitselt: *üks ~ va ~ igavene ~...] käsn*, mis oleks hinnanguliselt kahemõtteline, kui mitte juba selgelt negatiivne. Siin satutakse ühtlasi juba väga lähedale teisele, "mitteimavale" käsnale, mis on üheselt pejoratiiv ja sõimuvahend: *Sa kuradi käsn! Tema, vana käsn, pole mulle mingi konkurent* vmt. Käsnal puudub eluga ühine positiiv-neutraalne märgistus.

Elluhiilivaist ja põuepugevaist madudest

Vargsi hiilis ta mu ellu – taas on raske täpselt öelda, millise objektina elu meile siin tundub. Igatahes sinna sisenetakse ja kuna sisenemine on vargne ja hiiliv, siis konteineri või asukoha üldskeem kipub täpsustuma mingiks hallatavaks või kontrollitavaks territooriumiks, kus eeldatavasti paikneb midagi väärtuslikku, mida hiilija sooviks kätte saada, aga eluterritooriumi haldaja ei taha loovutada. St ellutulek meenutab siin pigem VARGUSEKS VALMISTUMISE, mitte niivõrd TAPMISEKS VALMISTUMISE skeemi. Kuid intervendi üleviimisega loomadomeeni võib lähendada ütluse ka viimasele tähendusvariandile.

Vargsi hiilis see siug mu ellu – siin ei peeta ilmselt silmas mitte elu kommunikatiivseid aspekte (suhete kehtestamine vmt), vaid mingit ohtlikku lähemalenihkumist, millest elu omanik ise teadlik ei ole. Kui *hiilis* asemel oleks *puges*, siis oleks selge, et siug jõudis oma eesmärgile, ja ühtlasi tahaks päralehiilimise kohta termineerida sel juhul enam mitte *eluna*, vaid *põuena*.

See siug puges vargsi mu ellu – siin on allusioon kliõeelisele põuepugemisele juba nii tugev, et kiirel *online*-vastuvõtul, kus pole aega teha rahulikult poeetilist süvaanalüüsi, võibki vähemkliõeeline elu hakata tunduma millegi põuetaolisena.

See siug puges vargsi mu põue – siin tekib mulle arusaamatute põhjustega paradoks. Ühest küljest näivad viimane ja eelviimane lause täiesti sünonüümsetena, umbes: 'See kurja kavatsev inimene poetas end vargsi mu lähedale', kuigi ehk tõesti tunneme, et *põue* istub siia konteksti paremini kui *ellu*. Kuid ainult seniks, kuni me ei ürita analüüsida nende lausete troobistruktuuri. Kui me seda teeksime,

peaksime kvalifitseerima *põue*-lause kindlasti täislauseliseks e. sententsiaalmetafooriks, *ellu*-lause aga üsna ebatüüpiliseks "taanduvaks" partsiaalmetafooriks. Lähemal vaatlusel, nagu peatselt näeme, pudeneks seegi uskumus koost.

Eestlane märkab kõigepealt, et neutraalse *mao* asemel on mõlemas näites tarvitatud murdelis-arhailis-poeetilist *siugu*, mille sõnastikessegi kandunud esmatähenduseks näib tänapäeval olevat juba pigem 'kuri, salakaval inimene' kui 'madu', ehk teisiti: kas tänapäeva eestlase jaoks on siug üldse enam konventsionaalnegi metafoor või ainult selle etümoloogiline jälg? Edasi tundub, et lausel *See siug puges vargsi mu põue* või ka *See madu puges vargsi mu põue* ei olegi üldse puhtotsest tähendust tõelises, pragmaatiliselt valiidses mõttes, kuna ta ainumõeldav kontekst peaks olema selline: keegi näitab kellelegi ussi ja väidab, et too puges talle kunagi lähimineviku põue. Kuid kõik eestlased jt eurooplased, asiaatidest ja aafriklastest rääkimata, teavad väga hästi ütlusi põue pugevate, rinnal toidetavate, rohus roomavate jm madude kohta ning madudega seonduvaid psühholoogilisi ohu- ja jälestusstereotüüpe üldisemalt. Kõik nad teavad, kui vähetõenäoline on bukvaalmao bukvaalne põuepugemine reaalse sündmusena ja kui veider liiati, et keegi sellest hiljem rahumeeli räägib.[6]

Kas me saame aga isegi lauset *See madu puges vargsi mu põue* käsitada laitmatu "loomakohase" sententsiaalmetafoorina? Vaatame selleks metafoorseid konstituente lähemalt.

Põuepugemine on see lookus, kus aktantide vahel toimuvaid suhtlemissündmusi ja/või eetilisi sündmusi ja/või sotsiaalseid sündmusi esitatakse (mõistestatakse, struktureeritakse) füüsiliste (allik)-sündmustena. Nii põuel kui ka pugemisel on lisaks omaette metaforisatsioonid, mis koondmetafoori mõtestumisel samuti kaasa mängivad, tuues ühtlasi kaasa uusi kontseptuaalseid linke.

Pugemine lähtub metafoorina eelkõige mitmesuguste alandumisaktide visualisatsioonidest (sünonüümid nt *lömitama*, *küürutama*, *kintsu kaapima*, *taldu lakkuma* jne), seega võib siin näha füüsilise allika projitseerimist sotsiaalsele sihile, konkreetsemalt ehk metafoori ALANDLIKKUS ON FÜÜSILINE MADALASEND. Meie kontekstis viimane ei sobi. *Pugemise* otsemõtestuses leiduvad komponendid LÄHENEMINE ja SISENEMINE annavad vihje, et võidakse viibida laiemas metafooriparadigmas, kus intiimsuse, sõpruse, emotsionaalse seotuse astmeid tavatsetakse esitada FÜÜSILISE DISTANTSINA, piir-

juhul FÜÜSILISE ÜHINEMISENA ~ SISENEMISENA ~ ÜMBRITSEMISENA. *Põu* on intiimsusdistantside skaalal samuti väga pooluselähedane lookus. Rõivaid kandev inimene on kahekihiline ja põues on pealiskiht juba läbitud ning kehaline kontakt vahetu. Põu on erootiliste assotsiatsioonidega küllastatud koht.

Kuid meie metafoor ei saa mingit emotiiv-aksioloogilist identiteeti, kuni selles ei osale kujutlused *mao* vingerdavast lähenemisest, külmast ja limasest puudutusest (kuigi osa neid on tegelikult kuivad ja soojad), keha ja riiete vahele liibumisest, mürgise hammustuse ja tulise valu ootustest. Siin asub jõusse ürgne opositsioon, millele on antud nimeks KULTURAALNE : NATURAALNE, KODU : METS vm. Madu põues on looduse tagasitung temalt vallutatud kultuurioaasi. Meenutagem, mida tegid omaaegsed külatüdrukud, kui neile hiir või konngi põue pisteti!

Need seigad vihjavad meile, et selle kujundi motivatsioon ja sihtskeemi mõistestus ei ole veel üldsegi selged. Alliksündmuses näib sisalduvat ilmne eeldus: põue omanik on tunnetus- ja/või teovõimetu, nt kinniseotud, halvatud või kõige pigemini magab. See annab aga alternatiivse või laiendava suuna ka *põue* mõtestamiseks. Põu hakkab meile näima mitte (ainult) suhtlemist, kultuurilist kontrolli, intiimsusastet määratleva lookusena, vaid (ka) **teadvusena, meelena**, ja *põuepugemine* vastavalt teadvusse, meelde tungimisena. Milliseid siinses geneerilises halos suplevate disjunktiivsete mõtestusvariantide omavahelistest suhetest tuleks lugeda metafoorseteks, milliseid aga metonüümilisteks, sellele ei anna senine teooria mingit vastust.

Allikskeemis osalevad inimene ja loom, sihtskeemis kaks inimest, põuepugemine pärineb kindlasti allikdomeenist. Kuid kuhu kuulub mõisteliselt *vargsi*? *Vargsi* on samajuurne sõnadega *varas*, *varastama* ja tema stiililiselt neutraalseim sünonüüm on *salaja*. *Vargsi* prototüüpseim mittemetafoorne tähendusväli jääb tahtlike, füüsiliste, kuulmatute-nägematute, kellegi jaoks pahade ja soovimatute aktsioonide alale, seega inimlikku domeeni. Sellest vaatevinklist võiks sõnaühendis *madu puges vargsi* näha ka personifikatsiooni või antropomorfisatsiooni alget, st "puhta sententsiaalmetafoori" kadu. Kuid adverbide *vargsi* või *salaja* tarvitused nt saagile ligihiilivate röövloomade kohta tunduvad samuti lootusetult mittemetafoorsed ja taas meenub Lakoffi ja Turneri väide loomametafooride "topeldatuse" kohta (vt ka märkuses 3). Igatahes suureneb

metafoorsustunne hüppeliselt mitte üleminekul "mõistuslikult intentsionaalsuselt" "instinktiivsele intentsionaalsusele", vaid üleminekul intentsionaalselt spontaansele. Minu keelevaist ei taju personifikatsiooni, kui öeldakse, et rebane lähenes *vargsi* kanadele, kass hiirele või lõvi antiloopidele, kuid tajub personifikatsiooni, kui öeldakse, et kevad või öö või armastus või haigus on *vargsi* tulnud või majanduslangus on *vargsi* alanud.

Sõnade *mina* ja *vargsi* taga olevad mõisted on seega oma domeeniliselt kuuluvuselt kahepaiksed. Taolised "nõrgad lülid" on metafooridele nähtavasti päris vajalikud, kuna etendavad nende mõtestamisel omamoodi lüüside või päästikute rolli.

Kuna meil on tegu loomametafooriga, kerkib veel üks lisaprobleem.

Selle lause juurde sobivat spetsiifilist allik- ja sihtsündmust pole eriti raske leiutada – näiteks:

a) alliksündmus 'madu puges Jürile põue',

b) sihtsündmus '(kuri ~ salakaval ~ saamahimuline ~ ...) Mari sokutas end Jürile naiseks'.

Kuid neid siduva geneerilise ühisosa juures pole mulle sugugi selge, millisel üldisusastmel peaksid olema kujutletud ja formuleeritud selles figureerivad abstraktse tasandi aktandid, st intervent (kellele allikas vastab madu ja sihis Mari) ja kannataja (kellele nii allikas kui ka sihis vastab mina-tegelane ehk Jüri). Invariantsi Printsiibi senised formuleeringud ei võimalda otsustada, kas geneerilise tasandi intervent peab kujutama endast lihtsalt n.ö impersonaliseeritud inimolendit (nt '(kuri, salakaval) inimene') või nii looma- kui ka inimeseüleseks abstraheeritud olendit. Kannataja näib esmapilgul võivat jääda inimolendiks (sest tal pole allikskeemis ontoloogilist vastet, mis nõuaks sellest ülemale siirdumist). Kuid kui intervent on abstraheeritud kõrgemale, tekiksid geneerilises skeemis endas põhjendamatud tasandivahed.

Ja veel, kui ikkagi lubada, et *põu* võiks mõtestuda ka TEADVUSE või MEELENA ja et *vargsi* oma inimlikus esma-tavatähenduses suudab evotseerida VARGUSE või üldisemalt KURITEO skeemi, mis üldsegi ei lange kokku ABIELLU MEELITAMISE skeemiga ega ole ka selle "ülemhulk", siis jääks üle nentida, et mängu puutuvate sihtskeemide (või -ruumide) hulk on julgelt enam kui kaks. Millised elemendid sellest paljumõõtmelisest sasikerast peaks paigutuma spetsiifilise, mis geneerilise hulka, mis aga kvalifitseeritama võib-olla

blendiks – see küsimus ületab mitmekordselt minu kognitiivsed kapatsiteedid.

Konklusioon

Uurimuse eesmärgiks oli katseliselt kontrollida, kas leidub optimaalne lõikepunkt olukordadele, kus
a) normaalmõõtmelise artikli pikkus on saavutatud;
b) on leitud *good reason* viidata 30-le valdavalt võõrkeelsele tööle;
c) on tehtud vajalik hulk märkusteks nimetatavaid lüürilisi kõrvalepõikeid ja
d) autor on oma jutuga pöördumatult rappa läinud.

Tehtud esialgne katse näitab, et täiesti ideaalset tulemust pole õnnestunud saavutada, kuigi lähendiks on lootustandvalt väike nelinurk. Artikkel on mahult veidi liiga suur, sellal kui viidatud tööde arv piirdub alles 29-ga ja kõigest 2/3 neist on ingliskeelsed. Lüürilisi kõrvalepõikeid on ehk keskmisest normist veidi vähem ja iga kõrvalepõike pikkus keskmisest normist mõnevõrra suurem. Kuid punkti (d) osas on tulem peaaegu ideaalilähedane, mis annab julgust edaspidisteks korduskatseteks täiustatuma metoodika alusel. Tuleb siiski rõhutada, et metoodikamuutused ei tohiks kindlasti tähendada loobumist LLP-st, mille igihaljast teovõimet siinne kirjutis on järjekordselt ja veenvalt tõestanud.

Märkusi

[1] LLP ehk nn *Lull-Laputa Principle* on heuristiline metoodika, mille kaasaegseks rakenduslikuks tippväljundiks on hästituntud RC ehk nn *Rubik's Cube* ning mille üht esileküündivat ajaloolist varianti on detailselt kirjeldanud juba Lemuel Gulliver (1726: Part III, Chapter V):

The first professor I saw was in a very large room, with forty pupils about him. After salutation, observing me to look earnestly upon a frame, which took up the greatest part of both the length and breadth of the room, he said perhaps I might wonder to see him employed in a project for improving speculative knowledge by practical and mechanical operations. But the world would soon be sensible of its usefulness, and he flattered himself that a more noble exalted thought never sprang in any other man's head. Everyone knew how laborious the usual method is of attaining to arts and sciences; whereas by his contrivance the most ignorant person at a reasonable charge, and with a

little bodily labor, may write books in philosophy, poetry, politics, law, mathematics, and theology, without the least assistance from genius or study. He then led me to the frame, about the sides whereof all his pupils stood in ranks. It was twenty feet square, placed in the middle of the room. The superficies was composed of several bits of wood, about the bigness of a die, but some larger than others. They were all linked together by slender wires. These bits of wood were covered on every square with paper pasted on them, and on these papers were written all the words of their language, in their several moods, tenses, and declensions, but without any order. The professor then desired me to observe, for he was going to set his engine at work. The pupils at his command took each of them hold of an iron handle, whereof there were forty fixed round the edges of the frame, and giving them a sudden turn, the whole disposition of the words was entirely changed. He then commanded thirty-six of the lads to read the several lines softly as they appeared upon the frame; and where they found three or four words together that might make part of a sentence, they dictated to the four remaining boys who were scribes. This work was repeated three or four times, and at every turn the engine was so contrived that the words shifted into new places, as the square bits of wood moved upside down.

[2] Muidugi on ülekohtune ja amoraalne püüda Feyaertsi niimoodi pilada, sest tegelikult peab ta silmas oma "possessioonide" mõtestumist nimelt millegi füüsilise omamisena eelkõige samas plaanis, milles meie siin vaatame oma elusid ja hingi. Omamissuhted nagu ruumis viibimise suhtedki võimaldavad siirdeid, mis on täiesti analoogilised: saamine vastab latiivsusele, omamine lokatiivsusele ja minetamine separatsioonile. Ja ta näidete hulgas on mitte ainult head asjad *luck*, *Verstand*, *mind*, vaid ka pahad *troubles* ja *Dummheit*.

Kujutleme, et jutu all on nt sellised mõisted, nagu õnn, mõistus ja mure(d). Kuivõrd selgesti need vm asjad tunduvad meile "omandina", sõltub lause konkreetsest metafoorsest disainist.

Põhimõtteliselt tuleks metafoor POSSESSIBLE OBJECTS lugeda tuvastatavaks ka siis, kui "omamise" märk on negatiivne: kellegi kohta öeldakse, et tal *pole* õnne ~ mõistust ~ muresid, kuid selline metafoorsus on puhtteoreetiline ja vaevalt tuleks kellelegi pähe, et see, mis niimoodi välja näeb, võiks olla mingi "metafoorse teekonna" bukvaalseks mandunud ots. Igatahes aga on nende situatsioonide aksioloogiline märk üheselt selge – õnne ja mõistuse puudumine on paha, murede puudumine hea. Neist "omanditest" separeerumist võib kujutleda nii "meelega" (agentiivselt, intentsionaalselt) kui ka "kogemata" (kogevalt, spontaanselt) toimuvana, kuid terve mõistus korreleerib asjad vaikimisi nii, et

headest asjadest separeerumist kujutletakse tahtmatuna, halbadest asjadest aga tahtlikuna. Separatiivne protseduur, mille kaudu vähemetafoorsesse staatikasse jõuti, võib aga olla vormistatud nudimate või eredamate leksikaalsete vahendite abil. Kui nt õnnest *ilma jäädakse*, mõistus *minetatakse*, muredest *vabanetakse*, on toimuva aksioloogiline kvaliteet endiselt selge, kuid pole veel eriti selge, kui agentiivne või passiivne on possessor ühe või teise verbi korral, ning öeldava kuulumine just füüsilisse maailma on endiselt problemaatiline. Kui mured *minema heidetakse*, mõistus *hüljatakse*, õnn *ära kaotatakse*, on need juba selgemini tajutavad füüsiliste toimingutena (st. metafooridena); kaks esimest mõtestuvad intentsionaalsetena, viimane spontaanse, "kogematasena", kuid mõistuse tahtlikku hülgamist on ülalmainitud korrelatsiooni tõttu üsna raske motiveeritult ette kujutada. Kui mured *aknast välja visatakse*, mõistus nt *garderoobi unustatakse* (Feyaertsi näites ongi saksa fraseologism *Er hat den Verstand in der Garderobe abgegeben*) või õnn *taskust välja pillatakse*, on materiaalne maailm kõikjal esindatud otse füüsiliste "asjade", st substantiivsete komponentide kaudu ning ka aksioloogilised ja intentsionaalsed faktorid on omavahel harmoonias: head asjad minetatakse kogemata, pahad eemaldatakse tahtlikult. Kui öeldakse vastupidi (nt *astub oma õnnele meelega peale*; vrd ka kaugemaid *otsib ise endale otsa*; *kaevab ise endale hauda* vmt), siis mõtestatakse neid ütlusi harilikult parastuse või irooniana. Esimese näite juures võivad meenuda "Kalevipoja" (Kp: laul XIX, 139–140) õrnakoorelised õnnemunad. Näidet iseendale hauakaevamise kohta on kognitivistlikus vaimus analüüsinud kõige esmalt vist Seana Coulson (1997: 238–240), edasi on sellest saanud üks musternäiteid nn kontseptuaalse segu e. blendi illustreerimiseks (vt nt Fauconnier 1997: 168–171; Fauconnier, Turner 1998a: 277, 278, 280; Fauconnier, Turner 1998b: 149–151; Ruiz de Mendoza Ibáñes 1998: 269–273; Krikmann, ilmumisel).

Kuid omaja ja omatava staatilisest vahekorrast (st. "omandi" olemasolust või puudumisest) väljumist võib kujutada mitte üksnes niimoodi, et kogejast saab aste-astmelt ilmselge agent, vaid ka hoopis teistsuguse protsessina, kus seni omanditena kujutletud asjad ise muutuvad aste-astmelt agentideks, st personifitseeruvad üha selgemini ja selgemini. Näiteks mured *tekivad* → *tulevad* → *tulevad uksest ja aknast* → *kargavad nelja jalaga selga*; mõistus *tuleb tagasi* → *tuleb koju* → *kargab jälle pähe krapsti*; õnn *tuleb* → *tuleb ~ pöördub tagasi su juurde* → *kohtab sind* → *naeratab sulle* → *võtab sind oma embusse* jne. Ja imelikul moel ütleb meie vaist meile, et mõistus saab ainult tagasi tulla, mure ja õnn seevastu võivad enne olla olnud või olemata. Samal viisil personifitseerudes võivad nad ka lahkuda: head asjad võivad lihtsalt

minna su juurest ära, jätta sind maha, sulle "adjöö" öelda või mida iganes, halvad asjad võivad samuti *tüdineda* sind vaevamast või su peale lihtsalt *halastada* jne. Igatahes tunduvad halbade asjade personifikatsioonid palju "tegelikumatena" kui heade asjade omad – tihti mingite zoo- või demonomorfsetena kujuteldavate olendite rünnete või ligitikkumistena.

Kõiki kolme on võimalik väljendada veel suure hulga muudegi metafooride kaudu. Veekogu, mere, lainetega seonduv kujundiklaster on üks selliseid: *lolluse ookean* öeldakse olevat ääretu; võidakse *ujuda* või *hõljuda õnnelainetel* või *õnnemeres*; võidakse *uppuda murelainetes* (*muremeri* aga, kuigi alliteratiivne, tundub miskipärast kohmaka ja kohatuna, aga lihtsalt *muredesse* või *muresse* võib jälle *uppuda* päris keelepäraselt). Lainete juurest on kerge jõuda tormide-tuulte jt meteoroloogiliste metafooride juurde jne jne.

Ehk lühidalt, kus on siin selged domeenivastavused, kus on korralikeks paarikuteks kokkuköidetud või selgeid voolusänge järgivad kontseptuaalsed metafoorid? Mida oskab taolises kontinuatiivses plasmas peale hakata vaene Invariantsi Printsiip?

[3] Loomametafoorid iseendast on muidugi ülihuvitav ja -lai uurimisala väga mitmes plaanis, näiteks:

1) milliseid "mentaalseid stereotüüpe" loomadele üldse on omistatud ja omistatakse eri kultuurides ja eri ajastuil: laial Euraasia alal nt on loomajutus kaval tegelane identifitseeritud rebaseks, Aafrika folkloorides aga täidab seda rolli tüüpiliselt just jänes;
2) mismoodi loomad karakteriseeruvad erinevais "poeetilistes žanrides" – nt lastekirjanduses, folkloorseis lauludes ja narratiivides, retoorikas ja fraseoloogias: meenutagem kas või valvenäidet, et *kana* ja *hani* on tänapäeval üheselt pejoratiivid rumala naisolendi kohta, eesti regilaulus seevastu võisid nad olla ema või naise aksioloogiliselt positiivsed poeetilised sünonüümid, st hellitusnimed;
3) kas on põhjust õigeks lugeda Lakoffi ja Turneri (1989: 193–197) oletust, et loomametafoorid on tegelikult "topeltmetafoorid", mis saadud sel teel, et inimene omistas kõigepealt teatud loomadele teatud inimlikud karakteristikud (kavaluse, lolluse, kurjuse, ülluse, vapruse, salakavaluse vm) ja seejärel "põrgatas" need loomadelt uuesti iseendale tagasi;
4) millised on totaalsete (sententsiaalsete) ja personifitseeriv-allegoorilise loomametafooride kontseptuaalsed vahekorrad, mille kohta leidub rohkelt empiirilist teavet nt vanasõnades – ja milline on personifikatsiooni kui "mandunud animismi" kvaliteet ja eriasend retoorikas üldse.

[4] Maksiimid kipuvad, nagu teada, üksteisesse sulama. Võib küsida näiteks, kas komponent, mis kuulub Lakoffil ja Turneril (1989: 172 jj) nende nn Suure Ahela Metafoori koostisse, on nimelt kvantiteedi- või relevantsimaksiim. Anna Papafragou (1996: 177) on metonüümia kognitivistlikku valvenäidet *The ham sandwich is getting restless* analüüsides arvanud, et Grice oleks näinud siin kvaliteedimaksiimi rikkumist. Samahästi võiks selles näha ka kvantiteedimaksiimi rikkumist: kui infot oleks antud piisavalt, poleks põhjust rääkijat valetamises süüdistada tekkinudki.

[5] Kui *alla* asemel oleks *kaela*, siis oleks tähendusnihe muidugi juba üpris suur. Aktantide konfiguratsioon vastaks küll juhule, mille kohta me parajasti näiteid toome, aga *elu* asemel oleks maandumiskohaks *kael*, st konteiner oleks kadunud ja midagi uut tekkinud. Ma ei oska sõnastada selle uue olukorra geneerilist skematiseeringut korrektses kognitivistlikus vaimus, igatahes peaks ta hõlmama mitmesuguseid pahu asju, muresid, probleeme või ülesandeid, soovimatuid külalisi vm inimesi, mis/kes võivad kellelegi *kaela sadada* või *kaelas olla* ning mida tahetakse *kaelast ära saada* ja vahel harva saadaksegi. Kui nad *on* kaelas, siis meie meel ei täpsusta eriti, kuidas nad just kaelas asetsevad, aga see täpsustus võidakse ka teha – nad pannakse nt kaela *rippuma*, ning siis näeb olukord projektsioonide tasandil jälle kaunis erinev välja sõltuvalt sellest, kas rippujaks on mõni tegemata või vastik ülesanne, tasumata võlg, probleem vm "asjaks" substantiveeritud olukord ja/või hingeseisund, või siis on selleks mõni inimolend. Kust domeenist siin kognitiivne inspiratsioon või motivatsioon on täpselt tulnud, on võimatu öelda. Võib-olla taanduvad kaelasrippuvad probleemid, kohustused jmt VANGIKS OLEKU skeemile ja pärinevad aegadest, kus vangidele tavatseti kaela ja jäsemetele riputada mitmesuguseid liikumist pärssivaid raskusi — vrd nt *See ripub mul kaelas nagu (tina)pomm*. Kui pommi asemel oleks kivi, võiks see vihjata pärinemisele hoopis teisest, UPUTAMISE skeemist. Kui rippujaks on inimolend, siis minu keeletaju järgi kõige salientsem skeem, mis siin võiks heiastuda, oleks KALLISTAMINE, mis toob kaaasa oletuse, et rippuja on naissoost ja kaela omanik meessoost ning nad mõlemad eeldatavasti noored. Muud sotsiaalsed stambid – vanurid, põetatavad, tüütud võlanõudjad jts – näivad selle põhitähenduse eest kaugele taanduvat. (Kas taolisi hämaroletusi tuleks nimetada presupositsioonideks, inferentsideks, mingit sorti pragmaatilisteks "lokutsioonideks" vm moel, seda mina küll ei tea.)

[6] Kui *mao* asemel oleks nt *harksaba*, oleks kogu teadmustaust vastupidine: harksaba põuepugemine on olnud varasemas ruraalses olustikus nii

tavaline ja ohutu sündmus ja harksaba üldse nii väheütlev putukas, et ei figureeri pea üldse metafoorides ega fraseologismides. Eesti dokumenteeritud kõnekäänuaines puudub harkasaba sootuks ja mul on ka omaenda virumaisest lapsepõlvest ainus ähmane mälestus, et *harksaba* tähendas poissi, poisslast ja vastandus *lehtsabale*, mis tähendas tüdrukut.

Kirjandus

Aisenman, R. A. 1999. Structure-mapping and the simile–metaphor preference. – Metaphor and Symbol 14:1, 45–51.

Arutjunova, N. D. 1979. Языковая метафора (Синтаксис и лексика). Лингвистика и поэтика. Ответственный редактор В. П. Григорьев. Москва: Наука. 147–173.

Bredin, H. 1998. Comparisons and similes. – Lingua 105:1/2, 67–78.

Coulson, S. 1997. Semantic Leaps: The Role of Frame-Shifting and Conceptual Blending in Meaning Construction. Unpublished P.H. dissertation. University of California, San Diego. http://hci.ucsd.edu/cogsci/grad_pubs/coulson_thesis.pdf

EKG I 1995 = Erelt, M.; Kasik, R.; Metslang, H.; Rajandi, H.; Ross, K.; Saari, H.; Tael, K.; Vare, S. Eesti keele grammatika I: Morfoloogia. Sõnamoodustus. Tallinn: ETA Eesti Keele Instituut.

Fass, D. 1997. Processing Metonymy and Metaphor. Greenwich, London: Ablex Publishing Corporation.

Fauconnier, G. 1997. Mappings in Thought and Language. Cambridge: Cambridge University Press.

Fauconnier, G.; Turner, M. 1998a. Principles of conceptual integration. – Discourse and Cognition: Bridging the Gap. Ed. by Jean-Pierre Koenig. Stanford, California: Center for the Study of Language and Information. 269–283.

Fauconnier, G.; Turner, M. 1998b. Conceptual integration networks. – Cognitive Science 22:2, 133–187.

Feyaerts, K. 2000. Refining the inheritance hypothesis: Interaction between metaphoric and metonymic hierarchies. – Metaphor and Metonymy at the Crossroads: A Cognitive Perspective. Ed. by Antonio Barcelona. Berlin, New York: Mouton de Gruyter. 59–78.

Grady, J. E. 1997. THEORIES ARE BUILDINGS revisited. – Cognitive Linguistics 8:4, 267–290.

Grady, J. E. 1999. A typology of motivation for conceptual metaphor: Correlation vs. resemblance. – Metaphor in Cognitive Linguistics: Selected Papers from the Fifth International Cognitive Linguistics

Conference, Amsterdam, July 1997. Ed. by R. W. Gibbs, Jr., G. J. Steen. Amsterdam, Philadelphia: Benjamins. 79–100.

Grady, J. E.; Taub, S.; Morgan, P. 1996. Primitive and compound metaphors. – Conceptual Structure, Discourse and Language. Ed. by A. E. Goldberg. Stanford, California: CSLI Publications. 177–187.

Gulliver, L. 1726. Travels into Several Remote Nations of the World. In Four Parts. London: Benj. Motte.

Indurkhya, B. 1992. Metaphor and Cognition: An Interactionist Approach. Dordrecht, Boston, London: Kluwer.

Kennedy, J. M.; Chiappe, D. L. 1999. What makes a metaphor stronger than a simile? – Metaphor and Symbol 14:1, 63–69.

Kp = Kreutzwald, Fr. R. Kalevipoeg: Tekstikriitiline väljaanne. I–II. Tallinn: Eesti Riiklik Kirjastus, 1961 ja 1963.

Krikmann, A. 1997. Üksteisesse sahteldatud konteinerid. – Eesti filoloogia poolsajand Teaduste Akadeemias. Toim. J. Viikberg. Tallinn: Eesti Keele Instituut. 365–366.

Krikmann, A., ilmumisel. Вклад современной теории метафоры в паремиологию. [Vilnius, 2001?]

Lakoff, G.; Turner, M. 1989. More than Cool Reason: A Field Guide to Poetic Metaphor. Chicago, London: The University of Chicago Press.

Nogales, P. D. 1999. Metaphorically Speaking. Stanford, California: CSLI.

Ortony, A. 1979. The role of similarity in similes and metaphors. – Metaphor and Thought. Ed. by A. Ortony. Cambridge, London, New York, Melbourne: Cambridge University Press. 186–201.

Papafragou, A. 1996. On metonymy. – Lingua 99:4, 169–196.

Ruiz de Mendoza Ibáñes, F. J. 1998. On the nature of blending as a cognitive phenomenon. – Journal of Pragmatics 30:3, 259–274.

Õim, H. 1971. Isikuga seotud sõnarühmade semantiline struktuur eesti keeles. – Keele modelleerimise probleeme 4. TRÜ toimetised 278. Tartu: TRÜ.

Õim, H. 1974. Semantika. Tallinn: Valgus.

Õim, H. 1983a. Inimene, keel ja arvuti ehk kompuuterlingvistika. Tallinn: Valgus.

Õim, H. 1983b. Семантика и теория понимания языка: Анализ лексики и текстов директивного общения эстонского языка. Doktoritöö. Tartu: TRÜ.

Õim, H. 1997. Eesti keele mentaalse maailmapildi allikaid ja piirjooni. – Pühendusteos Huno Rätsepale. Toim. M. Erelt, M. Sedrik, E. Uuspõld. Tartu: Tartu Ülikool. 255–268.

Personaalsufiksite alguskomponent (*)-*n*- uurali keeltes

Ago Künnap
Tartu ülikool

Soovides tervitada meie head juubilari, eesti keeleteoreetilise mõtte ja keeletehnoloogia julget novaatorit, oma kolleegi ja sõpra professor Haldur Õimu tema tähtpäeval, pean paratamatult jääma oma liistude juurde ja kõnelema ajaloolisest uralistikast. Kuid püüan seda teha samuti võimalikult novaatorlikult.

Üksainus pilk uurali keeltele veenab meid selles, et pole põhjust rääkida personaalsufiksite alguskomponendist (*)-*n*- kui ühisuurali nähtusest. Samuti mitte sellest, et ta oleks eri keelte ja murrete possessiivsufiksite alguses olnud ühtselt kas ainult omaja või ainult omatava arvutunnuseks. Mõlemad väited, kes, kus ja millal poleks neid ka esitanud, on ekslikud. Aga vaatame keeleandmeid veidi lähemalt. Kõigepealt nendiksin, et ei ole loogiline siduda personaalsufiksite alguskomponenti (*)-*n*- personaalsufiksitele järgeva (*)-*n*-ainesega, mis esineb näit. soome keele 3. isiku possessiivsufiksi murdelises kasutuses *-nsa-n*, *-nsä-n* (Collinder 1960: 302). Uurali keeled on küll põhiliselt aglutinatiivset tüüpi, kuid see ei tähenda mingit klotsimängu, kus klotse võib vabalt ümber paigutada, nagu paljud uralistid – eelkõige uurali keelte varasema staadiumi osas – näivad arvavat. Samuti pole põhjust ühendada kõnealust üksikkonsonandilist komponenti (*)-*n*- possessiivsufiksitele eelneva vokaalilõpulise arvutunnusega (*)-*n*V-, mis esineb mõnedes uurali keeltes (nagu pole alust siduda üksikonsonandilist genitiivisufiksit (*)-*n* vokaalilõpulise lokatiivisufiksiga (*)-*n*V).

Selline vokaalilõpuline pluuralisufiks on possessiivse deklinatsiooni korral olemas permi keeltes, näit. sürjakomi *tšerę-j* 'minu kirves' : *tšer-ni-m* 'meie kirves' (Osnovy 1974: 225–226, kus viimane vorm on analüüsitud küll kujul *tšer-n-im*), obiugri keeltes, näit. mansi *hāpu-mt* 'minu paadis' : *hāpu-nu-mt* 'minu paatides' (Collinder 1960: 300), samuti samojeedi keeltest selkupi keeles, kuigi ainult M. A. Castréni kirjapanekuis, näit. *loga-m* 'minu rebane' : *loga-i-m* ~ *loga-ji-m* ~ *loga-ni-m* 'minu rebased' (lk 299). Mainitagu, et mitmed uurijad on selkupi keele korral oletanud siin arengut

-ni- > *-ji-*, Castrén ise aga arengut *-ji-* > *-ni-* (vt. Künnap 1971: 42, vrd. ka 49–50). Kuigi areng *-ni-* > *-ji-* on üldiselt ootuspärasem, ei välistaks ma Castréni vastupidist oletust, sest on ju selkupi keeles tõenäoliselt aset leidnud ka areng **j* > *l'* (vt. Lehtisalo 1936: 48–55; Katz 1979), kuigi vastupidine areng oleks taas üldiselt ootuspärasem. Lisaksin veel, et permi keelte uurijate poolt laialt kasutatav sufiksile eelneva vokaali lugemine sufiksi koosseisu, näit. *tšer-n-i̯m* 'meie kirves' (Osnovy 1974: 226; vt. ka Osnovy 1976: 149–154, 172–193), on minu meelest põhjendamatu.

Eraldi tähelepanu nõuab mõnede uurali keelte ainsuse 2. isiku erandliku (*)*n*-elemendilise pöördelõpu probleem. Juha Janhunen rekonstrueerib protouurali ainsuse 2. isiku pöördelõpu kujul **-n ~ -t* ja jätkab: "An important point of dissimilarity between the systems of possessive suffixes and verbal personal endings exhibit a duality in the suffix consonant. Most of the present-day U[ralic] languages point to an original dental stop **-t*, while in the Eastern periphery (Komi, Ob-Ugric, Samoyed) the dental nasal **-n* is met. The nasal variant of the suffix obviously implies the previous existence of a 2. person pronoun with initial nasal, although only uncertain traces of the pronoun stem itself have otherwise been preserved (in Ob-Ugric only)." (Janhunen 1982: 34–35).

Péter Hajdú pühendas sellele probleemile oma plenaarettekande VI Rahvusvahelisel Fennougristide Kongressil Sõktõvkaris 1985. Ettekanne näitab, kui palju peavalu on 2. pöörde *n*-sufiks valmistanud uralistidele läbi aegade. Hajdú ise jõuab järeldusele, et "in der Grundsprache nur ein einziges Dentalexplosivlaut wurde gegentlich ... in den östlichen Dialekten der Grundsprache ... zum entsprechenden homorganen Nasalverschlusslaut in der Position vor dem nächstfolgenden Nasal assimiliert, wodurch zwei Allomorphe des Pronomens entstanden: **t*V*n* > **n*V*n* ... Letzteres (**n*V*n*) ist nur in den obugrischen Sprachen als selbstständiges Pronomen übriggeblieben, seine Spuren sind aber in allen samojedischen (und spärlich sogar in den permischen) Sprachen in der Personalbezeichnung nachweisbar, ebenso wie die Reflexe des Pronomens mit anlautenden **t-* mit entgegengesetzter Frequentz (im Ostjakischen und Permischen)." (Hajdú 1986: 6). Nii näeb Hajdú ausalt suurt vaeva, et jõuda – ettenähtud raamides – näivalt loogilise tulemuseni. Kuid milleks näeb ta seda vaeva? Nähtavasti sellepärast, et püüab uurali üldist *t*-personaalsufiksit ning komi, obiugri ja samojeedi

n-peronaalsufiksit tagasi viia ühele ühisele protouurali lähtele. See ei ole kerge ülesanne. Ja ammugi mitte kohustuslik.

Samojeedi keelte osas on Janhunen hiljaaegu konstateerinud: "From the proto-Uralic point of view, one of the most interesting features is that the second-person singular predicative ending seems to have been **n* in proto-Samoyedic, as opposed to **t* in most subbranches of Finno-Ugric. The simple shape **n* is, however, preserved only in in Nganasan, while the other Samoyedic languages have **n-t-ø*, possibly as a result of the influence of the corresponding possessive suffixes." (Janhunen 1998: 471).

Udmurdi keeles on ainukeseks 2. pöörde *n*-liseks verbivormiks eituspartikkel *en* imperatiivi singularis ja pluuralis, näit. *en nu* 'ära kanna!' ja *en ve·ralę* 'ärge kõnelge!' (Osnovy 1976: 180–181; Hajdú 1986: 2). Siin pole aga sugugi automaatselt selge, et oleks tegemist eitusverbiga *e-*, mille lõpus on komi verbiparadigmaga ühine pöördelõpp *-n*. Nimelt tuleb meil arvestada sellega, mida on udmurdi eituspartikli *en* kohta kirjutanud M. R. Fedotov. Viimane on toonud esile mitmeid asjaolusid, mis asetavad sellise automaatse järelduse kahtluse alla. Turgi keeltest on tšuvaši keel ainuke, kus imperatiivi 2. pööre on vormistatud partikli *an* abil, mis on aga tagasiviidav varasemale kujule **en*. On võimalik oletada (ja ongi oletatud), et tšuvaši *an* kujutab endast laenu udmurdi ja komi naaberkeeltest. Nende kolme naaberkeele vastastikused mõjusuunad on ka laiemas mõttes kõike muud kui selged. Fedotovil on õigus, kui ta nendib, et erinevalt teistest uurali keeltest on permi keelte jaatava kõne imperatiiv moodustatud turgi keeltega ühiste mallide järgi, st kasutades ainult verbitüve, näit. udmurdi *mi̯n* 'tule!', komi *mun* id., vrd. tšuvaši *kaj* 'mine ära!', tatari ja baškiiri *kit* id. Lisaks leidub vanatürgi keeles eituspartikkel *aŋ* 'ei', vrd. ka tunguusi-mandžu *ân* id. ja isegi korea keeles *an*(*i*) id (Fedotov 1980: 50–54). Seega pole imperatiivi moodustamise kontekstis välistatud tšuvaši eituspartikli **en* laenamine permi keeltesse, eriti kui arvestada udmurdi imperatiivi eituspartikli erandlikkust.

Seda udmurdi imperatiivi eituspartikli erandlikkust toob veelgi paremini esile asjaolu, et "Uralisches Etymologisches Wörterbuchis" on kõnealune eituspartikkel esitatud tagapoolse vokalismiga: *ę-*, *ö-*, *ü-*. Ainult Yrjö Wichmann on fikseerinud selle osaliselt *e*-vokaalsena (vt Rédei 1986–1991: 68). Kuidas kommenteerida sellist vokalismi? Kui Wichmann poleks esitanud *e*-list varianti, võinuksime täiesti

kindlalt öelda, et tagapoolne vokalism ei toeta partikli *en* udmurdi algupära. Olen toonud mõneti teises kontekstis esile, et "a part of of the Permic and Volgaic negation words with a velar vowel (first of all, those with *a*) drops out as original nouns ... and another part becomes questionabel as it could be made up of generalizations of these nouns (at least in the case of a velar stem vowel and *a* in particular)." (Künnap 1998: 85). Partikli *en* seletamist võimaliku ainsa või paremini säilinud oletatava protouurali eitusverbi **e-* jäljena udmurdi keeles ei saa võtta tõsiselt. Erandlikule udmurdi partiklile *en* leidub muidugi rohkesti vasteid komi keele eitusverbi *e-* vormide näol, komi imperatiivis sealhulgas udmurdi partikliga täpselt samas funktsioonis (vt. taas Rédei 1986–1991: 68). Kuid kooskõlas minu arusaamadega uurali keelte uurimisest ei muuda see udmurdi evidentsi. Ehk teiste sõnadega: mingi "tagasirekonstruktsioon" oletatavast protopermist udmurdi keelde on minu silmis täielik mõttetus. Kõik see teeb tõenäoliseks, et vaatlusalune udmurdi *en* on laenatud tšuvaši keelest.

Samojeedi keelte korral esineb nganassaani verbiparadigmas – eelkõige subjektilise ja refleksiivse konjugatsiooni indikatiivis – 2. pöörde lõpp *-ŋ* (vt nt Tereščenko 1979: 185). Eugen Helimski on pidanud võimalikuks sama lõpu esinemist – kuigi vaid ühelainsal juhul – ka matori keeles. Ta kirjutab: "Es ist nicht unmöglich, daß P[roto-]S[amojedisch] **-ŋ* (VxSg2 in der subjektiven Konjugation) als mat. *-ŋ* in dem folgenden Beispiel belegt ist: ... *i-* 'ich nicht' (Negationsverb): MM ***iüngchónda*** 'non dormio' ('du schläfst nicht' anstatt 'ich schlafe nicht'). Selbstverständlich ist es nicht die einzige Möglichkeit, die problematische Form ***iüng-*** zu interpretieren." (Helimski 1997: 166). Suulises vestluses minuga on Helimski lisanud, et ta on rekonstrueerinud vastava protosamojeedi võimaliku pöördelõpu *-ŋ* ainuüksi nganassaani ja matori andmete põhjal. Olen esitanud vaatlusaluse matori sõnapaari ühes oma varasemas artiklis koos mõnede teiste matori sõnapaaridega, kus eitava abiverbi lõpus on singulari 1. pöörde lõpp *-m*, näit. *ingümsia* 'non dormio' ja *igimchóndunschuk* id. Olen nende sõnapaaride kohta öelnud: "Was die Formen *igim-*, *ingüm-*, *iüng* angeht, so erinnern sie an die matorische Form *Igam-*: (Spasskij) *Igamdanem* 'ich weiß nicht', die von J. Janhunen mit einem Fragezeichen so rekonstruiert wird: **i-ŋå-m-*" (Künnap 1983: 382–383). Teise võimalusena on Helimski

tõlgendanud elementi *-ng-* vormis *iüng-* matori preesensi *ŋ*-sufiksina (Helimski 1997: 159).

Peab aga silmas pidama, et Piret Klesment on osutanud: matori *i*-tüvede sufiksi korral pole üheselt selge, kas konsonantainese *ng* häälduseks on *ŋg* või *ŋ* (Klesment 1995: 98). Lisaks esineb matori eitava abiverbi *i*-tüves vokaalaines *ü*. Kui oletada, et matori vormi *iüng-* lõpus on 2. pöörde lõpp *-ng-*, võib eeldada, et ajatunnus eelneb pöördelõpule nagu eespool toodud matori eitava abiverbi 1. pöörde vormides. Tuleb arvestada ka võimalusega, et matori *i*-tüvele on algselt järgnenud preesensi (või aoristi) sufiks *-j*V-. Klesmendi analüüsi põhjal on viimase sufiksi kujuks tavaliselt *-ja-* (lk. 96–97), kuigi üksikjuhtudel võivad esineda ka teised vokaalid, seejuures ühel juhul *ü*: *tschorgoiüm* 'stare' (Helimski 1997: 159). Seega on ootuspärasem, et matori *ŋ*(*g*)V või *g*V oleks preesensi (või aoristi) sufiks nagu ka nganassaani keeles *-ŋi-*, näit. *ni-ŋi-ŋ hoδətəˀ* 'sa ei kirjuta' (Tereščenko 1979: 261). Helimski on nähtavasti pidanud võimalikuks ka ajatunnuse *-ŋ-* esinemist matori vormis *iüng-* tüve *i*-järel, pakkudes tõlgendust "Präs. *-ŋ-* (?) + Sg2 (?)" (Helimski 1997: 251). Asjatundmatu matori keele kirjapanija ei tarvitsenud nimelt üldse kuulda häälikut *ŋ*, mis järgnes *i*-le: **iŋüŋ*(*g*)*-*. Rohkem pole Helimski esitatud oletusele midagi lisada, kuid rõhutaksin, et matori keele vormi *iüng-* korral on tegemist nganassaani 2. pöörde lõpu *-ŋ* ainukese võimaliku etümoloogilise vastega kogu samojeedi keelterühma ulatuses. Tuletaksin ühtlasi meelde tuntud koomilist seika Castréni uurimisreisidelt, kui tema samojeedi keelejuht tõlkis lause 'minu naine on haige' oma keelde kujul 'sinu naine on haige' (ning keeldus seda muutmast, kuna tema naine olevat tema teada terve). Samal kombel võis ka lause 'mina magan' matori keelde tõlkides saada kuju 'sina magad'.

Turgi keelte üldiseks singulari 2. isiku possessiivsufiksiks on *-ŋ* ~ *-n*, mida kasutatakse piiratult ka verbi vastava isiku pöördelõpuna. Turkoloogid tavatsevad tuletada selle possessiivsufiksi 2. isiku personaalpronoomenist, vrd. näit. vanaturgi *sän* 'sina'. Selline tuletusviis on väga spekulatiivne ja omalt poolt arvan ma, et küllap *-n* ~ *-ŋ* on turgi keelte algne üksikonsonandiline 2. isiku possessiivsufiks. Pole kindel, kas turgi sufiksi kaksikkuju on arengu *-n* > *-ŋ* tulemus, kuid selline areng pole vähemalt välistatud. Olen hiljuti osutanud sellele asjaolule, et ka eskaleuudi keeled tunnevad 2. isiku

n- ja *ŋ*-possessiivsufiksit, kusjuures Knut Bergsland oletab arengut *n* < **-nt* (Künnap 1997: 99).

Helimski suulise teate põhjal olevat ka komi 2. isiku pöördelõpu *n* lähteks konsonantühend **nt*, sest teisiti ei saavat selles keeles – nagu udmurdiski – tulemuseks olla *n*. Selline lähe ei sobi tema järgi aga nganassaani vastavale personaalsufiksile *-ŋ*, kuna nganassaani konsonantühend **nt* on üldiselt hästi säilinud. Helimski meelest võib nganassaani possessiivsufiksi mõeldavaks lähteks olla küll *-n*, sest see keel tunneb arengut *-n* > *-ŋ*.

Uurali keelte ülemal vaadeldud (*)*n*-ainelistest ainsuse 2. isiku pöördelõppudest saame kokkuvõttes seega järgmise pildi:

Udmurdi	*-n* (?) (? < **-nt*)	Neenetsi	**-nt*
Komi	*-n* (? < **-nt*)	Eenetsi	**-nt*
Ungari	–	Nganassaani	*-ŋ* (? < **-n*)
Handi	*-n*	Selkupi	**-nt*
Mansi	*-n*	Kamassi	–
		Matori	*-ŋ* (?)

Permi *n*-sufiksi rekonstrueerimist kujul **-nt* ei pea ma siiski obligatoorseks, sest olulised grammatilised markerid võivad käituda ennustamatult. Heaks näiteks on põhjaeesti ainsuse 1. pöörde lõpu *-n* säilimine, kuigi üldiselt on sõnalõpuline **-n* siin kadunud. Obiugri keelte korral on kõnealust *n*-i seostatud seal säilinud *n*-algulise personaalpronoomeniga. Omalt poolt pean seda *n*-i siin algupäraseks lahus nimetatud personaalpronoomeni säilimise faktist. Neenetsi, eenetsi ja selkupi korral peab lähtuma **nt*-st. Jääb küll võimalus, et tegemist on siiski ühisuurali 2. isiku tähistajaga *-t*, mille ette on sugenenud homorgaanne nasaal, nagu seda võib samojeedi keeltes sageli näha. Sel juhul langeks neenetsi, eenetsi ja selkupi vaatlusalune singulari 2. isiku personaalsufiks meie vaatluse alt üldse välja. Ei ole põhimõtteliselt välistatud, et nganassaani ja matori *-ŋ* võib olla saadud **n*-ist ja kuuluda seega kokku teiste uurali keelte singulari 2. isiku *n*-iliste personaalsufiksitega (kuid see **n* pole nganassaani ja matori keeltes häälikulooliselt tagasiviidav **nt*-le nagu mõnede muude samojeedi keelte korral).

Omaette probleemiks on komponendi (*)*-n*- kasutamine rea uurali keelte possessiivsufiksite alguses. Nähtus on iseloomulik eelkõige läänemeresoome, lapi (saami), mordva ja samojeedi keeltele. Mari keele korral võib pidada vaatlusaluse komponendi jäljeks ainult

mitmuse 1. isiku possessiivsufiksi konsonantainest, näit. *kiša-m* 'minu jälg' : *kiša-na* 'meie jälg' (Osnovy 1974: 226) < ? **-n-m*V. Kõnealuse komponendi mõeldava kasutuse selline piirang mari keeles pole loomulikult iseenesestmõistetav. Mari 1. isiku possessiivsufiksit *-na* pole põhjust pidada mingiks jäljeks volga algkeelest – vrd. ersamordva *uma-m* 'minu koppel' : *uma-n* 'meie koppel' (Osnovy 1975: 297) –, sest eriti pärast Gjbor Bereczki põhjalikke uurimusi on volga algkeele oletus suikunud nagu mingisse unne, kuigi temaga mõnikord siiski opereeritakse. Mordva keelte korral kirjutas Mikko Korhonen: "Hyvin johdonmukaisesti pl[uraalin] *n* esiintyy erzämordvan possessiivisessa taivutuksessa, esim. s[in]g[ularin] ... 2. *ťšora-zo* 'yksi poikansa' *ťšora-n-zo* 'useat poikansa' ..." (Korhonen 1981: 234). Korhose näite taolisi juhtumeid võib mordva keeltest leida, kuid eelkõige vaid mõnedest ersa murretest, samal ajal kui teistes ersa murretes ja mokša keeles on asi kaugel järjekindlusest, nt. ersa *kudo-st* 'nende maja, nende majad' (Osnovy 1974: 275), mokša *uma-ts* 'tema koppel' : *uma-sna* 'nende koppel' (Osnovy 1975: 298).

Läänemeresoome-lapi ühise algkeele jaoks rekonstrueeris Korhonen sirgjoonelise possessiivsufiksite süsteemi, milles possessiivsufiksitele eelnev **-n-* tähistas omaja mitmust, näit. 3. isiku singular **-nsa* ~ **-nsä*, duaal **-nsan* ~ **-nsän*, pluural **-nsak* ~ **-nsäk* (Korhonen 1981: 235). Nimetatud algkeelt pole ilmselt kunagi eksisteerinud ning see pole ainuüksi minu kurikuulsast uurali keelerühma algkeelte vastasusest johtuv seisukoht, vaid samale tulemusele jõudsid hiljuti ka näiteks Terho Itkonen (1999) ja Jorma Koivulehto (1999). (Üldisemalt rääkides ei leia me läänemeresoome keeltest muide üldse mingeid jälgi duaalist, kuigi neid on püütud seal näha.) Kuidas on aga lood läänemeresoome keelte possessiivsufiksite alguskomponendi (*)-*n*- (kunagise) pluraalse funktsiooniga? Ma ei pea põhjendatuks läänemeresoome keelte kunagise ühtse algkeele oletust, nende keelte evidentsist endast ei leia me nimetatud pluraalsusele rohkem tuge, kui mõnedest üksikutest soome murdevormidest ja sedagi omatava pluurali väljendamise kohta, näit. Iitti *tupa-s* 'sinu üks tuba' : *tuva-n-s* 'sinu mitu tuba' (Korhonen 1981: 234).

Teatavasti pole *n*-elemendi esinemine possessiivsufiksite alguses soome keeles, eriti murdeti, kuigi järjekindel (vt Mark 1925: 104–242). Samas kasutatakse soome keeles possessiivsufikseid läänemeresoome keelte rühmas kõige sagedamini. Karjala, vepsa ja isuri keeles on nende tarvitamine tunduvalt piiratum, seda eriti mitmuse

possessiivsufiksite osas. Eesti, liivi ja vadja keeles peetakse nende kasutamist rudimentaarseks. (Vt Laanest 1975: 116.) Kui me otsime eesti keelest possessiivsufikseid või vähemalt nende jälgi, ei ole kuigi palju leida. Julius Mark on osutanud, et kõneldavas eesti keeles leiduvad vaid 3. isiku possessiivsufiksi jäljed ning ainult sellistes adverbiaalsetes väljendites, nagu *laiutasa* ja *iseäranis*, samuti refleksiivpronoomenis *enese* (Mark 1925: 56–64). Possessiivsufiksite jälgi eesti rahvalauludes on põhjalikumalt käsitlenud Juhan Peegel (1966; 1974). Julius Mägiste on olnud possessiivsufiksite jälgede esinemise osas eesti keeles ja eriti rahvalauludes väga ettevaatlikul seisukohal (Mägiste 2000). Tema arvates pole kaugeltki kõik oletatavad possessiivsufiksite jäljed saadud possessiivsufiksitest. Samuti vajab veel lähemat selgitamist, kas *n*-algulised oletatavad possessiivsufiksite jäljed rahvalauludes pole vähemalt osaliselt kõigest prosoodilised täitesilbid (Mägiste 2000: 154).

Minu arusaamist mööda on possessiivsufiksite esinemus eelkõige eesti ja liivi keeles selline, et on põhjust tõsiselt kahelda, kas need keeled on üldse kunagi kasutanud possessiivsufikseid sellises ulatuses nagu näiteks soome keel, st kas nad on tundnud possessiivsufiksite n.ö täisparadigmat. Soomepärane põhjaeesti rannikumurre võttis varem enda alla tunduvalt ulatuslikuma ala kui tänapäeval, samuti elas Eestis läbi aegade rohkesti sisserännanud soomlasi, seda kuni Lõuna-Eestini ja tõenäoliselt ka liivi keelealani. Nii sellest murdest kui ka sisserändajate soome keelest võis üle possessiivsufiksiteta eesti ja liivi keeleala levida mõningat, sealhulgas lähteallika seisukohast ekslikku kasutust, mis on iseloomulik eelkõige eesti rahvalaulule. Nimetatud võimalus on loomulikult vaid hüpotees. (Vt eesti possessiivsufiksite probleemi kohta veidi lähemalt Künnap 2001b.) Ka pole eesti ja liivi keele korral enamasti vähimatki evidentsi oletatava 3. isiku possessiivsufiksi *s*-alguliste jälgede ette **n*-elemendi rekonstrueerimiseks.

Lapi possessiivse deklinatsiooni sõnavormide mitmesugused foneetilised seigad annavad ilmselt tunnistust sellest, et seal on kasutatud possessiivsufikseid nii ilma *n*-alguskomponendita kui ka selle komponendiga (Korhonen 1981: 233–245). Kõnealuse komponendi mingist selgest arvutunnuselisusest näitu aga pole. Permi keeltes üksikonsonandilist **n*-elementi possessiivsufiksite ees ei kasutata (Osnovy 1976: 149–152). Samasugune on olukord ka ugri keeltes (lk 282–284, 307–311, 386–387). Iseküsimus on see, et obiugri 2. isiku

possessiivsufiksi kujuks on *-n* (nagu muide ka turgi keeltes). Kuid viimane on obiugri keeltes ka verbi 2. isiku pöördelõpuks, nagu eespool juba nägime, nii et tegemist on õieti obiugri 2. isiku väljendamisega üldse personaalsufiksi *-n* abil, ilma et personaalsufiksile oleks kunagi eelnenud mingi lisakomponent **-n-*. Tõsi küll, ka 1. ja 3. isiku possessiivsufiksi kujuks võib handi keeles murdeti olla mõnikord *-n*, kuid siin on põhjendatult oletatud mingit hilist (analoogia)arengut (vt nt Osnovy 1974: 231–232, 276; Osnovy 1976: 310; Honti 1986: 124–140).

Käsitledes varem kokkusurutult kõigi samojeedi keelte possessiivsufikseid (Künnap 1971: 156–182) ja pöördelõppe (Künnap 1978: 11–84), olen vaadelnud ka neile eelnevat komponenti (*)*-n-* ning püüdnud selgitada selle päritolu (Künnap 1971: 173–182; Künnap 1978: 46–50, 54–56). Olen pidanud põhjendatuks seda seisukohta, et samojeedi (ja ka soome-ugri) verbivormid pole olnud sageli muud kui verbaalnoomenite possessiivsufiksilised vormid. See seletab ka ära, miks nende keelte pöördelõpud on sageli identsuseni sarnased possessiivsufiksitega. Koos possessiivsufiksitega on kandunud samojeedi deklinatsioonisfäärist konjugatsioonisfääri vaatlusalune **n*-komponentki. Jõudsin järeldusele, et samojeedi (ja soome-ugri) obliikvakäänetes esineb see komponent possessiivsufiksite ees sagedamini kui nominatiivis, puududes aga samojeedi (*)*m*-akusatiivilõpu järel algselt ehk hoopiski (Künnap 1971: 174–179). Kõik need asjaolud kallutasid mind paljudest seletustest eelistama üht: uurali keelte possessiivsufiksitele eelnev (*)*n*-komponent on saadud genitiivilõpust (*)*-n* (lk. 180–182). Sellel seisukohal olen ka praegu.

Nüüd, kus ma olen jõudnud veendumusele, et samojeedi keeled on tekkinud sel teel, et varem tõenäoliselt mingit paleosiberi keelekuju kõnelnud mongoliidsed samojeedid läksid üle europiididest soomeugrilaste läänemeresoome(-lapi) tüüpi keelekujule (vt. lähemalt Künnap 2001a), tõlgendaksin ma possessiivsufiksitele eelneva (*)*n*-komponendi esinemuse piirdumist, nagu olen püüdnud eespool näidata, praktiliselt läänemeresoome, lapi, mordva ja samojeedi keeltega nende keelte erandliku ühisjoonena, mis on samojeedidel saadud (koos terve hulga muude joontega) läänemeresoome(-lapi) tüüpi keelest sellele üleminekul.

Pärast kõige eelneva kirjutamist on mul olnud võimalus tutvuda Ulla-Maija Kuloneni artikliga "Zum *n*-Element der zweiten Personen besonders im Obugrischen" (Kulonen 2001), mida ma pean

ajaloolise uralistika seisukohast julgelt novaatorlikuks. Kulonen ei usu, et ühisuurali possessiivsufiksid ja pöördelõpud oleksid saadud vastavate personaalpronoominte aglutineerumise teel eelnevate noomeni- ja verbitüvedega, sest selleks on uurali personaalpronoomenid liiga ebaühtlased. Ta kirjutab, et "leksikaalsete ühikutena võivad personaalpronoomenid olla suhteliselt hilised innovatsioonid. [...] ... personaalafiksid ei tarvitse olla personaalpronoomenite tüvede aglutinatsiooni tulemus, vaid tegelik areng on võinud olla lausa vastassuunaline." (Kulonen 2001: 151).

Oma vaatlusaluse artikli alapeatükis "Sind die Personalendungen Ergebnis von Agglutination?" Kulonen selgitab: "Oben wurde bereits die Frage aufgeworfen, wie relevant die Hypothese ist, daß die Personalendungen und Possessivsuffixe der einzelnen uralischen Sprachen durch Agglutination aus den Personalpronomina entstanden seien, da sich bei näherer Betrachtung bald zeigt, daß die Endungen in verschiedenen Sprachen ein einheitlicheres Bild bieten und damit älter zu sein scheinen als die Personalprononomina, die zahlreiche lautliche Varianten und Unregelmäßigkeiten aufweisen." (Kulonen 2001: 168).

Kulonen läheb veelgi kaugemale: "Die Partikelstämme waren nach traditioneller Auffassung immer Verhältniswörter, wie es aus die Stämme der Postpositionen in den heutigen finnisch-ugrischen Sprachen sind. [...] Rédei (z.B. in UEW) verbindet sie [= mansi lokaalkaasuste lõpud – A.K.] den pluralischen Demonstrativpronomina der finnisch-ugrischen Sprachen. Diese Auffassung ist semantisch und funktional schwer zu begründen. Da alle nachweislich aus Postpositionen hervorgegangen Kasusendungen auf Partikel- oder Verhältnisvorstämmen auf bauen, ist es äußerst unwahrscheinlich, daß aus einem Demonstrativum, das als solches nur deiktische Bedeutung hat, eine Postposition entstehen können, die eine lokale Beziehung zum Ausdruck bringt." (Kulonen 2001: 169–170).

Omalt poolt eitan ma ühisuurali personaalsufiksite ja kaasussufiksite tekke Kuloneni poolt kirjeldatud traditsioonilist seletust (vt juba Künnap 1971; 1978; pöördelõppude kohta viimati Künnap 2001b). Kuloneni artiklis esitatud uudsed nüansid kinnitavad seda minu veendumust.

Kirjandus

Collinder, B. 1960. Comparative Grammar of the Uralic Languages. Uppsala.

Fedotov, M. R. 1980. Čuvašskij jazyk v sem'e altajskih jazykov. Čeboksary.

Hajdú, P. 1986. Personalbezeichnungen für die 2. Person im Uralischen. – SFU XXI: 1–8.

Helimski, E. 1997. Die matorische Sprache. Wörterverzeichnis – Grundzüge der Grammatik – Sprachgeschichte. Unter Mitarbeit von Beáta Nagy, Szeged (= Studia uralo-altaica 41).

Honti, L. 1986. Chrestomatia Ostiacica. Budapest.

Itkonen, T. 1999. Zur Herkunft der ostseefinnischen Sprachen und des Lappischen. – Sprachen in Finnland und Estland. Toim. P. Lehtimäki. Wiesbaden. 1–6.

Janhunen, J. 1982. On the structure of Proto-Uralic. – FUF 44: 23–42.

Janhunen, J. 1998. Samoyedic. – The Uralic Languages. Ed. by D. Abondolo. London and New York. 457–479.

Katz, H. 1979. Beitrag zur Lösung des Problems der Entwicklung von ursam. *j im Selkupischen und der hiemit zusammenhängenden Fragen der historischen Morphologie dieser Sprache und des Uralischen. – SFU XV: 168–176.

Klesment, P. 1995. Monosuffixal finite verb forms in Mator. – Minor Uralic Languages: Grammar and Lexis. Ed. by A. Künnap. Tartu, Groningen. 94–100.

Koivulehto, J. 1999. Das Verhältnis des Ostseefinnischen und des Lappischen im Lichte der alten Lehnwörter: Die Susbstitution des fremden Wortausgangs *-CVz im Lappischen. – Sprachen in Finnland und Estland. Toim. P. Lehtimäki. Wiesbaden. 7–22.

Korhonen, M. 1981. Johdatus lapin kielen historiaan. Helsinki.

Kulonen, U.-M. 2001. Zum n-Element der zweiten Personen besonders im Obugrischen. – FUF 56: 151.

Künnap, A. 1971. System und Ursprung der kamassischen Flexionssuffixe I. Numeruszeichen un Nominalflexion, Helsinki (= MSFOu 147).

Künnap, A. 1978. System und Ursprung der kamassischen Flexionssuffixe II. Verbalflexion und Verbalnomina, Helsinki (= MSFOu 164).

Künnap, A. 1983. Über die Hintergrund einiger samojedischer Negationsformen. – NyK 85: 381–385.

Künnap, A. 1996. On the origin of Uralic Languages. – Ünnepi könyv Domokos Péter tiszteletére. Budapest. 142–144.

Künnap, A. 1997. Über einige sich ähnelnde uralische, eskimoische und tschuktschische Suffixe. – LU XXXIII: 97–101.
Künnap, A. 1998. On the Uralic *ś-preterite and *k-present. – LU XXXIV: 81–86.
Künnap, A. 2001a. Keneltä samojedit ovat oppineet kielensä? – Keele kannul. Tartu Ülikooli eesti keele õppetooli toimetised 17. Toim. R. Kasik. Tartu. 177–185.
Künnap, A. 2001b. Mitmuse tunnus *-k? – Keel ja Kirjandus 6, 427.
Laanest, A. 1975. Sissejuhatus läänemeresoome keeltesse. Tallinn.
Lehtisalo, T. 1936. Über die primären ururalischen ableitungssuffixe, Helsinki (= MSFOu 72).
Mark, J. 1925. Die Possessivsuffixe in den uralischen Sprachen I, Helsinki (= MSFOU 54).
Mägiste, J. 2000. Possessiivsufiksite rudimentidest eestis, eriti vana eesti kirjakeele (1520–1739) adverbides jm. partiklites. – Julius Mägiste 100. Tartu Ülikooli eesti keele õppetooli toimetised 15. Tartu. 11–160.
Osnovy 1974 = Osnovy finno-ugorskogo jazykoznanija (voprosy proishoždenija i razvitija finno-ugorskih jazykov). Moskva.
Osnovy 1975 = Osnovy finno-ugorskogo jazykoznanija. Pribaltijsko-finskie, saamskij i mordovskie jazyki. Moskva.
Osnovy 1976 = Osnovy finno-ugorskogo jazykoznanija. Marijskij, permskie i ugorskie jazyki. Moskva.
Peegel, J. 1966. Eesti regivärsilise rahvalaulu keelest. 2. trükk. Tartu.
Peegel, J. 1974. Eesti regivärsilise rahvalaulu keelest. – Eesti rahvalaulud. Antoloogia 4. Toim. Ü. Tedre. Tallinn. 45–76.
Pusztay, J. 1995. Diskussionsbeiträge zur Grundsprachenforschung. (Beispiel: das Protouralische), Wiesbaden (= Veröffentlichungen der Societas Uralo-Altaica 43).
Rédei, K. 1986–1991. Uralisches Etymologisches Wörterbuch. Budapest.
Tereščenko, N. M. 1979. Nganasanskij jazyk. Leningrad.

Lühendid

FUF = Finnisch-Ugrische Forschungen. Helsinki.
LU = Linguistica Uralica XXVI– Tallinn 1990– (continuation of SFU).
MSFOu = Mémoires de la Société Finno-Ougrienne. Helsinki.
NyK = Nyelvtudományi Közlemények. Budapest.
SFU = Sovetskoe finno-ugrovedenie I–XXV. Tallinn 1965–1989.
UEW = Rédei 1986–1991.

Kõneleja-spetsiifiliste tunnuste otsingul[1]

Einar Meister
TTÜ Küberneetika Instituut

1. Sissejuhatus

Inimkõrv on üks hämmastavalt täiuslik organ – oleme võimelised tajuma helisid ja hääli, lokaliseerima signaaliallika suunda ja asukohta, klassifitseerima erinevaid akustilisi signaale, tuvastama liigikaaslasi ja nende sugu, vastu võtma ja dekodeerima inimkõnet, eristama tuttavaid ja võõraid ning identifitseerima kõneleja isikut. Loomulikult ei toimu kõik need protsessid mitte ainult kõrvas, vaid eelkõige peaaju vastavates piirkondades erinevate tunnuste ja tajumehhanismide alusel. Kuigi kõne- ja kõnelejatuvastus on oma olemuselt erinevad, lähtutakse mõlemal juhul ühest ja samast signaalist, s.o kõneleja poolt produtseeritud kõnesignaalist. Kõneleja eesmärgiks on edastada kuulajale eelkõige lingvistilist informatsiooni – sõnu ja lauseid, mis kuulaja arvates parimal viisil väljendavad tema mõtteid või sõnumit. Kuid paratamatult kodeeritakse kõnesignaali lisaks lingvistilisele informatsioonile ka kõneleja isikut kirjeldavat informatsiooni. Artikuleeritud kõne puhul on võimatu lahutada lingvistilist informatsiooni tema akustilisest vormist, ka sünteeskõne puhul saame rääkida "isikupärasest" häälest. Seega on kõne- ja kõnelejatuvastus teineteisega tihedalt seotud ja baseeruvad suures osas samadel akustilistel tunnustel, mis kannavad nii lingvistilist kui ka kõneleja-spetsiifilist informatsiooni. Kui kõnetuvastuse puhul on oluline eristada tunnustes sisalduv lingvistiline informatsioon ja kõnelejast tingitud tunnuste variatiivsus – kui müra – maha suruda, siis kõnelejatuvastuse puhul on viimane just isiku tuvastust võimaldavaks informatsiooniks.

Inimtaju on võimeline juba üksiku sõna põhjal otsustama, kas kõnelejaks on mees, naine või laps, kas tegu on tuttava või võõra isikuga, samuti tuvastama kõneleja isikut. Samade ülesannete realiseerimine arvutil on aga vägagi keerukas. Automaatsete kõnelejatuvastussüsteemide loomisel on viimasel aastakümnel saavutatud häid tulemusi eelkõige tänu kiirele tehnoloogilisele arengule, samas

[1] Töö on teostatud Eesti Teadusfondi grandi nr 4154 ja Soome Tehnoloogiaagentuuri TEKES grandi nr 40285/00 toetusel.

pole kaugeltki selge tajumehhanismide ja kõneleja eripära kandvate tunnuste olemus.

Käesoleva artikli eesmärgiks on käsitleda kõnelejatuvastuse ülesandeid, meetodeid ja rakendusi ning esitada käimasoleva uurimisprojekti raames eesti keele foneetilisel andmebaasil teostatud akustiliste tunnuste analüüsi esialgseid tulemusi.

2. Mis on kõnelejatuvastus?

Kõnelejatuvastus üldise mõistena tähendab kõneleja isiku kindlakstegemist tema kõnehääle alusel. See on eri teadusvaldkondi integreeriv uurimissuund, mis ühelt poolt kuulub keeleteaduse, täpsemalt idiolekti foneetika valdkonda, teiselt poolt on tegu biomeetrika valdkonnaga. On ju kõnehääl üks inimese biomeetrilistest tunnustest sõrmejälje, silmaiirise, näo- ja kõrvageomeetria jm. kõrval.

Eristatakse kolme põhilist ülesannet:

Eristamine (ingl. *discrimination*) – võrreldakse kahte kõnenäidet ja selgitatakse, kas kõnenäidetest leitavad tunnused on nii lähedased, et need võiksid kuuluda ühele ja samale isikule või on tegemist kahe erineva isiku kõnega.

Identifitseerimine (ingl. *identification*) – võrreldakse otsitava isiku kõnenäidet N isiku kõnenäidetega ja otsustatakse, millisele kõnelejale N kõneleja hulgast antud kõnenäide kuulub, st otsitakse kõige lähemat mudelit. Identifitseerimisülesanne jaguneb omakorda kaheks alamülesandeks: kinnise hulga (ingl. *closed set*) ja avatud hulga (ingl. *open set*) ülesandeks. Kinnise hulga korral on teada, et otsitav isik on kindlasti nimetatud N isiku hulgas, avatud hulga puhul seda ei teata.

Verifitseerimine (ingl. *verification*) – otsustatakse, kas end registreeritud kõnelejana esitleva isiku kõnenäide aktsepteerida või tagasi lükata, st võrreldakse registreeritud isikuna esineva kõneleja kõnenäidet tema poolt varem salvestatud mudeliga; kui need on lubatud vea piires kokkulangevad, tunnistatakse kõneleja verifitseerituks.

Põhiline vahe identifitseerimis- ja verifitseerimisülesannete vahel on võimalike alternatiivide arvus – identifitseerimise puhul on alternatiivide arv võrdne registreeritud kõnelejate arvuga; verifitseerimise puhul on tegemist ainult binaarse otsustusega, sõltumata registreeritud kõnelejate arvust. Seega on vea tõenäosus kõneleja identifitseerimisel oluliselt suurem kui verifitseerimisel.

Kõnelejatuvastuse ülesannete lahendamisel on võimalikud järgmised variandid:

- korrektne tuvastus – leitakse isik, kellele uuritav kõnenäide kuulub;
- korrektne eitus – leitakse, et uuritav kõnenäide ei kuulu kahtlusalusele isikule;
- vale tuvastus – uuritav kõnenäide tunnistatakse kuuluvaks valele isikule;
- vale eitus – otsitavale isikule kuuluv kõnenäide tunnistatakse temale mittekuuluvaks.

3. Kõnelejatuvastuse rakendusvaldkonnad

3.1. Fonograafiline ekspertiis

Maailma kohtupraktikas on juba aastakümneid rakendatud kõnesalvestusi asitõenditena kuriteo sooritanud isikute tuvastamiseks. Ka Eestis on kõnesalvestuste ekspertiis saamas tavapäraseks kohtueelse uurimise meetodiks. Viimastel aastatel on uurimisorganid mitmetel juhtudel pöördunud TTÜ Küberneetika Instituudi foneetika ja kõnetehnoloogia labori poole kõnesalvestuste fonograafiliseks ekspertiisiks. Kuna mitmesugused helisalvestusseadmed on laialdaselt kättesaadavad, siis pole haruldased juhtumid, kus telefoni teel teostatud väljapressimised ja ähvardused on salvestatud kannataja poolt telefoniautomaatvastaja kassetile või nelja-silma kohtumistel toimunud kõnelused on registreeritud diktofoni abil. Politsei tehnilised võimalused jälitustegevuseks on järjest paranenud ja olulise informatsiooni salvestamine ka väga keerulistes tingimustes on võimalik, samuti salvestatakse Päästeameti häirekeskuse numbrile tulnud kõned. Pankades salvestatakse telefonipanga teenuseid kasutavate klientide ja pangatöötajate vahelised dialoogid, paljudes firmades on paigaldatud kaasaegsed turvasüsteemid, mille abil registreeritakse nii pilt kui ka heli.

Fonograafiline ekspertiis hõlmab erinevaid alaliike, millest olulisemad on järgmised:

- kõneleja isiku tuvastamine,
- salvestuse stenogrammi koostamine,
- kõnelejate arvu, soo, vanuse, tervisliku ja emotsionaalse seisundi määramine,
- salvestuse koha ja aja määramine, helistaja asukoha määramine,
- salvestuses esinevate müradе iseloomu määramine ja müradest pu-

hastamine,

- salvestuse autentsuse hindamine: kas tegemist on originaalsalvestuse või koopiaga, kas salvestus on terviklik, st teostatud ühe pideva sessioonina, või koosneb see mitmest eri aegadel salvestatud lõikudest; kas salvestust on elektroonselt või mehaaniliselt töödeldud või kas see on kokku monteeritud.

Kõnesalvestuste ekspertiisi korral on tavaliselt tegemist kas kõneleja identifitseerimis- või eristamisülesandega, mille lahendab vastava kvalifikatsiooniga ekspert. Isiku tuvastamiseks peab eksperdi käsutuses olema vähemalt kaks kõnenäidet: üheks kõnenäiteks on tavaliselt kuritöösituatsiooni salvestus, teiseks on kuritöös kahtlustatava(te) isiku(te) kõnesalvestus(ed), mis on teostatud uurija poolt ülekuulamise või võrdlusmaterjali võtmise käigus. Sageli teeb ekspertiisi keerukaks asjaolu, et võrreldavad kõnenäited on salvestatud erinevates akustilistes tingimustes ja kõneleja emotsionaalne seisund (seega ka tema kõne) on kuritöö toimepanemise ja võrdlusmaterjali võtmisel ajal oluliselt erinev.

Ekspertiisi teostamiseks viiakse uuritavad signaalid analoog-digitaalmuunduri abil arvutisse ja seejärel rakendatakse auditiivset ja/või akustilis-foneetilist analüüsimeetodit.

Auditiivne meetod põhineb eksperdi oskusel jälgida ja analüüsida kõnesignaalis esinevaid prosoodilisi, segmentaalseid ja spektraalseid (tämbrilisi) erinevusi, samuti sõnavara kasutust, lauseehitust ning dialoogi struktuuri. Prosoodilistest tunnustest on olulisemad kõnetempo ja -rütm, kõnemeloodika, rõhkude ja pauside asetus kõnes. Paljudel inimestel on selgelt tajutavad erinevused kõnetämbris, näiteks hääle kähedus, nasaalsus jm. Kõnedefektide olemasolu on samuti kindlaks tunnuseks kõnelejate eristamisel. Kõneleja isikupära väljendub hästi ka sõnavaras, eriti nn täitesõnade või häälitsuste (*mhmm, mmm, aaa, noh, on ju, eks* jne) süstemaatilises kasutamises.

Auditiivse analüüsi teostamisel kasutatakse signaaliredaktorit, mis võimaldab eksperdil valida uuritavatest signaalidest huvipakkuvaid kõnelõike ja neid sobivas järjekorras kuuldavaks teha.

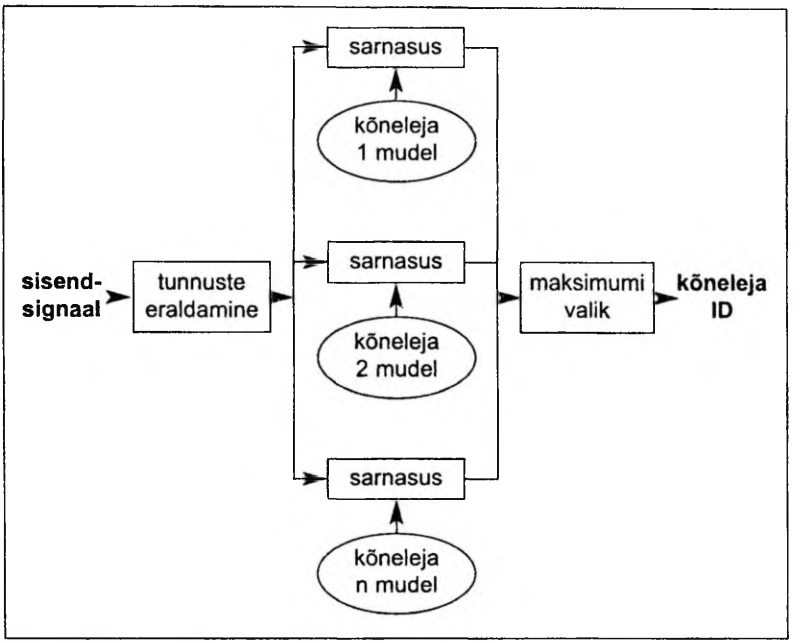
Kuigi auditiivse tuvastusega on alati seotud teatud subjektiivsus (eksperdi teadmised ja kogemused, suhe uuritavate kõnelejatega), on see siiski fonograafilise ekspertiisi kõige olulisem ja usaldusväärsem meetod. Auditiivne meetod on kasutatav ka küllalt kõrge müratasemе puhul, sest inimtaju, erinevalt signaalitöötlussüsteemidest, on võimeline käsitlema kõnet ja müra erinevate signaalidena.

Akustilis-foneetilise meetodi puhul kasutatakse abivahendina mingit signaalitöötlussüsteemi, mis võimaldab kõnesignaale mitmekülgselt analüüsida. Juba 1940. aastate lõpul hakati rakendama **spektrogrammide visuaalset võrdlust**, mis osutus võimalikuks tänu spektraalanalüüsi meetodite ja vahendite väljatöötamisele Bell'i laboritcs USAs. Oli lootus, et hääle spektrogrammi ehk nn häälejälje (ingl. *voiceprint*) näol saadakse sõrmejäljega võrdväärne isikut kirjeldav informatsioon. Kuid vaatamata intensiivsetele uuringutele 1960.–70. aastatel, jäid tulemused siiski ebarahuldavateks. Spektrogrammide "lugemine" ja võrdlemine on suurt kogemust ja aega nõudev protseduur ning seetõttu praktikas kaasajal harva kasutatav. Spektrogrammide võrdluse asemel on leidnud kindla koha erinevate signaalitöötlusmeetodite abil kõnest leitavate tunnuste kvantitatiivne võrdlus. Enamkasutatavad tunnused on pikaajaline spekter, vokaalide ja nasaalide spektrid, keskmine põhitooni sagedus, põhitooni diapasoon jm. Ekspert peab igal konkreetsel juhul otsustama, milliseid tunnuste analüüsimeetodeid kasutada ja kuidas interpreteerida saadud tulemusi. Selleks peab ekspert omama teadmisi erinevate tunnuste muutumispiirkondadest ja tunnuseid mõjutavatest faktoritest.

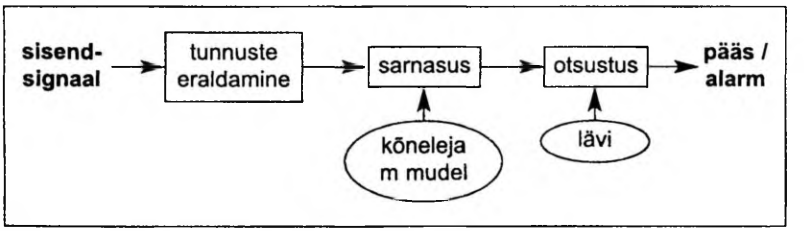
Eksperdi töö hõlbustamiseks ja subjektiivsete vigade vältimiseks on mitmel pool maailmas välja töötatud poolautomaatseid kõnelejatuvastussüsteeme, näiteks IdentiVox (Hispaania), IDEM (Itaalia), SIVE (Leedu), Dialekt ja Phonexi (Venemaa). Nimetatud süsteemid on kasutatavad, kui uurimiseks on esitatud piisavalt palju (minut või rohkem kõnet iga uuritava kohta) heakvaliteedilist kõnematerjali. Lõpliku otsuse langetab siiski ekspert, esitades selle tavaliselt tõenäosusliku hinnanguna (vähe tõenäone, tõenäone, väga tõenäone), kategoorilised otsused (tegemist on kindlasti ühe ja sama isikuga, tegemist on kindlasti erinevate isikutega) on võimalikud ainult üksikutel juhtudel.

3.2. Automaatsed tuvastussüsteemid

Automaatsete süsteemide (joonis 1–2) puhul tuvastatakse kõneleja inimeksperdi abita süsteemi treenimisel loodud kõnelejamudelite ja tundmatu isiku kõnest leitavate tunnuste võrdlemise teel. Kasutatakse erinevate signaalitöötlusalgoritmide abil leitud tunnuseid, näiteks pika- ja lühiajalist spektrit, spektri töötlemisel saadavaid parameetreid ja mitmeid kõnetaju mudelitest lähtuvaid tunnuseid.



Joonis 1. Kõneleja identifitseerimine



Joonis 2. Kõneleja verifitseerimine

Tunnuste klassifitseerimisel rakendatakse põhiliselt Markovi varjatud mudeleid (ingl. *Hidden Markov Models*), erinevaid neuronvõrgu tüüpe ja vektor-kvantimise meetodeid. Automaatse süsteemi rakendamisele eelneb selle treenimine suure hulga erinevate kõnelejate kõnenäidetega. Treenitud süsteem on võimeline korrektset otsust langetama juba küllalt lühikese (1–3 sekundit) kõnenäite põhjal.

Automaatset kõnelejatuvastust on otstarbekas kasutada pääsu reguleerimise vahendina strateegilistel objektidel, infosüsteemides, samuti näiteks pangaoperatsioonide teostamisel. Sel juhul on tegemist verifitseerimisülesandega, kus lisaks personaalse parooli või PIN-

koodi sisestamisele kasutatakse isiku kindlakstegemiseks ka tema kõnenäite võrdlemist varem salvestatud kõnenäitega.

Automaatsed kõnelejatuvastussüsteemid jagatakse tekstist sõltuvateks ja sõltumatuteks. Tekstist sõltuvate süsteemid puhul rakendatakse nii treeningul kui ka tuvastusprotsessis üht ja sama teksti – kindlaid paroollauseid või teatud hulka sõnu, mida kasutaja võib valikuliselt ette lugeda. Tekstist sõltumatute süsteemide puhul ei oma konkreetne tekst süsteemi treenimisel ja kasutamisel tähtsust.

Mõlemat liiki süsteemid on suhteliselt lihtsalt rünnatavad, kuna aktsepteerituks võib osutuda ka registreeritud isiku poolt magnetofonilindile salvestatud kõnet või paroole ettemängiv isik. Seetõttu on oluliselt turvalisem rakendada süsteemi poolt etteantava juhusliku tekstiga tuvastussüsteemi, kus lisaks kõneleja tuvastamisele kontrollitakse ka kõneleja poolt loetud teksti vastavust süsteemi poolt etteantud tekstile.

Võrreldes teiste biomeetriliste tunnustega nagu sõrmejälg, näo profiil jt, on kõnehääle eeliseks tema kasutatavus ka telefoniteenuse puhul.

4. Tunnuste olemus

Kõnelejatuvastuse aluseks on kõneleja orgaanilised, õpitud ja hetkeolukorrast tingitud tunnused.

1. Orgaanilised tunnused baseeruvad kõnelejate kõneorganite – kopsud, häälekurrud, suuõõs, ninaõõs, jm – füsioloogilistel erinevustel, samuti kuuluvad siia rühma kõneleja sugu, vanus, kehakaal, pikkus, kõne- ning kuulmishäired.
2. Õpitud tunnused lähtuvad keele õppimise ja kasutuse käigus omandatud kõneproduktsioonimallidest, kõneleja sotsiaalsest ja haridusliikust taustast, keelekeskkonnast (murdekeele ja võõrkeelte mõjutused).
3. Hetkeolukorrast tingitud tunnused kajastavad kõneleja tervislikku ja emotsionaalset seisundit, alkoholi, ravimite ja narkootiliste ainete mõjutusi kõneloomele jms.

Tuvastamisel kasutatavad tunnused peaksid (Nolan 1983):

- võimalikult palju varieeruma eri kõnelejatel;
- võimalikult vähe varieeruma samal kõnelejal;
- olema moonutamis- ja matkimiskindlad;
- olema sagedasti esinevad;

- olema mõõdetavad;
- olema signaalitöötlust taluvad.

Tunnuste töötlemisel tuleb tingimata arvesse võtta järgnevaid tunnuseid mõjutavaid tegureid:

- kasutatav seadmestik (mikrofon, sidekanal jms);
- salvestustingimused (ümbruskonna müratase);
- hääle moonutamine ja matkimine;
- sagedusriba piirangud;
- analüüsitehnika.

Kõnelejatuvastuseks sobivate tunnuste leidmiseks on teostatud hulgaliselt uuringuid, võimalike tunnustena on välja pakutud järgmisi:

- nasaalide spekter,
- nasaali ja vokaali siire,
- r-konsonandi formandid,
- l-konsonandi formandid,
- pikaajaline spekter,
- pikaajalise spektri üla- (>1000 Hz) ja alaosa (< 1000 Hz) amplituudi suhe,
- vokaalide formandid,
- põhitooni sageduse keskmine kõrgus ja hajuvus (Nolan 1983).

Automaatse tuvastuse puhul on osutunud efektiivseteks tunnusteks kepstri ja delta-kepstri koefitsiendid (Furui 1981; Furui 1994; Soong & Rosenberg 1988; Rosenberg *et al* 1994) ja kõnetaju mudelitest lähtuvaid tunnuseid, nagu näiteks MFCC (ingl. *mel-frequency cepstral coefficients*) (Carey, Parris 1992; Gish 1990; Openshaw *et al* 1993; Rose, Reynolds 1990), RASTA-PLP (ingl. *RelAtive SpecTrA – Perceptual Linear Prediction*) tunnused (Hermansky *et al* 1991; Koehler *et al* 1994), Seneff'i auditoorse mudeli (Seneff 1988) parameetrid (Zue *et al* 1989).

Kõnelejatuvastusel on tõsiseks probleemiks parameetrite ajalised variatsioonid. Need on tingitud nii kõnelejast endast, erinevatest signaali salvestus- ja edastustingimustest, kui ka taustmüra iseloomu ja taseme muutumisest. Kõneleja ei ole võimeline identselt kordama ühte ja sama lauset, ühe salvestussessiooni jooksul salvestatud laused on omavahel rohkem korreleeritud kui eri aegadel salvestatud laused, hääles toimuvad pikaajalised, vanusega seotud muutused.

Vähendamaks eri parameetrite variatsioonide mõju, rakendatakse kõnelejatuvastussüsteemides mitmeid tunnuste normaliseerimise meetodeid.

5. Kõnelejatuvastuse uuringud TTÜ Küberneetika Instituudis

TTÜ Küberneetika Instituudi foneetika ja kõnetehnoloogia laboris on kõnelejatuvastuse uuringuid teostatud alates 1994. aastast. Uuriti mitmete tunnuste (F0, spektraalsed tunnused) omadusi ja neuronvõrkude rakendamist kõnelejate tuvastamisel (Altosaar, Meister 1995; Meister 1998). Alates 2000. a. käivitus vastav Eesti Teadusfondi grant (No. 1454) ja koostöö Helsingi Ülikooli foneetikaosakonnaga (prof. A. Iivonen) eesmärgiga välja töötada automaatse kõnelejatuvastussüsteemi prototüüp ning kõnesalvestuste fonograafilise ekspertiisi metoodika.

5.1. Uurimismetoodika

Töö teostamisel kasutatakse uurimismaterjalina EU COPERNICUS-programmi raames loodud eesti keele foneetilist andmebaasi (Eek, Meister 1999), mis sisaldab 70 diktori kõnesalvestusi (35 naist, 35 meest, kokku ligi 12 tundi kõnematerjali). Lisaks sellele kasutatakse ka uurimisorganite operatiivsalvestusi ja telefoniteenuste (nt telefonipank) kasutamisel salvestatud dialooge.

Projekti käigus analüüsitakse süstemaatiliselt mitmeid spektraalseid ja prosoodilisi parameetreid; uuritakse erinevate tunnuste kõnelejaid eristavat võimet; leitakse tunnuste muutumispiirkonnad ning hinnatakse tunnuse kõnelejasisest ja kõnelejatevahelist variatiivsust; tunnuste klassifitseerimiseks rakendatakse erineva konfiguratsiooniga neuronvõrke. Uurimaks tunnuste kõnelejasisest ajalist variatiivsust, salvestatakse samu isikuid täiendavalt erinevate ajaintervallide järel.

Kõnesignaalide analüüsil kasutatakse ja võrreldakse erinevaid töökeskkondi (Computerized Speech Lab Model 4300, SoundScope, Praat, Matlab) ning tunnuste eraldamise algoritme.

5.2. Multiparameetriline kõnelejaprofiil

Kõneleja-spetsiifiline informatsioon on kodeeritud praktiliselt kõigisse tunnustesse, mida erinevate analüüsimeetodite abil on võimalik kõnesignaalist leida. Loomulikult ei ole kõik tunnused võrdse kaaluga, mõned tunnused kannavad endas rohkem informatsiooni kui teised. Kõneleja individuaalsust on otstarbekas esitada mitmetest nor-

maliseeritud akustilistest tunnustest koosneva multiparameetrilise kõnelejaprofiilina (Iivonen *et al* 2001).

Järgnevalt on esitatud eesti keele foneetilise andmebaasi salvestustel teostatud erinevate akustiliste parameetrite analüüsitulemusi, mida hiljem rakendatakse kõnelejaprofiili koostamisel.

5.2.1. Pikaajaline spekter

Pikaajaline spekter (ingl. *LTAS, Long-Term Average Spectrum*) saadakse mingi kõnelõigu, näiteks sõna või lause, piires leitud lühiajaliste spektrite keskmistamise teel. Sel juhul ei kirjelda saadud spekter üksiku hääliku spektraalseid omadusi, vaid on eelkõige kõneleja kõnetrakti iseloomustavaks tunnuseks. Mida pikem on lause ja mida rohkem erinevaid häälikuid selles sisaldub, seda paremini pikaajaline spekter antud kõnelejat kirjeldab.

Häid tulemusi pikaajaliste spektrite kasutamisel on saanud mitmed uurijad (Hollien, Majewski 1977; Meister 1998; Niemi 1999; jt).

Uurimaks pikaajalise spektri omadusi, on teostatud vastav analüüs 20 mees- ja 20 naiskõneleja salvestustest CSL[2]-keskkonnas. Pikaajalised spektrid on leitud iga kõneleja 16 lausest (iga lause kohta üks spekter) kolme erineva analüüsiakna korral (256, 512 ja 1024 punkti, vastavad spektrid on tähistatud edaspidi LTAS-256, LTAS-512 ning LTAS-1024).

Erinevatest lausetest leitud pikaajaliste spektrite sarnasust on hinnatud eukleidilise kauguse abil. Spektrite $x = \{x_i\}$ ja $y = \{y_i\}$ vaheline eukleidiline kaugus EK on arvutatud järgmiselt:

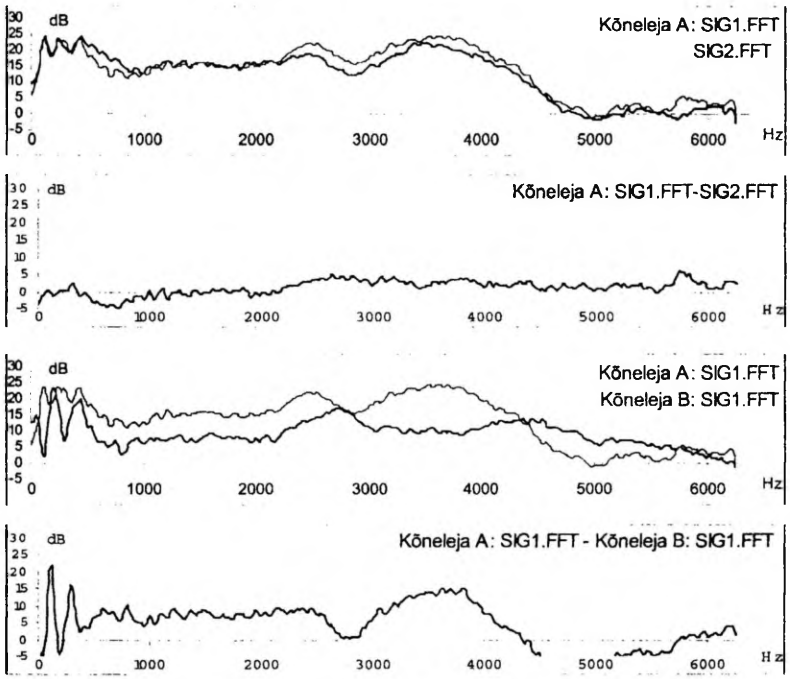
$$EK = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

kus n on spektri resolutsioon (n = 128 LTAS-256 korral, n = 256 LTAS-512 korral, n = 512 LTAS-1024 korral).

Kui ühe kõneleja erinevate lausete spektritevaheliste eukleidiliste kauguste väärtused ning jaotusfunktsioon kirjeldavad antud tunnuse kõnelejasisest (ingl. *intra-speaker*) variatiivsust, siis eri kõnelejate lausetest leitud spektrite vahelised kaugused ja nende jaotus-

---

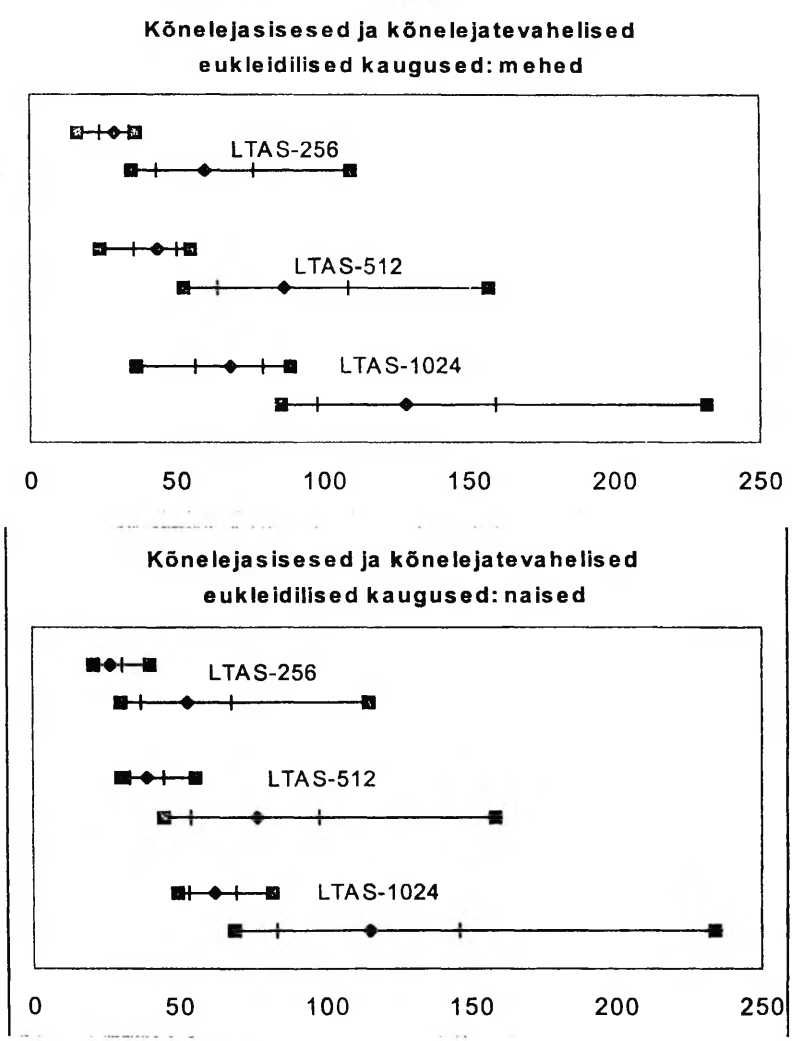
2 CSL – firma Kay Elemetrics Corp. (USA) kõneanalüüsisüsteem Compterized Speech Lab, Model 4300.



Joonis 3. Kahe kõneleja pikaajaliste spektrite võrdlus
Esimene aken: kõneleja A kahe eri lause pikaajalised spektrid;
teine aken: kõneleja A eri lausete pikaajaliste spektrite vahe;
kolmas aken: kõnelejate A ja B sama lause pikaajalised spektrid;
neljas aken: kõnelejate A ja B sama lause pikaajaliste spektrite vahe

Tabel 1. Mees- ja naiskõnelejate pikaajaliste spektrite eukleidiliste kauguste statistilised näitajad

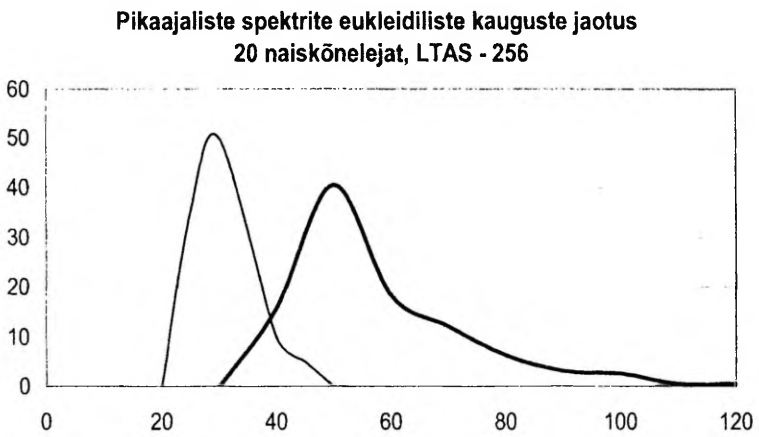
	Mehed			**Naised**		
LTAS tüüp	**256**	**512**	**1024**	**256**	**512**	**1024**
Keskmine kõnelejasisene EK	28,7	43,0	67,7	26,5	38,8	61,4
EK standardhälve	5,2	7,6	11,5	4,4	5,8	8,4
Minimaalne EK	16,3	24,0	36,6	20,5	30,5	49,2
Maksimaalne EK	36,4	54,6	88,7	40,1	55,8	82,1
Keskmine kõnelejatevaheline EK	59,8	86,3	128,4	52,5	76	114,7
EK standardhälve	16,8	22,3	30,6	15,5	21,6	31,2
Minimaalne EK	34,8	52,8	85,7	30,1	44,4	68,7
Maksimaalne EK	109,7	156,8	231,5	114,7	158,5	234,1



Joonis 2. Eukleidiliste kauguste keskväärtused (♦), standardhälbed (|) ja minimaalsed-maksimaalsed väärtused (■) erinevate spektri resolutsioonide korral

funktsioon kajastavad aga kõnelejatevahelist (ingl. *inter-speaker*) variatiivsust (vt joonis 3).

Joonisel 3 näitena esitatud spektritevaheliste eukleidiliste kauguste jaotusfunktsioonid on iseloomulikud enamikele tunnustele.



Joonis 3. Eukleidiliste kauguste jaotusfunktsioonid
Vasakpoolne kõver kirjeldab kõnelejasiseste ja parempoolne kõnelejatevaheliste kauguste jaotust

Saadud analüüsitulemused näitavad, et pikaajaline spekter kannab endas olulist kõnelejaspetsiifilist informatsiooni, kusjuures suurem spektri resolutsioon tagab parema eristusvõime.

5.2.2. Põhitooni sagedus

Kõne põhitoon tekib häälekurdude võnkumise tulemusena, mis omakorda käivitub kopsudest väljahingatava õhuvoolu toimel. Põhitooni sagedus sõltub võnkuvate häälekurdude pikkusest, massist ja pingsusest, samuti õhusurvest häälekurdude all. Põhitooni olemasolu või puudumine on heliliste/helitute häälikute vastanduse aluseks, selle pikemaajalist muutust tajub kuulaja kõne meloodiana (intonatsioonina).

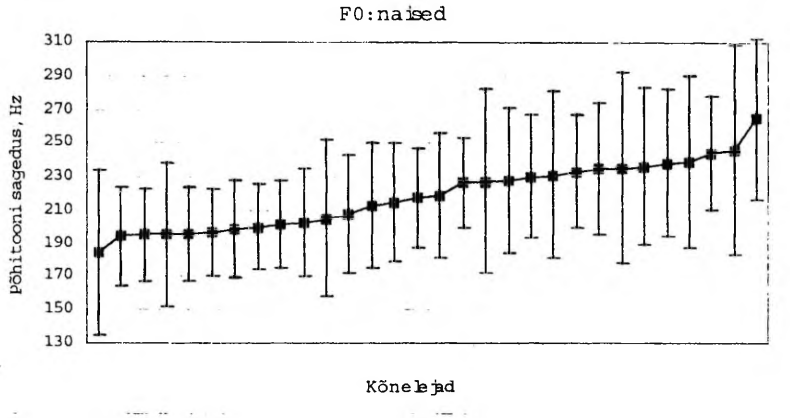
Teadaolevalt kõigub põhitooni sagedus meestel keskmiselt vahemikus 70 kuni 200 Hz ja naistel 150 kuni 400 Hz, kuid eri indiviididel võib see varieeruda ka suuremates piirides (Hess 1983).

Põhitooni analüüs on teostatud 30 mees- ja 30 naiskõneleja signaalidest, mis on salvestatud etteantud teksti (5-lauselised fraasid

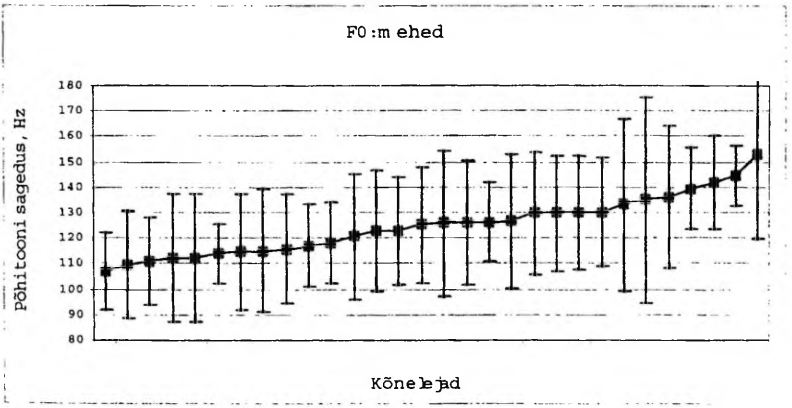
[3] PRAAT on Amsterdami Ülikooli foneetikaosakonna teadurite P. Boersma ja D. Weenink poolt loodud kõneanalüüsi keskkond, saadav Internetist vabavarana (http://www.praat.org).

Tabel 2. Analüüsitud kõnelejate keskmise põhitooni sageduse minimaalsed, keskmised ja maksimaalsed väärtused [Hz]

	F0 min	F0 keskmine	F0 max
Mehed	107	125	153
Naised	184	219	265



Joonis 4. 30 naiskõneleja keskmised põhitooni sagedused ja standardhälbed



Joonis 5. 30 meeskõneleja keskmised põhitooni sagedused ja standardhälbed

BABELi tekstikorpusest) lugemisel. Analüüsitava kõnelõigu pikkus on iga kõneleja puhul keskmiselt 28 sekundit, analüüs on teostatud PRAAT[2]-keskkonna vastava algoritmi abil. Kuna tegu on neutraalsete tekstide lugemisel saadud kõnenäidetega, siis peaks leitav põhitooni keskmine sagedus iseloomustama iga kõneleja keskmist, nn füsioloogilist põhitooni sagedust.

Joonistel 4 ja 5 on esitatud nais- ja meeskõnelejate keskmised põhitooni sagedused koos standardhälbega, keskmised statistilised näitajad on esitatud tabelis 2.

5.2.3. Muud tunnused kõnelejaprofiilis

Peale eelkäsitletud tunnuste on kavas analüüsida mitmeid teisi tunnuseid ja hinnata nende sobivust kõnelejate eristamiseks.

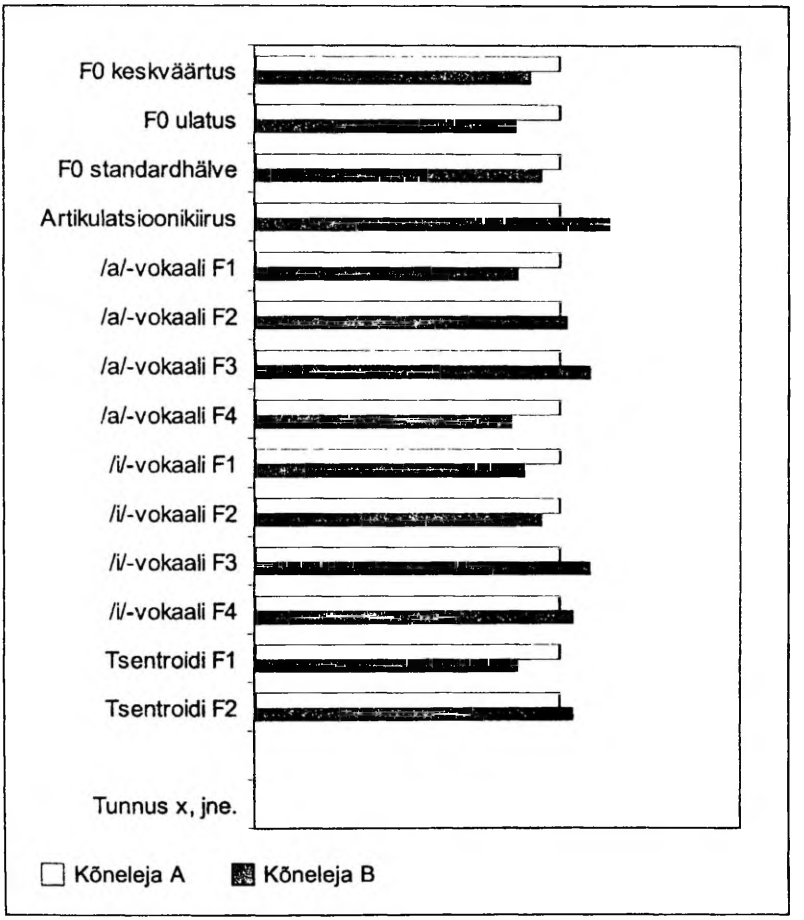
Näiteks:

- artikulatsioonikiirus – keskmine silpide arv sekundis;
- vokaalide lühiajalised spektrid – eestikeelses kõnes sagedamini esinevate vokaalide (/a/, /i/, /e/, /u/) spektrite võrdlus, kui vokaalid asuvad samas häälikuümbruses;
- tsentroidsagedused – kõnes esinevate vokaalide formantsageduste keskväärtus F1/F2 koordinaadistikus. Kuigi tsentroidsageduste väärtused sõltuvad otseselt vokaalide esinemissagedusest, saavutavad tsentroidid teatud stabiilse väärtuse pikema kõnelõigu korral ja kirjeldavad sellisena kõneleja vokaaltrakti omadusi;
- "jitter" – põhitooni perioodi fluktuatsiooni kirjeldav parameeter;
- "shimmer" – põhitooni amplituudi fluktuatsiooni kirjeldav parameeter;
- "HNR" (ingl. *Harmonic-to-Noise Ratio*) – vokaali harmoonilise ja müralise komponendi suhe, iseloomustab hääle kähedust;
- jm.

5.2.4. Kõnelejaprofiili graafiline esitus

Kõnelejaprofiili ideed selgitab kõige paremini selle graafiline esitus (vt joonis 6).

Kõneleja A kõigi tunnuste väärtused normaliseeritakse ja neile kõigile antakse sama väärtus, kõneleja B tunnused saavad kõneleja A vastavate tunnuste suhtes positiivseid või negatiivseid väärtusi. Otsuse langetamiseks, kas tegu on ühe ja sama või kahe erineva kõnelejaga, on vajalik eelnevalt teada, millised on erinevate kõnele-



Joonis 6. Kahe kõneleja virtuaalsete kõnelejaprofiilide võrdlus

japrofiili moodustavate tunnuste kaalud, tunnuste kõnelejasiseste ja kõnelejatevaheliste variatsioonide piirid, tunnuste usaldusväärsus.

Erinevate tunnuste leidmiseks on vajalik rakendada erineva keerukusega algoritme – seega on igal tunnusel hind, mida saab mõõta tunnuse leidmiseks teostatud arvutuste hulgaga. Oluline on valida iga tuvastusülesande tarvis piisav sobilike tunnuste hulk, mis optimaalse arvutusvõimsuse korral tagab soovitud tulemuse.

6. Kokkuvõte

Käesolevas töös esitatud tulemused pikaajalise spektri ja põhitooni sageduse kohta kinnitavad nende tunnuste kõneleja-spetsiifilist iseloomu, tunnuste statistilised näitajad annavad hea ettekujutuse kõnelejasisestest ja kõnelejatevahelistest piiridest. Lisaks teiste potentsiaalsete kõnelejaprofiili komponentide analüüsile on kavas välja töötada meetodid kõnelejaprofiili automaatseks moodustamiseks ning tuvastamiseks, oluline on ka erinevate profiilide sobivuse testimine erinevate tuvastusülesannete korral.

On raske prognoosida, kas ja millal käimasoleva uurimistöö tulemused leiavad rakendust näiteks pangaautomaatides kliendi identifitseerimiseks tema kõnelejaprofiili järgi, kuid fonograafilise ekspertiisi teostamisel rakendatakse seniseid tulemusi juba täna.

Kirjandus

Altosaar, T.; Meister, E. 1995. Speaker recognition experiments in Estonian using multi-layer feed-forward neural nets. – Proceedings of Eurospeech'95, Vol. 1. 333–337.

Carey, M. J.; Parris, E. S. 1992. Speaker Verification Using Connected Words. – Proceedings of Institute of Acoustics, Vol. 14. 95–100.

Eek, A.; Meister, E. 1999. Estonian speech in the BABEL multi-language database: Phonetic-phonological problems revealed in the text corpus. – Proceedings of LP'98. Vol II. Ed. by O. Fujimura *et al.* Prague: The Karolinum Press. 529–546.

Furui, S. 1981. Cepstral Analysis Technique for Automatic Speaker Verification. – IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. ASSP-29:2, 254–272.

Furui, S. 1994. An overview of speaker recognition technology. – Proceedings of ESCA Workshop on Automatic Speaker Recognition Identification and Verification. Martigny, Switzerland, April 5–7. 1–9.

Gish, H. 1990. Robust discrimination in automatic speaker identification. – Proceedings of International Conference on Acoustics, Speech and Signal Processing, S 5.9. 289–292.

Hermansky, H.; Morgan, N.,;Bayya, A.; Kohn, P. 1991. Compensation for the effect of the communication channel in auditory-like analysis of speech (RASTA-PLP). – Proceedings of European Conference on Speech Communication and Technology, Genova. 1367–1370.

Hess, W. 1983. Pitch Determination of Speech Signals. Berlin, Heidelberg, New York, Tokyo: Springer-Verlag:

Hollien, H.; Majevski, W. 1977. Speaker identification by long-term spectra under normal and distorted speech conditions. – Journal of Acoustical Society of America 62:4, 975–980.

Iivonen, A.; Harinen, K.; Keinänen, L.; Liisanantti, H.; Meister, E.; Tuuri, L. 2001. Moniparametrinen puhujantunnistus. – 21. Fonetiikan Päivät, Turku 4.–5.1.2001. Publications of the Department of Finnish and General Linguistics of the University of Turku. Ed. by S.Ojala, J.Tuomainen. Turku. 81–95.

Koehler, J.; Morgan, N.; Hermansky, H.; Hirsch, H. G.; Tong, G. 1994. Integrating RASTA-PLP into speech recognition. – Proceedings of International Conference on Acoustics, Speech and Signal Processing. 421–424.

Meister, E. 1998. Kõnelejatuvastuse eksperimendid neuronvõrkudel. Magistritöö. Tallinna Tehnikaülikool.

Niemi, T. 1999. Keskiarvospektrit ja Euclidean Distance-arvo forensisessa fonetiikassa. – Out Loud: Papers from 19th Meeting of Finnish Phoneticians. Ed. by J. Järvikivi, J. Heikkinen. Studies in Languages, Vol. 33. 65–75.

Nolan, F. 1983. The Phonetic Bases of Speaker Recognition. Cambridge: Cambridge University Press.

Openshaw, J. P.; Sun, Z. P.; Mason, J. S. 1993. A Comparison of composite features under degraded speech in speaker recognition. – Proceedings of International Conference on Acoustics, Speech and Signal Processing, Vol. 2. 371–374.

Rose, R. C., Reynolds, D. A. 1990. Text independent speaker identification using automatic acoustic segmentation. – Proceedings of International Conference on Acoustics, Speech and Signal Processing, S51.10. 293–296.

Rosenberg, A. E., Lee, C.-H., Soong, F. K. 1994. Cepstral channel normalisation techniques for HMM-based speaker verification. – Proceedings of International Conference on Spoken Language Processing. 1835–1838.

Seneff, S. 1988. A joint synchrony/mean-rate model of auditory speech processing. – Journal of Phonetics 16, 55–76.

Soong, F. K.; Rosenberg, A. E. 1988. On the use of instantaneous and transitional spectral information in speaker recognition. – IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. ASSP-36:6, 871–879.

Zue, V.; Glass, J.; Phillips, M.; Seneff, S. 1989. Acoustic segmentation and phonetic classification in the SUMMIT system. – Proceedings of International Conference on Acoustics, Speech and Signal Processing, S8.1. 389–392.

Emergent nature of morphological paradigms: Plural inflection in Swedish and Finnish[1]

Sinikka Niemi, Jussi Niemi
University of Joensuu

1. Introduction

Recurrent patterns of morphological relationships tend to create analogical or schematic links within and across inflectional paradigms (see, e.g., the classic instances of paradigmatic leveling and pairings like the English *cling* : *clung*, hence also *bring* : **brung*, see, e.g., Bybee and Slobin (1982) for the notion of schema analogy). The present study targets plural inflection in two languages in which phonology partially determines the inflectional category of the lexeme: in one language, Swedish, the inflectional category membership of most nouns is practically unambiguously determined by the last segment(s) of the word after its gender properties have been set. The other language of the present study, Finnish, carries no grammatical gender. Instead, it has more comprehensive, word-shape (or, super-rime, or word-body) analogies within the lexical system (e.g. Paunonen 1976; Skousen 1989). The experiments below will be the first ones to assess the path(s) of the late stages of first-language acquisition of noun inflection in these languages. Using the *wug* paradigm, where the subjects are required to inflect potential, but non-existing words of their

[1] The present study has been partially supported by two Academy of Finland grants (projects: Genetic Language Impairment in Finnish, 1998–2000, Late Language Acquisition, 1998–1999, Principal Researcher: second author), and it is also part of the Canadian Social Sciences and Humanities Research Council MCRI study on the Mental Lexicon (1997–2001), Principal Investigator: prof. Gonia Jarema. We also wish to thank Maarit Silvén and her associates at the University of Turku for collecting the Finnish 7-year-old group data and the staff at the two teacher training schools at the Universities of Vasa and Joensuu for allowing us to collect the remaining data. Thanks also go to Janne Heikkinen and Juhani Järvikivi for their assistance in the construction for the Finnish stimuli and in the data collection of the Finnish portion of the study.

native language (Berko 1958), we compare the emergent development of the expected adult patterns among homogeneous groups of monolingual normal speakers (schoolchildren) of ages 7–8, 10–11 and 14–15 (see the section on *Subjects* below for details).

There are three basic views on the learning of what can be called irregular forms: they are either memorized and stored as such, or they are (or at least the bulk of them, perhaps excluding fully opaque suppletive forms, like *go* : *went*) linked to base forms through structure-changing rules that "generate" the surface forms (e.g. Hoard, Sloat 1973; for the rule-like view, see especially Pinker's work and theoretical discussions in, e.g., Pinker 1991 and Pinker, Prince 1988). A third alternative is based on views of language and cognition, in which units (e.g., words, whether morphologically simple or complex) are associated through probabilistically flexible and ontogenetically emergent links, which can be embedded in parallel distributed processing networks (for first attempts on simulations of Finnish morphology in this type of modelling, see Tikkala, Eikmeyer, Niemi, Laine 1997). The schema-analytical approach adopted by, e.g., Bybee and Slobin (1982) can be subsumed under this third alternative, according to which we may say that (some) irregular forms may be memorized and learnt as such. This does not, however, preclude the speaker from (subconsciously) creating generalizations and novel forms on the basis of these networks of associations. (See also Derwing and Skousen 1994 for a functionally similar account based on exemplar modeling.)

2. Relevant Characteristics of Noun Inflection in Swedish and Finnish

Swedish. Plural inflection of Swedish nouns, which belong to one of the two genders (utrum/*en*, or neuter/*ett*), is relatively simple with its traditional five inflectional classes (seven according to SAG 1999, Vol. 2: 63) compared for example with Finnish. The number inflection is highly dependent on the gender of the lexeme and on the phonological nature of the last segment or segments of its base form (indefinite singular). Like with the Finnish data, a choice of the theoretically most interesting word-types will be presented in this paper; for a fuller description of the Swedish experiment, see S. Niemi (in press). The present inflectional categories are exemplified through: (a) *en* gender nouns with final /a/, e.g., *flicka* : *flick+or* 'girl',

(b) *en* gender items with final /e/, e.g., *pojke* : *pojk+ar* 'boy', (c) *en* gender items with final /eR/ and with subsequent /e/ syncope in the plural, e.g., *syster* : *systr+ar* 'sister', (d) *en* gender words with /are/, e.g., *bagare* : *bagare* 'baker', and (e) *ett* gender words with final /e/, e.g., *äpple* : *äpple+n* 'apple'.

Finnish. Depending on one's level of abstraction and metatheoretical underpinnings, Finnish nouns may be claimed to have as many as 82 (e.g., *Nykysuomen sanakirja* ['Dictionary of Present-day Finnish'], 1973, 2401 pp.) or 49 inflectional classes (e.g., *Suomen kielen perussanakirja* ['Basic Dictionary of Finnish'], 1990, 2008 pp.) or as few as half a dozen (Karlsson 1982). Whatever the number of the paradigms in the descriptions of autonomous morphology, it is a pretheoretical fact that multi-member phonological strings (partially) determine the various surface inflectional patterns in this language (for mathematical modeling of Finnish verb inflection, see the analogy model of Skousen 1989; for an optimality-theoretical view of non-categorical noun inflection in this language, see Anttila 1997). Another feature that makes Finnish an inflectionally highly complicated language is that speakers often have to choose among *both* stem *and* suffix allomorphs to arrive at the correct output. (For the importance of stem allomorphs in processing Finnish, see Niemi, Laine, Tuominen 1994; Järvikivi, Niemi (in press, a); Järvikivi, Niemi (in press, b).) As for nouns, they (and other noun-like, the so-called nominal, classes like adjectives, numerals and pronouns) have 14 productive cases, often with formal differences between plural suffixes of the same case. Since the experiment under discussion deals with the partitive plural, the plural examples below are given in that form only. For the present brief exposition and a comparison with Swedish, the results from select theoretically interesting inflectional categories will be presented, from viz., (a) the agglutinative *pullo* 'bottle' (nom.sg.) : *pullo+ja* (part.pl.; agglutination with no stem change, due to terminal vowel quality), (b) the two competing low vowel stems, *naama* : *naamo+ja* 'face' (stem-final /a/ to /o/ change), and *kuula : kuul+ia* 'bullet' (stem-final /a/ deletion due to the round vowel of initial syllable), and the two competing /s/ stems: *lihas* : *lihaks+ia* 'muscle' (stem-final /s/ to /kse/ change, and with subsequent /e/ deletion in the present context), *kirves* : *kirve+itä* 'axe' (unproductive and lexically marked /s/ type with stem-final /s/ deletion). For a detailed discussion

of these and for an account of the other inflectional categories used in the Finnish experiment, see J. Niemi (submitted).

3. Method

3.1. Stimuli

In both languages sets of disyllabic (Finnish) or di- and trisyllabic (Swedish) pseudonouns were constructed. When creating the pseudowords for these two languages, the critical portions of the wordshapes were left untouched. Thus, for Swedish, following the classic predictions of e.g., SAG (1999), the critical final segment(s) of real words were left unaltered, while one or two phonemes (actually: graphemes) were changed elsewhere. For Finnish, following the predictions of, e.g., Skousen (1989), the pseudonouns were generated by changing the initial consonant (slot) only, producing items like **vullo* from *pullo* 'bottle' and **vuula* from *kuula* 'bullet'. Note that these languages carry relatively transparent orthographies, and that in the items eventually selected for the written *wug* task, complete isomorphy was required to avoid any intervening grapheme to phoneme transformation effects. The source words of the test items were all concrete and frequent nouns in each language. The number of pseudonouns in the Finnish experiment was 180, evenly distributed across the nine inflectional categories analyzed, and the Swedish experiment had 25 test items evenly distributed across the 5 inflectional types chosen.

3.2. Procedures

The classic *wug* procedure (Berko 1958) was adopted for the present experiments. In all its simplicity, the *wug* task requires the subject to inflect a given item embedded in a carrier frame in a form that can be inserted in the subsequent target sentence, e.g., *Here we have one wug. There we have many* _____ (expected: *wugs*).

For the Finnish task, the quantifier *paljon* 'a lot of' was selected, instead of *mon+ta* 'many' (sg.) (cf. the English *many a horse*) and *mon+ia* (pl.) (cf. the E. *many horses*), since we know that speakers, especially children, are sensitive to surface copying of affixal allomorphy, i.e., they are sensitive to schema concord (for schema concord in Finnish, see Laalo 1995; Heikkinen, Järvikivi, Niemi 2000). In both languages, the subjects were given explicit written and oral

instructions and they were free to respond with no temporal pressure. The order of the items on the test sheets was randomized with the Excel spreadsheet program also used in the analysis. (For further details, see S. Niemi (in press) and J. Niemi (submitted).)

3.3. Subjects

The Swedish-speaking subject populations were tested at the Swedish-speaking University of Vasa Teacher Training School (*Vasa övningsskola*) in the bilingual coastal town of Vaasa/Vasa (population 56,600, out of which 25,6% are Swedish-speaking). These subjects were *ex post facto* selected to the monolingual and bilingual groups on the basis of a thorough language proficiency questionnaire (for details, see S. Niemi 2001). For the present discussion, only the monolingual subjects will be analyzed. The Finnish-speaking subjects were tested in Turku and Joensuu. All in all, 63 7–8-year old, 16 10–11-year old and 16 14-year-old Finnish-speaking children as well as 24 8–9-year-old, 25 10–11-year-old and 11 15-year-old Swedish-speaking children participated in the experiments that we now report. For brevity, the groups with two age values (years) will below be reffered to by using the lower age value of the two.

4. Results and Discussion

The results, first of all, show that in both languages speakers around age ten are still moving towards an adult-like internalization of the basic phonological cues to category-assignment, since in both languages the youngest subjects produce significantly fewer expected responses than their 10-year-old peers (Figure 1) (two-tailed t-tests with infinite degrees of freedom: Swedish-speaking groups: $t = 8.778$, $p < 0.005$; the two youngest Finnish-speaking groups: $t = 7.601$, $p < 0.005$), not to speak of the oldest groups. Moreover, the practically 100% correct responses of the oldest Finnish group show that we are dealing with real effects. In other words, subjects treat these pseudo-items as if they were real words of their language. The full mastery of Finnish morphology – as measured with the present procedure – appears to stabilize during the first half of the second biological decade – quite late , if one considers on the research effort that is generally focused on the early years of language development. Furthermore, the morphological patterns of Swedish are opaque well into (or

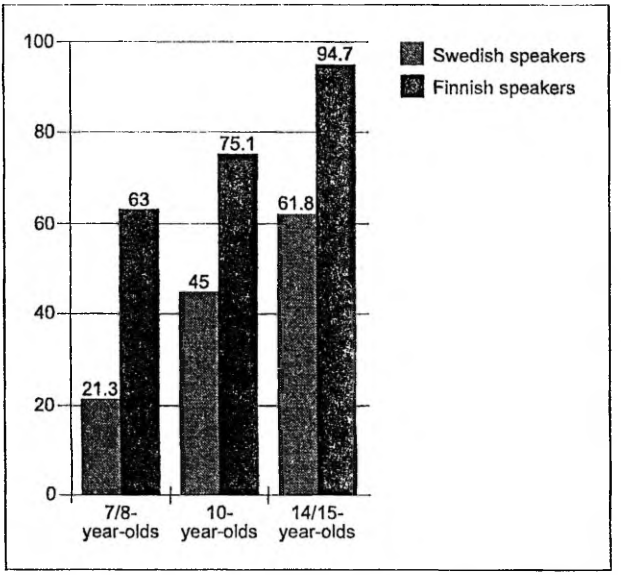


Figure 1. The overall number (%) of expected responses in the plural inflection experiments with pseudowords

N's for the groups: Swedish 8-year-olds 600, 10-year-olds 625 and 15-year-olds 275; Finnish 7-year-olds 1926, 10-year-olds 1536 and 14-year-olds 1524.

even beyond) the latter part of the second decade, since the overall correctness score of the oldest group is as low as 61.8%. This is striking, since the key cues for unambiguous paradigm assignment – the gender (as signaled through the indefinite article) and the last critical segment(s) of the stem – were visible to the subjects (see below for further discussion). At this stage, one of the messages that we have is that theoretically interesting changes take place in one's first-language grammar well after the beginning of the school-age.

Accordingly, the morphological mastery of a language may stabilize quite late in ontogenesis, much later than any model of the critical period would predict. Moreover, we have to assume that the speaker groups of the same age levels are pairwise totally equivalent as regards their general cultural and cognitive background. For instance, they are all monolingual residents of Finland attending publicly funded urban comprehensive schools using the same basic curricula. Accordingly, the dividing notion between the age-matched peers is their first language, Swedish versus Finnish. If we accept this

line of argumentation about the comparability of the age groups, it is interesting to observe that the Finnish subjects exhibit a higher level of morphological mastery than the Swedish groups (Figure 1) in spite of the fact that Finnish is definitely morphologically the more complex one of the two. There is, we claim, an elegant and theoretically sound way out of this apparent paradox: speakers of Finnish perform better in the morphological task simply because the language has more phonological material to signal category membership than Swedish. If this language-related distinction is embedded within the frameworks of distributed processing and connectionist associative networks, it can be hypothesized that the better mastery of the Finnish speakers is due to the larger number of units and hence also connections that bridge the singular and plural forms (see, e.g., Tikkala et al 1997). In other words, in Finnish the body or the super-rime contains more cues than the last segment(s) of Swedish. Observe that this view receives some additional support from within Swedish, since the inflectional class that shows better scores in the oldest age group is the one with more than one terminal segment signaling the type of plural, and this effect appears to be cumulative, as will be discussed below (see Figure 2 for breakdown of the data by morphological paradigm). The difference between the correctness scores of **tindare* and **pyster*, 89.1% and 74.5% respectively, is statistically significant (one-tailed *t* test with infinite degrees of freedom, $t = 1.984$, $0.025 > p > 0.01$). The difference between **pyster* and **jaspe* (56.4 % correct) is also significant ($t = 2.000$, $0.025 > p > 0.01$). In **jaspe* the only cue of the stem to show paradigm membership is the final /e/, while in **pyster* we may claim that the cue is composed of the pre-terminal vowel and the resonant, perhaps even of the consonant preceding the /e/, thus eventually creating /CeR/ as the cue schema. It is interesting that this word-type is processed as well as it is, since its plural inflection does involve the deletion of the pre-terminal /e/. Finally, in the **tindare* type we have a phonologically stable terminal sequence /are/ to guide correct plural assignment, and it is this word-type that receives the highest scores, although semiotically the plural inflection of the **tindare* : **tindare* type is the least natural, or the least transparent of the types now analyzed, since the added morphological, or rather morphosemantic complexity is not accompanied by any phonological change in this category (see, e.g., Dressler 1985 for phonological and morphological naturalness from a semiotic point of view).

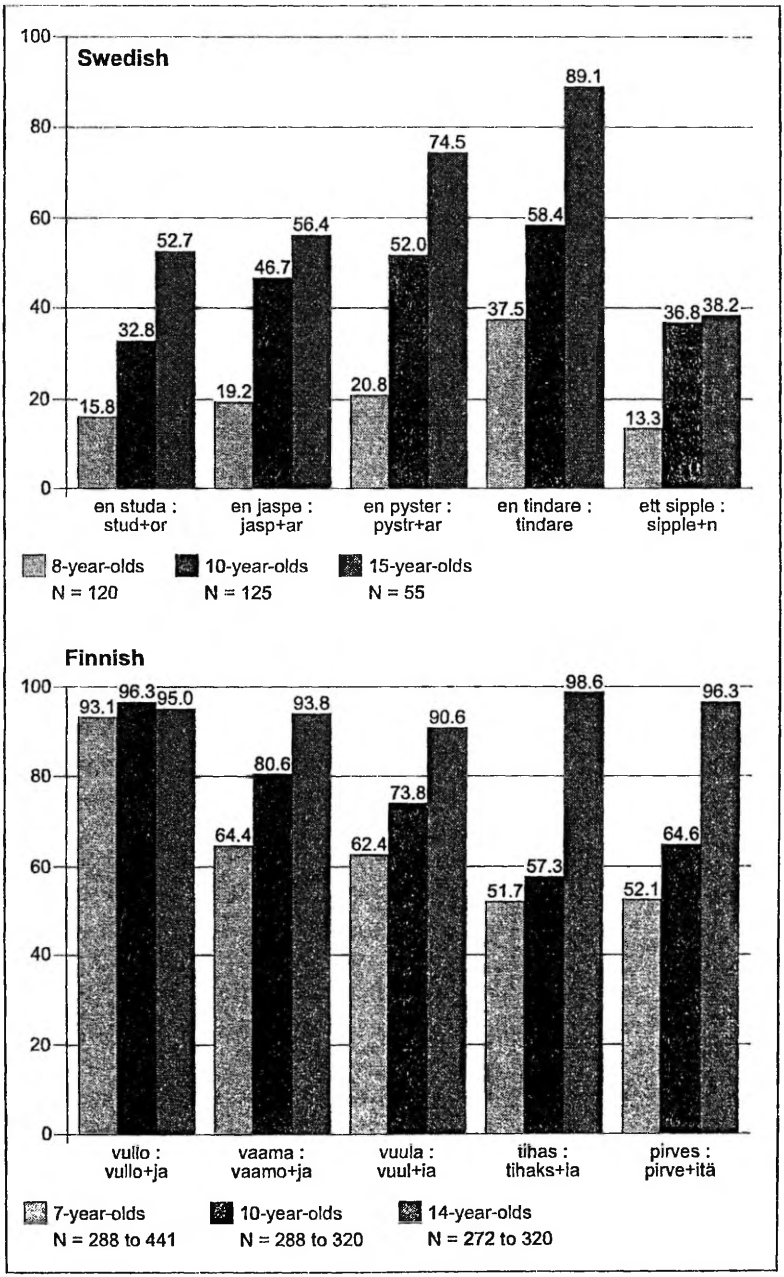


Figure 2. The percentages of expected responses in the plural inflection experiments with pseudowords broken down by morphological category

The breakdown of the Finnish data (Figure 2) shows that, not unexpectedly, the agglutinative plural marking of type **vullo : *vullo+ja* receives the highest correctness scores in the two younger groups, while the paradigms with morphophonological changes in the stem *and* with potential competitors obtain lower hit rates. When comparing the two /low/ vowel types (**vaama* : **vaamo+ja* vs. **vuula* : **vuul+ia*), one would expect a better performance on the one with vowel alternation, since cognitively the alternating pattern is the more natural alternative, as it better retains the meaning-to-form isomorphy than does the deletion type, in which the affix is also phonologically assimilated with the stem syllable (**vuu.lia*), while the alternating type retains the disyllabicity of the stem and adds a syllable *cum* affix to it (**vaamo+ja*, syllabified as **vaa.mo.ja*) (see Dressler 1985; for observations on the higher productivity of the alternating type at the *early* stages of the acquisition of Finnish, see Niemi, Niemi 1987). It thus appears that by age 7 Finnish children have encountered so many exemplars of both categories that their "division of labor" is relatively firmly established. In network-theoretical terms, one could say that their network is equally saturated with input from both paradigms. However, the performance of the two younger Finnish groups on the remaining two competing paradigms, viz., those with /s/ (**tihas* : **tihaks+ia* vs. **pirves* : **pirve+itä*) is somewhat less adequate as their overall scores are lower than those of the /a/ paradigms. The emerging stability at this age is, however, shown by the fact that in the /s/ paradigms we also obtain identical correctness scores for the two competing categories, i.e., for the productive *tihas* : *tihaks+ia* type and for the lexically marked and unproductive *pirves* : *pirve+itä* type. Like with the /low/ vowel paradigms, the early stages of acquisition show clear signs of bias towards one of these classes, viz., towards the productive *s* : *kse* type (Dasinger 1997; Niemi, Niemi 1987). However, the unproductive nature of the latter alternation type, /s/ with phonological zero, does not prevent the 7- to 14-year-old subjects from applying it to novel items. We assume that, like in the /a/ stems, the numbers of instances encountered even among the unproductive type are high enough to make it possible for the speaker to (subconsciously) establish relatively stable, or highly activated associations between phonological strings (bodies, super-rimes) like [*k*]*irves* : [*k*]*irveitä* (hence also [**p*]*irves* : [**p*]*irveitä*). Accordingly, we would predict that a novel word, e.g., a commercial brand name,

like *Tirves* could *also* receive a considerable proportion of the allegedly unproductive type *Tirveitä*, in addition to *Tirvesejä* and *Tirveksiä*, although the latter two are productive and, furthermore, they better retain the isomorphy relationships favored especially in recent proper names (cf., e.g., the established first name *Johannes* : *Johanneksia* vs. *Tekes* : *Tekesejä*/?*Tekeksiä*/**Tekeitä*, a new governmental body in its common acronym form).

In sum, the present data from two typologically different languages show that the acquisition of first-language morphology is a slow, emergent process that extends well into the second biological decade (see, e.g., Elman 1999 for a review of emergence and references therein to brain mapping studies showing retention of plasticity of the human cortex into adulthood). Moreover, the paradigm-internal form-based cues become stronger or more transparent as more phonological material (segment(s), schematic sequences) is used in the relevant morpholexical associations or links. This conclusion is based both on the language-internal (cf. Swedish **tindare* > **pyster* > **jaspe*) and cross-language (Finnish word-bodies > Swedish final segmen(s)) results obtained.

Finally, the view of autonomous morphology on the categorical nature of morphological category-membership, transparency and related notions, like "all Swedish *en* gender nouns with a terminal /a/ fall in Category 1, and receive /or/ in the plural", is seriously questioned by the present results, according to which morphological relationships and categories are better seen through, e.g., non-categorical associations and schematic analogies, thus also implying and licensing fluctuation in function of word-type and ontogenetic stage.

References

Anttila, Arto. 1997. Variation in Finnish Phonology and Morphology. PhD Dissertation, Stanford University.

Berko, Jean 1958. The child's learning of English morphology. – Word 14, 150–177.

Bybee, Joan 1985. Morphology: A Study of the Relation between Meaning and Form. Amsterdam: Benjamins.

Bybee, Joan; Slobin, Dan 1982. Rules and schemas in the development and use of English past tense. – Language 58, 265–289.

Dasinger, Lise 1997. Issues in the acquisition of Estonian, Finnish, and Hungarian: A crosslinguistic comparison. – The Crosslinguistic Study of Language Acquisition, Vol. 4. Ed. by Dan Slobin. Mahwah, NJ: Lawrence Erlbaum. 1–86.

Derwing, Bruce; Skousen, Royal 1994. Productivity and the English past tense: Testing Skousen's Analogy Model. – The Reality of Linguistic Rules. Ed. by Susan Lima, Roberta Corrigan, Gregory Iverson. Amsterdam: Benjamins. 193–218.

Dressler, Wolfgang 1985. Morphonology. Ann Arbor: Karoma.

Elman, Jeffrey 1999. The emergence of language: A conspiracy theory. – The Emergence of Language. Ed. by Brian MacWhinney. Mahwah, NJ: Lawrence Erlbaum. 1–27.

Heikkinen, Janne; Järvikivi, Juhani; Niemi, Jussi 2000. NP agreement and government in familial language impairment: Evidence from Finnish. – Poster presented at the Second International Conference on the Mental Lexicon, Montreal, October 2000.

Hoard, James; Sloat, Clarence 1973. English irregular verbs. – Language 49, 107–120.

Järvikivi, Juhani; Niemi, Jussi (in press, a). Stem allomorphs in on-line processing. – To appear in the Proceedings from the 34th Linguistics Colloquium, University of Mainz, September 1999.

Järvikivi, Juhani; Niemi, Jussi (in press, b). Form-based representation in the mental lexicon: Priming (with) bound stem allomorphs in Finnish. – Brain and Language.

Karlsson, Fred 1982. Suomen kielen äänne- ja muotorakenne. Porvoo: WSOY.

Laalo, Klaus 1995. Skeemakongruenssi lapsenkielessä. – Virittäjä 99, 153–172.

Niemi, Jussi 1999. Production of grammatical number in Specific Language Impairment. – Brain and Language 68, 262–267.

Niemi, Jussi (submitted). Allomorph selection in plural formation: Late stages of normal acquisition and Familial (Genetic) Language Impairment.

Niemi, Jussi; Laine, Matti; Tuominen, Juhani 1994. Cognitice morphology in Finnish: Foundations of a new model. – Language and Cognitive Processes 9, 423–446.

Niemi, Jussi; Niemi, Sinikka 1987. Acquisition of inflectional marking: A case study of Finnish. – Nordic Journal of Linguistics 10, 59–89.

Niemi, Sinikka (in press). Schemas and competing paradigms in Swedish plural formation. – Brain and Language.

Niemi, Sinikka 2001. Swedish Syntax at Late Stages of Language Acquisition: Normal Monolinguals and Bilinguals and SLI Speakers. Publications in the Humanities. University of Joensuu.

Paunonen, Heikki 1976. Allomorfien dynamiikkaa. – Virittäjä 80, 82–107.

Pinker, Steven 1991. Rules of language. – Science 253, 530–535.

Pinker, Steven; Prince, Alan 1988. On language and connectionism: Analysis of parallel distributed processing model of language acquisition. – Cognition 28, 73–193.

SAG 1999: Svenska Akademiens Grammatik. Ed. by Ulf Teleman, Staffan Hallberg, Erik Andersson. Uddevalla: MediaPrint.

Skousen, Royal 1989. Analogical Modeling of Language. Dordrecht: Kluwer.

Tikkala, Anneli; Eikmeyer, Hans-Jürgen; Niemi, Jussi; Laine, Matti 1997. The production of Finnish nouns: A psycholinguistically motivated connectionist model. – Connection Science 9, 295–314.

Kas tesaurus ja tekstid lähevad kasutuses kokku?

Heili Orav, Kadri Vider
Tartu ülikool

Sissejuhatus

1960. aastate lõpul moodustati Tartu ülikoolis juristidest, keeleteadlastest ja matemaatikutest uurimisrühm, mis töötas välja tesaurusel põhineva süsteemi juriidilise info otsimiseks (Koit, Õim 1998). Haldur Õim võeti 1969 selle rühma "keele-eksperdiks". Süsteemi nimetati JURIOS ("Juriidiline Infootsisüsteem") ja selle oluliseks komponendiks oli tesaurus, mis toimis vahendajana programmide ja tekstimassiivide vahel. JURIOS tesauruses oli 15 000 märksõna (lähemalt vt Valge, Õim 1976) Tollal oli süsteem enam-vähem ainus keele automaattöötluse praktiline rakendus Eestis.

1970. aastail hakati maailmas jõudsalt arendama freimisemantika nime all tuntud semantilisi võrgustikke. Sellest uuest suunast said inspiratsiooni ka H. Õim ning tema kolleegid 1980. aastal rajatud TRÜ tehisintellekti laborist. Töögrupis tegeldi venekeelse teksti automaattöötlusega, eesti keele automaatse süntaktilise ja semantilise analüüsi ja sünteesiga, samuti teksti mõistva süsteemi TARLUS väljatöötamisega.[1]

1993. aastal moodustati Tartu Ülikooli juurde arvutuslingvistika uurimisgrupp, kus jätkub ühe allteemana keele semantiline analüüs. 1997. aastal alustasime eesti keele põhisõnavara tesauruse koostamist WordNeti eeskujul. 1998.–99. aastal koostasime H. Õimu eestvedamisel eesti keele tesaurust EuroWordNet-2 projektis osalejatena ja Euroopa Komisjoni toetusel. Mahuka semantilise andmebaasi loomise aeganõudvat tööd on tehtud ka H. Õimu juhitud Eesti Teadusfondi grantide ja Eesti Informaatikakeskuse projektide raames.

Loodav tesaurus põhineb olemasolevatel traditsioonilistel sõnaraamatutel ja tekstikorpusel (mis annab teavet sõnakasutusest), seega võib semantilist informatsiooni, mida andmebaas sisaldab, pidada keelelisel teadmisel põhinevaks.

---

[1] Süsteemi kohta vt TRÜ toimetised nr. 594 (1981), 621 (1983), 688 (1984).

Käesoleval artiklil on kaks eesmärki: kirjeldada eesti keele põhisõnavara tesauruse ehk eesti wordneti (EstWN) hetkeseisu ja anda ülevaade sellest, kuidas käsitsi semantilisest ühestamisest on kasu tesauruse parandamisel ja täiendamisel.

Eesti keele tesaurus praegu

EstWN[2] 39. versioonis (23.11.2001) on 9729 sünohulka,[3] 13 733 erinevat sõna ja sõnaühendit ning 17 493 erinevat tähendust. Enamus sünohulki on ühendatud hierarhiaid tekitavate ülem- ja alamsuhetega, teistele semantilistele suhetele on pööratud vähem tähelepanu.

Tabel 1. EstWN 2001. aasta novembris

	Nimisõnades	Verbides	Adjektiivides	Kokku
Sünohulki	6734	2757	307	9798
Sõnade tähendusi (variante)	11230	5745	518	17493
Tähendusi sünohulga kohta	1,67	2,08	1,69	1,79
Erinevaid sõnu ja sõnaühendeid	9518	3796	419	13733
Tähendusi sõna kohta	1,18	1,51	1,24	1,27
Semantilisi suhteid	14549	5596	538	20683
Semantilisi suhteid sünohulga kohta	2,16	2,03	1,75	2,11

Algselt pidi EstWN katma eesti keele põhisõnavara, mis tehti kindlaks korpuse statistilise analüüsi käigus. Siiski ei ole alati võimalik panna piiri nende sõnade tesaurusesse viimisele, mille esinemissagedus on väga väike. Sõnavarastatistika põhjal tehtud sõnaloend kaasab ka sagedaste sõnade vähem- ja väga vähe olulised tähendused, mis omakorda toovad kaasa vähesagedasi sõnu sünonüümidena. Tekstides vähe esinevad on ka võimalikult üksikasjalikuks arendatud

[2] Eesti wordneti (EstWN) põhjalik kirjeldus on antud artiklis "Eesti keele tesaurus" (Vider jt 2000).

[3] Sünohulk (*synonym set, synset*) on Wordneti elementaarosake, mille moodustavad ühte mõistet väljendavad sünonüümsed sõnad ja sõnaühendid. Termini sünohulk oleme võtnud kasutusele seepärast, et erinevalt sünonüümisõnastiku sünonüümireast võib sünohulk olla ka ühesõnaline.

hierarhiates leiduvad peaaegu terminilaadsed sõnad (selliseid hierarhiaid püüti saavutada mõnede semantiliste alade, näiteks muusikariistade kohta EuroWordNet-2 raames).

Rakendasime oma tesaurust kõigepealt tekstisõnade semantilise ühestamise katsetamisel.

Semantilise ühestamise ülesande püstitus

Et mingil keeletehnoloogia rakendusel oleks võimalik teksti "mõista", olgu tegemist infootsisüsteemi või masintõlkega, on vajalik, et tekstis olevate sõnade tähendused oleksid üheselt määratud ehk tekst peab olema *semantiliselt ühestatud* (mitmetähenduslike sõnade puhul tähendab see antud juhul realiseeruva tähenduse kindlakstegemist).

Semantilise ühestamise aktuaalsuse järsust kasvust annavad ilmeka pildi arvud: kaks aastat tagasi korraldatud SENSEVAL-1 (*Evaluating Word Sense Disambiguation Systems*) projektis osales 3 keelt (inglise, prantsuse, itaalia), sel aastal toimunud SENSEVAL-2-s 12 keelt (ka näiteks korea ja jaapani keel, lisaks muidugi eesti keel).

Automaatse semantilise ühestamise programmi koostamine koosneb tavapäraselt kahest allülesandest. Esiteks tuleb valitud tekstides sõnad käsitsi ühestada, st luua semantiliselt ühestatud korpus, mis on kasutatav treening- ja testkorpusena. Teiseks tuleb koostada automaatse semantilise ühestamise programm, mida seejärel rakendatakse samadel tekstidel. Tulemuste võrdluse põhjal programmi täiustatakse. Tesauruse täiustamise töös on aga võimalik ära kasutada ka esimese allülesande lahendamise tulemusi.

2001. aasta lõpuks on olemas nii arvestatava mahuga testkorpus kui ka automaatse ühestamise programmi *semyhe* (vt Kahusk, Kaljurand siinses kogumikus) töötav variant.

Testkorpuse katsetamine ja käsitsi ühestamine

2001. aastal on käsitsi semantiliselt ühestatud umbes 30 000 tekstisõnaga korpus, milles ühestatavaid sõnu on ligikaudu 10 000. Seda korpust kasutasime ka SENSEVAL-2 eestikeelse ülesande püstitamisel. Tekstid on võetud Eesti Kirjakeele Korpusest (TÜKK). Eesti keele tesauruse tähendusnumbrite põhjal ühestati tekstides ainult substantiive ja verbe, kuna adjektiivide sisestamine tesaurusesse on alles algusjärgus.

Iga teksti ühestas käsitsi kaks inimest. Sõna tähenduse number märgiti talle vastava numbriga EstWN-s. Kui sõna tesauruses puudus, märgiti tähendusnumbriks "0" ning "+1" märgiti juhtudel, kui sõna küll esines tesauruses, aga puudus konkreetsele kasutusjuhule vastav tähendus. Kui ühestajatel tekkis eriarvamusi, vaadati need koos läbi ja lepiti kokku, milline variant valida. Erinevaid tõlgendusi tekstitähendustele oli algul rohkesti, peale sagedamini esinevate erimeelsuste analüüsi jäi neid vähemaks, kuid keskmiselt ühestati erinevalt umbes 28% juhtudest. Sagedasemate verbide mitmetitõlgendatavust aitas lahendada ka süntaktilise informatsiooni arvestamine: verb 'olema' sai erinevad, kindlad tähendusnumbrid eksistentsiaal-, possessiiv- ja predikatiivlausetes; modaalse tähendusega olid verbi 'pidama' 12 tähendusest kaks ja verbi 'saama' 12 tähendusest üks.

Üks ühestajate arusaamade suurte erinevuste põhjusi oli see, et sõnatähendus tesauruses oli ühestajate arvates kas liiga spetsiifiline või liiga üldine. Esines nii tähenduste **üle-eristamist**, milles kahe tähenduse vaheline erinevus on väga ähmane (sageli tegelikult vaid kasutuskontekstist sõltuv), kui ka **ala-eristamist** – seda maksab kahtlustada, kui tesauruse sõnaseletuses esineb osalauseid rinnastav 'või', mis viitab tegelikult kahele erinevale tähendusele.

Omaette probleemiks kujunesid mõistelised püsiühendid ja nende tuvastamine tekstides. Tesauruses leidub märksõnadena 984 sellist sõnaühendit, kolmveerand neist on ühend- ja väljendverbid. Mõistelised püsiühendid tekivad tesaurusse sünonüümidena (näiteks 'meenutama, meelde tuletama' või 'üllitama, välja andma') või spetsiifiliste sõlmedena hierarhiates (nt 'bioloogiline protsess', 'must turg'). Et nimisõnaliste ühendite komponendid paiknevad tekstis eeldatavasti üksteise vahetus läheduses, on neid lihtsam eelnevalt automaatselt üles leida ja ka tesauruse sõnaloendiga võrrelda. Verbiühendite leidmine käib aga lemmatiseerijale praegu veel üle jõu. Ka käsitsi ühestajatel tekkis verbiühendite ülesleidmisega raskusi ja erimeelsusi nende tuvastamisel.

Kui palju sõnu tekstides tesaurus praegu katab?

Tesaurusest puuduvaid tähendusi (tähendusnumbriks märgiti "+1") oli tervelt 17% sõnadest. Põhjus on ajalooline – need sõnad on tesaurusesse lisatud kui mingi sagedase sõna sünonüümid ja nende teisi tähendusi pole jõutud veel läbi vaadata. Tesauruses esindatud sõnade puuduvate tähenduste leidmine tekstidest on semantilise käsitsiühes-

tamise olulisemaid tulemusi, kuna neid pole võimalik leida automaatsete meetoditega. Semantiliselt ühestamata (tähendusnumbriks märgiti "0") jäid substantiivid ja verbid, mida tesauruses polnud.

Ligi pooled sellistest sõnadest on liitsõnad. Keelekasutajal on lihtne välja mõelda uusi liitsõnu, mida ei leia ühestki sõnastikust, kuid mida teistel eesti keelt rääkivatel inimestel on väga kerge mõista. Kõigi selliste liitsõnade lisamine tesaurusse pole otstarbekas. Lisamise määravaks tingimuseks võiks olla liitsõna mõne komponendi vähene esinemine iseseisva sõnana.

Pärisnimesid pole EstWN-sse veel sisestatud. Kuna morfoloogiline lemmatiseerija analüüsib neid tekstis substantiivideks, moodustasid 17,5% nulltähendusega sõnadest pärisnimed. Tesauruses on pärisnimede jaoks omaette sõnaliik ja lähitulevikus on plaan tekstides esinenud pärisnimed ka sisestada.

Tabel 2. Tesauruse katvus tekstides

	Tekstides
Erinevaid lemmasid	3912
Erinevaid tähendusi	4714
Tesaurusest puuduvaid lemmasid	1697
Tesaurusest puuduvaid tähendusi	383
Lemmasid 1 tähenduses	1789
Lemmasid 2 tähenduses	292
Lemmasid 3 tähenduses	106
Lemmasid 4 tähenduses	32
Lemmasid rohkem kui 4 tähenduses	32

Mitte kõik tekstides esinenud sõnade tesauruses esitatud tähendused polnud kasutamist leidnud. Seda näitab tabel 3, mis võrdleb mõne lemma tähenduste hulka tekstides ja tesauruses. Erandlikud olid 'tegema' ja 'tulema', mille tähenduste arv tekstides oli suurem kui tesauruses, sest tekstidest selgusid tesaurusest puuduvad tähendused.

Tabel 3. Tähendusrikkamad lemmad tekstides

Sõnaliik	Lemma	Tekstitähendusi	Tesaurusetähendusi
Verb	saama	12	12
Substantiiv	asi	10	11
Verb	käima	9	23
Verb	olema	8	9
Verb	pidama	7	12
Verb	panema	7	11
Verb	leidma	7	8
Verb	nägema	7	7
Verb	võtma	7	7
Verb	tegema	7	6
Verb	ajama	6	17
Verb	minema	6	17
Verb	andma	6	13
Substantiiv	maa	6	8
Substantiiv	päev	6	8
Substantiiv	elu	6	7
Verb	tulema	6	5

Kokkuvõte

Tekstide käsitsi semantilise ühestamise tulemusel leidsime ja lisasime suure hulga puuduvaid sünohulki ning tähendusi meie tesaurusesse. Erinevad arvamused inimeste vahel, kes ühestasid neid tekste, näitasid meile EstWN-i kõige problemaatilisemaid kohti.

Artiklis mainitud põhjustel me eeldame, et EstWN sisaldab juba praegu rohkem kui ainult baassõnavara. Käsitsiühestamise tulemused osutavad aga nii semantiliselt kui sõnakasutuse seisukohalt katmata aladele meie tesauruses.

Kirjandus

Erelt, Mati 1984. Doktoriväitekiri teksti mõistmise alalt (H. Õim). – Keel ja Kirjandus 3, 188–189.

Koit, Mare; Õim, Haldur 1998. Arvutuslingvistika mujal ja meil. – Keel ja Kirjandus 1, 1–7.

Valge, Jüri; Õim, Haldur 1976. Teooria praktikasse: automatiseeritud infosüsteemid. – Keel, mida me uurime. Koost. M. Mäger. Tallinn: Valgus. 18–22.

Vider, Kadri; Kahusk, Neeme; Orav, Heili; Õim, Haldur; Paldre, Leho 2000. Eesti keele tesaurus. – Arvutuslingvistikalt inimesele TÜ üldkeeleteaduse õppetooli toimetised 1. Toim. T. Hennoste. Tartu: Tartu ülikool. 127–152.

EuroWordNet – http://www.hum.uva.nl/~ewn/

SENSEVAL-2 – http://www.sle.sharp.co.uk/senseval2/

TÜKK – http://www.cl.ut.ee/ee/corpusb/

WordNet – http://www.cogsci.princeton.edu/~wn/w3wn.html

Keelemudel

Silvi Tenjes
Tartu ülikool

1. Sissejuhatus

Artiklis vaadeldakse kolmetasandilist keelemudelit. Mudelit seostatakse semantika, metafooride, vähem ka keeletekke ja žestidega. Mudel põhineb W. von Humboldti ideedel keelest. Artiklis tuvastatakse nende mõtete edasiarendusi H. Õimu semantikakäsitlustes. Artikli lõpuosas on arutluse all metafooride ja käežestide ühendus kolmetasandilise keelemudeliga ning H. Õimu seisukohad nendega seoses.

Kui ma mõtlen prof. Haldur Õimu fenomenile, siis on mind paelunud tema võime leida teoreetilise keeleteaduse suundade, seisukohtade ja vaadete paljususest kõige edumeelsemad ning antud ajas olulisemad. Samuti tema oskus leida minevikust viljakaid liine tänapäeva keeleteooria jaoks. Ühe sellise näitena võiks nimetada Wilhelm von Humboldti keeleteoreetilisi seisukohti, mida prof. Õim on minu jaoks huvitavaks kõnelenud. Humboldti seisukohad keelest pärinevad 19. sajandist, kuid osasid neist pole ei veenvalt tõestatud ega ümber lükatud. Käsitleksin järgnevalt lühidalt Humboldti vaateid, eelkõige tema ja tema interpreteerijate seisukohti keele 'sisevormist', ning mõnesid prof H. Õimu huvitavaid tähelepanekuid nendega seoses.

2. W. von Humboldti keeleteoreetilised seisukohad

W. von Humboldt (1767–1835) oli oluline hariduselu edendaja Preisi kuningriigis, oli palju reisinud ning oskas mitmeid nii Läänes kui Idas kõneldavaid keeli. Ta kirjutas mitmeid keeleteoreetilisi artikleid, millest kõige olulisemaks peetakse tööd pealkirjaga "Inimkeele struktuuri mitmekesisusest" (Humboldt 1949[1]). Artikkel ilmus pärast

[1] W. von Humboldti "Über die Verschiedenheit des menschlichen Sprachbaues" ilmus 1836. Siin on viidatud artikli kordustrükile: W. von Humboldt "Über die Verschiedenheit des menschlichen Sprachbaues", Darmstadt, 1949.

tema surma pikema sissejuhatusena Jaava saarel muiste kõneldud kaavi keele kirjelduse ees.

Humboldti keeleteooria aluseks on loov keelevõime, mis sisaldub loomupäraselt iga kõneleja ajus või meeles (ingl *mind*) (Robins 1967: 174). Keel on lahutamatult seotud selle elava võimega, mille kaudu kõnelejad toovad esile ja mõistavad lausungeid, mitte kõneaktide või kirjutamise nähtava produktiga. Humboldti enese sõnul on see loov võime (*energeia, Tätigkeit, Erzeugung*), mitte lihtsalt tulemus või saadus (*ergon, Werk, Erzeugtes*) (Humboldt 1949: 43–44). Keelevõime on inimvaimu olemuslik osa. Võimelisuse olemuse poolest võivad keeled muutuda ja kohaneda, kui olukord seda nõuab. Ainult niimoodi võib selgitada keele keskset omadust (ja müsteeriumi!), et kõnelejad võivad lõpmatult kasutada keelevaramu lõplikkust enda jaoks igal ajal[2]. Seepärast jääb vaatamata keele kirjeldustele ja analüüsimistele midagi tema olemuslikust loomusest seletamata. Võib olla see ongi üks oluline põhjus, miks Humboldti teooriad – vähemalt osaliselt – on köitvad ka tänapäeval.

Kuigi keelevõime on universaalne, läheb Humboldt siin kaugemale ja väidab, et on olemas iga erineva keele individuaalsus kui omapärane rahva või seda keelt kõneleva grupi omadus. Kõnelemise artikulatoorne baas on ühine kõikidele inimestele, kuid hääl on vaid passiivne materjal keele vormilisele moodustusele või struktuurile (*innere Sprachform*) (Humboldt 1949: 89–98, 269). Humboldti *innere Sprachform* – sisevorm – on keele semantiline ja grammatiline struktuur, kehastunud elemendid, mustrid ja reeglid, millega on mõjutatud kõne toormaterjal (Robins 1967: 175). Osaliselt on see ühine kõikidele inimestele, olles kaasatud inimese intellektuaalsesse varustusse, kuid teisalt moodustab iga keele eraldatud *Sprachform* selle keele vormilise identsuse ja erinevuse kõikidest teistest keeltest (seda võib mõnes mõttes võrrelda F. de Saussure'i *langue-parole* eristusega). See iga keele organiseeriv printsiip haldab tema silbistruktuure, grammatikat ja leksikoni (Humboldt 1949: 48). Humboldt väljendab ka mõtte ja keele seotust: 'inimeste kõne on nende vaim, ja nende vaim on nende kõne'[3]. Humboldti sellesuunalised seisukohad

[2] 'Die Sprache muss von endlichen Mitteln einen unendlichen Gebrauch machen' (Humboldt 1949: 103).

[3] 'Ihre Sprache ist ihr Geist und ihr Geist ihre Sprache' (Humboldt 1949: 41).

on olnud aluseks Ameerika keeleteadlaste F. Boasi, E. Sapiri, B. L. Whorfi jt. arusaamadele keelest.

Humboldti arvamusi on mõjutanud I. Kanti filosoofiline süsteem. Inimese vaim pole Kanti järgi passiivne vaha, millele kogemus ja aisting kirjutavad oma kirja. Ka pole see ainult abstraktne nimi hingeseisundite tarvis. See on tegev, aktiivne organ, mis kujundab ja korrastab aistingud tajudeks ja kujutlusteks, see on elund, mis transformeerib kogemuse kaootilise paljuse mõtte korrastatud ühtsusesse (Künnapas 1992: 169). Selles aistingu toormaterjali ümbertöötamise protsessis on mõtlemise valmissaaduseks kaks järku või astet. Esimeseks astmeks on aistingute koordineerimine taju vormide – aja ja ruumi – rakendamise kaudu neile. Teiseks astmeks on nii saadud tajumite koordineerimine mõtlemise vormide, mõtlemise kategooriate rakendamise kaudu (Künnapas 1992: 170). See oli universaalne filosoofiline teooria, mille Humboldt kohandas relativistlikult ja lingvistiliselt *innere Sprachform'i* jaoks. Iga keele *innere Sprachform* vastutab kogemuse andmete korrastamise ja kategoriseerimise eest, nii et erinevaid keeli kõnelevad inimesed elavad osaliselt erinevates maailmades ja neil on erinev mõtlemissüsteem. Humboldt kasutas seoses keele toimimisega kolme mõistet: *Anschauen*, *Denken* ja *Fühlen* (tajumine, mõtlemine ja tundmine) (Humboldt 1949: 90).

3. H. Õimu süvakihi idee semantikakäsitlustes

Juba professor H. Õimu varasemates töödes võime leida jälgi, milles on idee keeles peituvast "kihist", võib olla isegi mitmest "kihist". Järgnevalt on Õim lausete moodustamise ja mõistmise protsessi kujutamisel väljendanud ideed teatud alusstruktuurist. "Kuid kas meie teadmised, mida me lausete abil vahendame, ongi mälus esitatud alusstruktuuride kujul? Seda saab uurida lausete mõistmise abil. /.../ Et me mõne aja möödudes mäletame, mis meile öeldi, on üsna tavaline. Kuid kui sageli mäletame, missuguse lausega seda meile öeldi? Kui loeme või kuuleme jutustust mingist sündmusest, siis on tavaline, et mäletame jutustust mõnda aega ja võime seda teistele edasi rääkida, reprodutseerides isegi üksikuid detaile. Kuid me ei mäleta peaaegu kunagi – ega püüagi mäletada –, kuidas jutustus oli lauseteks jagatud, missuguste "tükkide" kaupa vastav sündmus meile esitati; ja veel vähem mäletame, kuidas konkreetsed laused välja nägid. Kui keegi mäletab iga üksikut lauset, siis on see pigem haruldane kui tavaline. /.../ Kuulajale jääb meelde sisuline

fakt, mitte see, missuguse lausega tõsiasi teatavaks tehti." (Õim 1974: 58–59) Ilmselt ei säili info mälus vastavate lausete alusstruktuuride kujul. Seega, "lisaks lausete sisulise esituse tasandile peab eksisteerima veel mingi sügavam tasand, kus meie teadmised on esitatud "objektiivses" vormis, sõltumatult konkreetsetest lausetest, mille vahendusel need teadmised on saadud. See ongi tasand, kuhu laused lähevad ja kust nad tulevad. /.../ Ühelt poolt kuuluvad sellele tasandile teadmised, mis me ümbritseva maailma kohta nii või teisiti oleme saanud, teiselt poolt aga võib oletada, et samale tasandile kuuluvad ka sõnade tähendused." (Õim 1974: 60)

Ka leksikaalsest semantikast kõneldes kasutab Õim mõtet teatavast sügavamast tasandist keeles. "On ammu täheldatud, et sõnad ei jaotu keeles sugugi ühtlaselt, vaid koonduvad kindlate mõistete või teemade ümber, moodustades semantilisi välju. Näiteks sõnad, mis tähistavad inimese intellektuaalseid tegevusi – *mõtlema, arvama, uskuma, lootma, oletama, kavatsema, järeldama, tõestama* jne; või sõnad, mis koonduvad nn sotsiaalse suhtlemise teema ümber: *kiitma, ülistama, sõimama, noomima, hurjutama, kritiseerima, häbistama, ähvardama, alandama, mõnitama, teotama* jne. Samuti on hästi tuntud teiselaadsed tõigad. Näiteks on mõnes Okeaania keeles üle saja sõna erisuguste banaanide (nii erinevate liikide kui ka eri küpsusastmes olevate banaanide) tähistamiseks, samal ajal kui on küllalt keeli, mis ajavad läbi üheainsa sõnaga või milles sellist sõna üldse pole." (Õim 1974: 18–19) Ungarlastel on palju erinevaid sõnu mõiste 'viinamari' jaoks, eesti keel tunneb ainult üht ja eristab omadussõnadega kahte liiki: 'heledad viinamarjad' ja 'tumedad viinamarjad'. "Palju olulisem kui too väline rühmitumine on aga see, et sama ebaühtlus ilmneb ka sügavamal semantilisel tasemel. Siin võib leida kindlad abstraktsed semantilised struktuurid, mis on aluseks konkreetsetele väljenditele ning nende rühmadele. Nood struktuurid ei iseloomusta mitte niivõrd üht või teist konkreetset keelt, kuivõrd inimkeelt, selle semantilist ehitust üldse. /.../ Ja nende abstraktsete semantiliste struktuuride põhjal võime otsustada, missugused tegurid on inimesele tema maailmas olulised /.../" (Õim 1974: 19).

Semantika põhiteese on, et enamiku sõnade tähendused pole terviklikud üksused, vaid on analüüsitavad teatud komponentideks. "Lausete moodustamine ja mõistmine, samuti mingi lause mõistmisel saadud informatsiooni edasine töötlemine – mõtlemine laias mõttes – seisneb suurel määral just opereerimises nende elementaarsete kom-

ponentidega." (Õim 1974: 50) "...Inimesel on kalduvus organiseerida meeldejäetav informatsioon struktuuridesse, mis tuginevad ruumilistele vahekordadele. Kujutluspildid on niisuguste struktuuride otseseks näiteks. Kuid selgub, et ka abstraktsema sisulise informatsiooni säilitamisel kasutatakse sageli mingit kvaasiruumilist esitusviisi. Teatud mõttes võib sellele kinnitust leida juba keelest endast, nimelt rohketest metafooridest (aga samuti väljenditest, mis kunagi võib olla olid metafoorid, kuid nüüd on ammu lakanud seda olemast), kus mingi ruumiline seos on üle kantud hoopis muule alale: me räägime *kõrgemast* seltskonnast ja *all*ilmast; eksamil *kukutakse läbi*; mingi probleem mõeldakse *läbi* ja andmed töödeldakse *läbi*: on olemas *ülemused* ja *alamad*; mõni töö on *all*pool arvestust; on *kaugemaid* ja *lähemaid* sugulasi jne jne." (Õim 1974: 66) Oma semantikakäsitlustes väljendab H. Õim arvamust keeles olevast alusstruktuurist või sügavamast kihist.

4. Metafoorsus keeles ja mõtlemises

Edasi vaataksime metafooride, keeletekke ja žestide seoseid kui vahendeid, mis teavitavad meid "kihilisest" keelemudelist. Kõigepealt vahendame mõnesid J. Kaplinski seisukohti metafoorist ja inimteadvusest. "Metafoor kujundliku sõnakasutusena teeb temast kõikjaloleva, aga ka ähmastab metafoori mõistet. Metafoori põhifunktsioone on see, et tema abil saame nimetada asju teiste asjade kaudu. See võimaldab kirjeldatavaid asju seostada parajasti oluliste tunnuste järgi, ignoreerides teisi. Sõnade tähendused on mitmetahulised ja nii poeetilise kui ka filosoofilise mõtlemise sisuks ongi vahel eeskätt tähenduste ja tähenduskomponentide kaleidoskoopiline seostamine omavahel ikka uutes kombinatsioonides. Sellise kaleidoskoobimängu üks eesmärk on kahtlemata korra ja stabiilsuse säilitamine inimese suhtumistes." (Kaplinski 1997: 220) "Metafoor on oluline abiline ja mõndagi saab kõige täpsemalt väljendada vaid metafoorselt. Metafoor on oluline osa mõtlemisprotsessis. Võib-olla enim on ta seotud nägemismeelega. Mõtlemine on ajas, aegruumis seda, mis nägemine ruumis. Nende seos on aja ja ruumi vältimatu seose kajastus meis." (Kaplinski 1996: 74) "Meelte olemasolu loob selle, mille inimlikku avalduskuju nimetame teadvuseks – eluruumi, selles sisalduvate oluliste asjade koopia, peegelduse, von Uexkülli *sisemaailma* (*Innenwelt*) loomas; selleks eriti kujunenud närvirakkudes ja elundis – ajus. Teadvus on algselt ruumi peegeldus,

teadlikkus ruumist. Et ruum ja aeg on seotud, on selles esialgu veel pelgalt ruumilises teadvuses avardumisvõimalusi aegruumi, aja poole." (Kaplinski 1996: 77)

Metafoor kuulub eelkõige keele juurde ja on oletatavasti seotud meie mõtlemisega. Keele areng ongi kulgenud käeliselt osutuselt sümboolse tähenduse juurde ning metafoor annab juba pisut kulunud sõnale uue (tähenduse) näo. Kui metafoor kuulub keele juurde ja kui spekuleerida teemal, et keel arenes käežestidest (vt lähemalt Tenjes 2001b), siis on nii ikoonilisus, metafoorsus kui sümbol keeles täiesti mõistetavad. Käe osutavad ja viitavad žestid olid esmased vahendid suhtluses. Üleminek häälelisele kõnele toimus järk-järgult ja käe roll jäi väiksemaks. Esmane "metafoorne ülekanne" toimus siis, kui käežestide ikoonilised tähendused läksid üle häälega väljendatud tähendusteks. Neid hakati kasutama kui sümboleid, millel on kindel tähendus. Aja jooksul toimunud väljendi permanentne ülekandumine teistesse samasugustesse situatsioonidesse oli "teine metafoorne ülekanne". Õigem oleks öelda, et sellest alates toimubki "metafoorne ülekandumine" kui teatud protsess, kus sümbolid muutuvad uuteks metafoorideks ja tekivad uued tähendused. Käsi ei saa selles protsessis olla enam esmase rolli kandja, kuna ta on selle rolli delegeerinud keelele.

Ükski keel ei ole teise keelde täielikult tõlgitav ja osa verbaalseid metafoore eesti keeles pole seda nt inglise keeles. Uurides, kuidas metafoorsed väljendid ja metafoorsed žestid funktsioneerivad eesti keeles (vt Tenjes 2001a: VII: 1–15), pidin ma endalt küsima *Miks see ikkagi nii on? Kus on see koht, kus žest ja keel kohtuvad enne väljendi lausumist* (seda, et nad kohtuvad, on näidanud paljud žestiuuringud)*?*

5. Kolmetasandiline keelemudel

Vastuse võiks anda järgnev mudel keelest. Võib oletada, et keelemudel koosneb mitmest kihist: pealmisteks on kõne ja žest. Kuid vaatamata käega viipamisele kuhugi suunas ja olenemata ütlusest on veel MIDAGI – mingi psüühiline pool, tasand, mille välised avaldused on teatud konkreetsel viisil esindatud eelnevalt nimetatud kõne ja žesti tasandil.

Professor H. Õim juhtis mu tähelepanu viisile, kuidas ma võiksin eelnevalt esitatud küsimustele vastused leida. See on idee VAHEKIHIST keeles. Idee pärineb, nagu me teame, 19. s. keeleteadlaselt

W. von Humboldtilt. Keel liigendab maailma ja iga keel liigendab seda erinevalt. Keeles võib olla sügavam kiht, sisevorm Humboldti järgi. Sealt tulevad žestid ja seal sünnivad nt keelespetsiifilised metafoorid. Meie igapäevases kasutatavas keeles – verbaalses kihis – on vormistatud sõnad ja žestid. Metafoorid on küll verbaalses kihis, aga metafoor *sünnib* vahekihist.

Vahekihi olemasolust annavad aimu nt mõned sisihäälikud eesti keeles, mis tulevad keelde, kuid ei püsi seal ja teisenevad: z ja ž jäävad küll kirjakeelde, kuid kõnes muutuvad helituteks häälikuteks. Meie keelele omane astmevaheldus muudab uued sõnad samuti astmevaheldus*likeks*, nt *internet-interneti-internetti*, *šoppama-šopata-šoppan* jms. Selline sisihäälikute "rahulolematus" eesti kõnekeelega või astmevahelduse "käitumine" on teatud **sisevormi** avaldus, meie keele sisevorm.

Teatud keeleline relativism võib olemas olla, samuti võib olla mingi üldine skeem keeles. M. Johnsoni (1987) jt *image schemata* metaforiteoorias on hea idee, aga seda ei pea tuletama inglise keelest. Viidates äsjaöeldud vahekihile keeles, võib öelda, et nt keelespetsiifilised metafoorid pole keeles "kinni". Metafoor pole universaalne. Igas keeles on keelespetsiifiline kiht, millest metafoorid ja žestid sünnivad. See pole vormistatud pindkiht, vaid vahekiht. Vahekiht on ka žestidel. Vahekihis kõik murdub ja peegeldub. Seega, nagu eespool öeldud, metafoorid mitte niivõrd ei asu, vaid sünnivad vahekihist väga erinevate ja keeruliste "murdumiste" tulemusena. "Murdumine" on nagu valguse murdumine sfääril või veepiisal: valguskiir ei lähe murdudes tagasi samasse punkti, kust ta tuli. Seepärast ongi aluse leidmine metafooridele, žestidele ja üldse sõnade tähendustele nii keeruline ja huvitav. Keeles võib olla ka n.ö keele-alune süvakiht, mis võiks olla universaalne. Universaalne kiht tuleb vahekihti, murdub ja killud lendavad pindmisse keelekihti laiali. Seega, keel on mitmekihiline (Õim, isiklik vestlus). Kuidas siis leida, kuidas antud keeles universaalne väljendub, eks ole!? Keelemudeli "töölehakkamiseks" oletame, et kihtidevahelise info edastamine võib toimuda **infovoo** ehk teatud kommunikatsiooni kaudu. Sellele võib saada tuge käežestide ja keele uuringutest. Keele ja käežestide puutepunktid on seotud inimese üldise kognitiivsusega. Vastavalt inimese üleüldisele kognitiivsusele võib aluseksolev seos žesti ja sõna vahel olla *protsess* või teatud liiki *informatsioon*. Žestid võivad olla kehastunud info tõlgendused kavatsusliku ja arusaava vaimu vahel (Bouissac 2000).

Ka žestidel ja mõistetel peab olema kattuv ala. See näitab seoseid sügaval inimmõistuse või inimvaimu psühholoogilisel tasandil. Selline kattuv ala märgitseb teatud heterogeensust. Heterogeensus on inimteadvuse igipõline omadus ja selle teadvuse mehhanismile on tingimata vajalik vähemalt kahe süsteemi kohalolek, mis poleks lõpuni teineteiseks tõlgitavad (Lotman 1999: 198).

6. Kokkuvõtteks

Artiklis vaatlesime kolmetasandilist keelemudelit süvakihi, vahekihi ja vormistatud pindkihiga. H. Õim on arendanud keele vahekihi-ideega edasi W. von Humboldti mõtteid keele sisevormist. Tasanditevaheliseks liikumapanevaks jõuks ja sidujaks võib olla info ehk teatud kommunikatsioon.

Kirjandus

Humboldt, Wilhelm von 1949. Über die Verschiedenheit des menschlichen Sprachbaues. Darmstadt.

Bouissac, Paul 2000. Information, Imitation, Communication: An Evolutionary Perspective on The Semiotics of Gestures. Based on a plenary lecture of the same title given at the Conference Gestures: Meaning and Use. 1.– 4. April. Oporto, Portugal.

Johnson, Mark 1987. The Body in the Mind: The Bodily Basis of Meaning, Imagination and Reason. Chicago: University of Chicago Press.

Kaplinski, Jaan 1996. See ja teine. Tartu: Tartu Ülikooli kirjastus.

Kaplinski, Jaan 1997. Võimaluste võimalikkus. Tallinn: Vagabund.

Künnapas, Teodor 1992. Suured mõtlejad. Tallinn: Olion.

Lotman, Juri 1999. Semiosfäärist. Tallinn: Vagabund.

Robins, R. H. 1967. A Short History of Linguistics. London: Longman.

Tenjes, Silvi 2001a. Nonverbal Means as Regulators in Communication: Sociocultural Perspectives. Tartu: Tartu University Press. (Doktoridissertatsioon)

Tenjes, Silvi 2001b. Keele žestilise päritolu hüpotees. – Keel ja Kirjandus 10, 683–690; 11, 756–764.

Õim, Haldur 1974. Semantika. Tallinn: Valgus.

Eesti keele põhisõnavara operaatoritest. Katseid verbide ja kaassõnadega[1]

Ann Veismann, Ilona Tragel, Renate Pajusalu
Tartu ülikool

Sissejuhatus

Pikka aega on arvatud, et keele tähendusstruktuurid koosnevad kahest osast: sõnavarast, mis kannab leksikaalseid tähendusi, ja grammatikast, mis kannab grammatilisi tähendusi. Kognitiivses paradigmas on nende pooluste jäigast eristamisest loobutud ja rõhutatud pigem keele ühtset mõistestruktuuri, sest keel on nii leksikaalsete kui grammatiliste vahendite osas seotud inimese tunnetuse ja käitumismehhanismidega (vt nt Õim 1990). Täielikult ei saa sellisest eristusest muidugi loobuda: keele mõistete hulgas on ühelt poolt skemaatilisi mõisteid, mida eelkõige väljendab grammatika, ja teiselt poolt leksikaalselt väljendatud mõisteid, mille parimaks esindajaks on konkreetse tähendusega substantiivid. Nende vahele jääb suur hulk keele väljendusvahendeid, mis pole õieti ei leksikaalsed ega grammatilised. Käesolevas vaatlemegi ühelt poolt selliseid keele üksusi, mis on oma morfoloogiliselt distributsioonilt iseseisvad täistähenduslikud sõnad, kuid tähenduselt tunduvalt skemaatilisemad kui prototüüpne sõna. Teiselt poolt käsitleme grammatilisi sõnu, mille tähendus öeldakse selguvat alles koos täistähenduslike nimisõnadega, ent mis ise on keeles arenenud just täistähenduslikest sõnadest. Juttu tuleb keele põhisõnavara operaatoritest: keele sagedasematest ja olulisematest sõnadest, milleta kommunikatsioon on võimatu.

Värvi- ja lõhnasõnavarast lähtudes on eesti keele põhisõnavara käsitlenud U. Sutrop (2000, 1998, 1995) ja emotsioonide poolelt E. Vainik (2001). Põhisõnavara operaatoreid otsides kerkib esmalt küsimus, kas nn grammatilisi sõnu üldse põhisõnavara hulka saab lugeda. Kas saab grammatilistele sõnadele rakendada Berlini ja Kay määratlust *Põhisõna on psühholoogiliselt esiletulev, enamasti morfoloogiliselt lihtne omasõna, mis kuulub oma valdkonna prototüüpse(te) liikme(te)ga samasse klassi ning millel on viimastega ühesugune grammatiline potentsiaal. Üldiselt tähistab põhisõna mingit*

[1] Uurimistööd on toetanud Eesti Teadusfond (grant nr 4405).

objekti, omadust või nähtust ning on kasutatav kõigis asjakohastes situatsioonides (Berlin, Kay 1969: 5–7; Sutrop 1998: 61)? Määratlus annab üsna selgelt ette ka põhisõnavara vormilise külje, need on nimi- ja omadussõnad; ei verbid ega kaassõnad saa tähistada objekti, omadust või nähtust (kui jätta kõrvale verbi käändelised vormid). Samuti on verbi, kaassõna ja pronoomeni kategooriad liiga laiad ja keele-(eriti süntaksi)-põhised. Seepärast on meie projekti anaüüsi eesmärk ja objekt laiem, pigem võib seda võrrelda Ogdeni *BASIC english*'iga (vt Ogden 1933) või Anna Wierzbicka loomuliku semantilise metakeelega.

Verbide osas on põhisõnavara operaatorid määratud vastavate kriteeriumidega (vt Tragel 2001). Selles artiklis on verbide osas vaatluse all see, kas keelekasutajad (-keeleuurijad, vt punkt 3.2.3) produtseerivad samu verbe, mida kriteeriumide väljatöötamisel eeldati. Teisena analüüsitakse, missugusi tulemusi annab verbide loetelukatse. Kaassõnade osas lähtutakse nendest verbide jaoks välja töötatud tingimustest, mis kesksete kaassõnade väljaselgitamiseks sobivad.

Teiste operaatorite osas esitatakse vastavate alaliikide operaatorite kandidaadid, peamisteks tingimusteks on suur esinemissagedus ja polüseemsus, lähtutakse ka *BASIC englishi* vastavatest operaatoritest (Ogden 1933) ning Wierzbicka semantilise metakeele primitiividest (Wierzbicka 2000).

Kõigepealt esitame väljavalitud põhisõnavara operaatorid, nii nagu meie neid mõistame. Valikus oleme lähtunud eesti kirjakeele korpuse 1990. aastate ilukirjanduslike ja ajakirjanduslike tekstide põhjal koostatud lemmade sagedustabelites (Hennoste, Muischnek 2000: 213–217) saja sagedasema hulka kuuluvatest sõnadest ning neile lisanud sõnu varemkirjeldatud kriteeriumide (Tragel 2001) alusel, eelkõige lähtudes skemaatilisest tähendusest või polüseemsusest. Välja on seega jäetud substantiivid, kui suhteliselt iseseisva leksikaalse tähendusega sõnad, ja sidesõnad, relatiivpronoomenid ning pragmaatilised partiklid, kui puhtgrammatilise tähendusega sõnad. Operaatorid on liigitatud oletatava esmatähenduse järgi, nt *üks* on kogusesõnade hulgas, kuigi see funktsioneerib ka pronoomenina, ja *siis* ajasõnade hulgas, kuigi see on ka sidesõna ja partikkel. Liigitus ei lähtu jäigalt sõna morfoloogilisest klassist, vaid pigem tähendusest, sest vaadeldava sõnavararühma omaduste hulka kuulub kahtlemata mitmefunktsioonilisus. Seejärel tutvustame mõningaid katseid ja nende tulemusi, mis olid kas abiks põhisõnavara operaatorite

selgitamisel või andsid informatsiooni nende sõnade tähenduste/ esilduvuse kohta kõnelejate teadvuses.

1. Põhisõnavara operaatorid

1. 1. Verbid

Ülevaatlikkuse huvides esitame põhisõnavara operaatorite hulka kandideerivate **verbide** esialgse liigituse, mis on tehtud sünkroonilise prototüüptähenduse alusel:

a) üldverbid: *olema, tegema*;
b) liikumis- ja asendiverbid: *tulema, minema, käima, seisma, istuma*;
c) omandamisverbid (tüüpilise omandamise või loovutamise situatsiooniga seotud verbid): *saama, võtma, andma, panema, viima, tooma*;
d) modaalverbid: *pidama, võima,* (rühma täielikkust silmas pidades tuleks lisada ka *saama,* mis on loetletud eelmises rühmas);
e) algusverb *hakkama*;
f) kognitiivverbid: *nägema, tahtma, mõtlema, teadma, vaatama, kuulama, tundma, arvama*;
g) suhtlusverbid: *ütlema, rääkima, küsima, vastama*;
h) kausaalverbid *ajama, laskma, jääma.*

Neid verbe vaadeldes joonistub välja ettekujutus tüüpilisest inimelust: inimene on ja tegutseb; mõnikord liigub, mõnikord püsib paigal; omandab midagi ja loovutab midagi füüsilises ja mentaalses maailmas (kõnelemine on ju mentaalse maailma loovutamine, kognitiivverbid aga väljendavad millegi omandamist mentaalses maailmas); peale selle on ta ühiskonna liige, millega teda seovad kohustused ja lubadused. Ja kõike seda võib alustada, võib põhjustada, aga võib ka jääda äraootavale seisukohale.

Osa loetletud verbidest kuulub eesti keele tuumverbide hulka (rühmad a–e (v.a *seisma* ja *istuma*), rühmast f esimesed kolm ning suhtlusverb *ütlema*. Nende kohta oleme teinud eraldi katse, et selgitada sellise tuumverbide kogumi paikapidavust (vt p 3.2.3).

Eesti keele kesksesse sõnavarra kuuluvate verbide määratlemise puhul on muuhulgas huvitav jälgida, millised verbid uurijate tähelepanu on köitnud. Erinevatelt lähtealustelt on uuritud näiteks verbe *saama, tulema* ja *pidama* (Õim 1965), verbi *pidama* (Erelt 2001), verbe *ajama, panema, laskma* (Kasik 1999, 2001), verbe *hakkama* ja *saama* (Metslang 1994), verbi *seisma* (Pajusalu 2001)

2. 2. Pronoomenid

Sagedasimad pronoomenid on
a) isikupronoomenid: *mina, sina, tema*;
b) demonstratiivid: *see, nii, selline*;
c) refleksiivpronoomenid: *oma, ise*;
d) indefiniitpronoomenid: *miski, keegi, mingi*;
e) *sama*.

Pronoomenid kordavad, aga teatud mõttes ka täiendavad verbide poolt loodud maailmapilti. Inimene tegutseb kas enesekeskselt ise (olgu siis nii, et subjekt domineerib ISE üle või et ISE domineerib subjekti üle või tagatipuks hoopis ISEga partner olles, vt Õim 2001) või koos teiste inimestega, keda ta olenevalt nende suhtluslikust staatusest peab "sinadeks" või "temadeks", peale selle ümbritsevad inimest osutatavad, seega konkreetsed objektid, aga ka umbmäärased "keegid" ja "mingid". Inimese on elu ka pidev kordumine, võrdlemine ja äratundmine, ikka ja jälle tuleb esile "sama".

Ogdeni põhisõnavara operaatorite hulgas on oodatult kõik ainsuslikud isikupronoomenid ja demonstratiivid *this, that, such* ja *so*, millele eesti keeles vastavad *see* (*too* on kirjakeeles liiga haruldane), *selline* ja *nii*. Eesti *sama* vaste on loetelus samuti olemas, kuid omaduste, mitte operaatorite hulgas. Eesti indefiniitpronoomenite vasteks on Ogdenil *some,* teataval määral võib nende esindajaks lugeda ka indefiniitset artiklit *a*. Eesti *ise* ja *oma* vasteks on Ogdenil *self*, mis on küll tema loetelus rubriigis "things". Seega on Ogdeni põhisõnavara operaatorid pronoomenite osas praktiliselt samad, mis eesti keele sagedasemad pronoomenid.

Wierzbicka semantiliste universaalide hulgas on isikupronoomenitest ainult I ja YOU, mis on ka loogiline, sest tegemist pole ju mitte tegeliku keele, vaid minimalistliku metakeele sõnadega. Tõepoolest on kõik muud isikupronoomenid kirjeldatavad nende kahe ja indefiniitsete SOMEONE ning SOMEBODY abil. Kvantifikaatorite hulgas on ka indefiniitne SOME, millele teatud kontekstides vastavad eesti keeles indefiniitsed pronoomenid *mingi* või *keegi*. Demonstratiividest on esitatud samuti ainult üks, lähedaleviitav THIS. Refleksiivsust pole samuti peetud eraldi kategooriaks, kuid see mahub ilmselt pronoomeni I funktsioonide hulka. Eraldi kategooriana on esitatud THE SAME.

2. 3. Suhtesõnad

Sagedasimad ja kesksemad ruumi- ja ajasõnad (kui tavalisemad suhtesõnad) on

a) demonstratiivid: *siin, seal,*

b) ajasõnad: *siis, nüüd, praegu, juba, pärast,*

c) muud ruumilisi suhteid väljendavad adverbid ja kaassõnad: *alla, peale, üle, vastu, läbi, eest.*

Inimese eksistentsile on olemuslik veel olemine ruumis ja ajas, põhisõnavara operaatorid määratlevad ruumiliselt nii lähema kui kaugema, ajaliselt aga ainult deiktilise nullpunkti ja järgnevuse. Osaliselt on see tingitud muidugi sellest, et aega tajutaksegi ruumi kaudu, nii et kaugemad hetked nimetatakse tihti ruumisõnade abil. Võrdluseks võib tuua, et Ogdeni *BASIC english* sisaldab näiteks järgmisi suhtesõnu (Ogden nimetab neid operaatoriteks, nagu ka verbe jm): *about, across, after, against, among, at, before, between, by, down, from, in, off, on, over, through, to, under, up, with.* Kokku 20, mida on võrdlemisi palju. Wierzbicka on oma loomulikku semantilisse metakeelde viimati (2000) arvanud sellised ruumi- ja ajasõnad nagu AFTER, BEFORE, ABOVE, BELOW, INSIDE, FAR, NEAR, SIDE (Wierzbickat sõnaliigid ei huvita, ta lähtub sõnade sisust või semantilisest funktsioonist).

2. 4. Kogusesõnad

Sagedasimad kogusesõnad on *kõik, veel, enam, ainult, pool, palju, mõni, väga, kogu, iga* ja arvsõnad: *üks, teine, kaks, esimene.*

Kogusesõnade suur hulk sagedaste sõnade hulgas on mõneti üllatav ja osutab ilmselt kvantifikatsiooni olulisusele.

Ogdenil esinevad kogusesõnadest *about, all, any, every, no, other, some, almost, enough, even, little, much, only, very.* Mõnevõrra üllatavalt puudub vaste sõnale *veel* (*more*), mis on ilmselt siiski oluline operaator. Arvsõnu Ogdenil loendis ei esine.

Wierzbicka loetelus on kvantifikaatorite all toodud ONE, TWO, SOME, MANY/MUCH, ALL ja intensifikaatorite all VERY ja MORE. Eesti keele sagedastest kogusesõnadest on selles loetelus ilma vasteta seega *enam, ainult, pool, iga, esimene* ja *teine.* Need sõnad on Wierzbicka metakeeles kirjeldatavad olemasolevate primitiivide abil.

3. Katsed kaassõnade ja verbidega

3.1. Kaassõnad kui tüüpilised suhtesõnad

Kuigi, nagu eelpool öeldud, määratakse kognitiivset esilduvust harilikult täistähenduslike substantiivide puhul, üritasime siiski proovida, kas õnnestub kaassõnade (kaassõna kui suhtesõna sisaldab nii ruumiaja kui muude suhete väljendusi) ja verbide puhul selgitada välja kognitiivne esilduvus. Selleks viisime läbi loetelukatse. Katsealustel paluti kirjutada kolme minuti jooksul kõik pähe tulevad kaassõnad ja seejärel lehe teisele poolele kolme minuti jooksul tegusõnad. Täpsustuseks öeldi, et kaassõnad võivad olla nii pre- kui postpositsioonid ning käivad koos nimisõnaga. Kirjutajad olid esimese aasta filoloogiatudengid. Kirjutajaid kokku oli 77, neist kaassõna puhul läks arvesse 45 loetelu, 24 oli kas jätnud lehe tühjaks või kaassõna täiesti vääriti mõistnud, 8 lehte olid kas liialt soditud või tundusid muul põhjusel ebapiisavalt spontaansed. Lisaks on võetud kaassõna katsetulemustesse veel kahe katseisiku andmed, (mõlemad on umbes 30-aastased filoloogid) kellega katse tehti telefoni teel, paludes loetleda pähetulevaid kaassõnu 3 minuti jooksul.

Loetleti kokku 68 erinevat kaassõna, neist 17 esines vaid ühe korra. Esiletuleku indeks arvutati neile sõnadele, mis esinesid vähemalt viies loetelus. Indeksi arvutamise valemi osas võeti eeskuju U. Sutropi artiklist (Sutrop, ilmumas), kus selleks on pakutud S=F/(N mP). Valemis on S indeks, F arv, mitu korda sõna loeteludes esineb, N kõigi loetelude arv (=47), mP sõna keskmine astak (loetelude järgi, kus see sõna esines). Tulemused on esitatud tabelis 1.

Tabel 1. Kaassõnade loetelukatse tulemused

A – astak eelneva veeru näitaja järgi, S – sagedus, kA – keskmine astak; indeks – kognitiivse esiletuleku indeks

Sõna	s	A	kA	a	Indeks	a
peal	36	1–2	4,1	2	0,186	1
all	36	1–2	4,3	3	0,177	2
kõrval	33	3	5,788	10	0,121	3
juures	27	4	4,96	6	0,116	4
sees	26	5	6,5	15–16	0,085	5
taga	24	6	6,375	14	0,08	6
ees	21	7	6,238	13	0,072	7

Sõna	s	A	kA	a	Indeks	a
alla	14	9	4,714	4	0,063	8
koos	16	8	5,56	7	0,061	9
ilma	11	12	4,82	5	0,048	10
peale	13	10	6	11	0,046	11
üle	12	11	5,75	9	0,044	12
taha	5	20–23	3,8	1	0,0279	13
mööda	8	16	6,125	12	0,0278	14
kohal	9	13–15	7,22	19	0,0265	15
läbi	9	13–15	7,67	21	0,025	16
sisse	7	17	6,71	17	0,022	17
küljes	6	18–19	6,5	15–16	0,0196	18
pärast	5	20–23	5,6	8	0,019	19
vahel	9	13–15	10,44	23	0,018	20
keskel	6	18–19	7,33	20	0,017	21
kõrvale	5	20–23	6,8	18	0,0156	22
ümber	5	20–23	9,4	22	0,011	23

Nagu näha, on suurima indeksiga lokatiivsed kaassõnad (*peal, all, kõrval, juures, sees, taga*), seejärel tulevad nende samade kaassõnade latiivsed vasted (*alla, peale, taha; sisse*) ja mõned teised liikumisega seotud kaassõnad (*üle, mööda; läbi*). Vastustelehtedest 7-l olidki ainult lokatiivsed vormid, lisaks leidus 3 lehte, kus kõigi lokatiivide seas esines üks muud sorti kaassõna. Ainult latiivseid vorme olid loetlenud kolm küsitletut (+ üks ühe muu vormiga). Separatiivsete kaassõnade esinemine oli tunduvalt harvem. Tüüpilise (prototüüpse) kaassõnana meenutatakse niisiis eelkõige lokatiive. Kui võrrelda esiletulevaid sõnu sagedaste suhtesõnadega (Hennoste ja Muischneki andmetel mahuvad 100 sagedasema sõna hulka *välja, üle, vastu, tagasi, läbi, eest, pärast*, ent kaas ja määrsõnade sagedusloendeid on ka mõnevõrra teistsuguseid), siis kattuvus kuigi suur ei ole, vaid *üle*, *läbi* ja *pärast*. Sellel on ilmselt omad põhjused, sagedus kirjakeeles ei pruugi määrata esilduvust. Sellepärast tasub vaadelda korraks kaassõnu ka polüseemsuse aspektist. P. Palmeos on oma uurimuses loetlenud kõige enam tähendusi (üle kolme tähenduse, pre- ja postpositsioonile kokku, kui sõna mõlemana esineb) järgmistele kaassõnadele: *alla, eest, kohta, läbi, peale, pärast, vastu, üle*. Kõik need kaassõnad võivad esineda ka adverbina, mis suuren-

dab nende polüseemsust veelgi. Kaheksast väljatoodud sõnast kõige tähendusrikkam on *peale*, talle järgnevad *üle* ja *vastu*. Nendest järgmised on *eest* ja *läbi*. Siin on kattuvus sagedaste kaassõnadega tunduvalt suurem (*eest, läbi, pärast, vastu, üle*). *Välja* rangelt võttes polegi (Palmeose järgi ka mitte) kaassõna, *tagasi* sattumine sagedaste hulka on raske seletada, võimalik, et selle põhjuseks on sage esinemine verbi partiklina (e afiksaaladverbina).

Kui võrrelda esiletuleku indeksit ja polüseemsust, siis jäävad ühistena sõelale *alla, peale, üle, läbi* ja *pärast*. Polüseemsetest leidis *vastu* loetelukatses nimetamist vaid ühel korral, *eest* ja *kohta* aga üldse mitte. See on hõlpsasti seletatav nii, et separatiivne vorm ei tule kolmeaspektiliselt orienteeritud sõnadest esimesena pähe, olgu ta kuitahes mitmetähenduslik, ning *kohta* on ruumisuhte märkimiseks (mis valdavalt kõrge esilduvusindeksiga sõnad olid) liiga üldise tähendusega. Kõigi kolme vaadeldava loendi ühisosaks on *üle, läbi* ja *pärast*. Neist *pärast* on erakordne seetõttu, et tal puudub ruumitähendus; kõik ülejäänud polüseemsed ja sagedased suhtesõnad väljendavad ühe tähendusena ruumisuhteid, samuti on loetelukatse tulemustes valdavad ruumitähendusega kaassõnad. Operaatorite hulka on kõike kokku võttes mõistlik lugeda *alla* ja *peale* kui esileküündivaimad n.ö kohasõnad (*peale* peaks olema tegelikult ka sagedane, vt kasvõi nt murdekorpuse andmeid Lindström *et al* 2000). Nii *peale* kui *alla* kuuluvad ka varakult omandatavate suhtesõnade hulka (vt Vija 2001; Vider 1996; Kõrgvee 2001). Seejuures märgivad nii Vija kui Kõrgvee, et lapse jaoks ei pruugi nende kahe sõna tähendus kohe päris selge olla, tarvitatakse kas ühte neist mõlema tähenduses või üht ainult kindlates konstruktsioonides. Kohasõnade kõrvale võiks seada nn trajektoori või liikumistee sõnad *üle* ja *läbi*, mis igasuguste kriteeriumide järgi esile tõusevad (olles võrdlemisi varased ka Kõrgvee uuritud lapse keeles). Erandlikuna peaks juurde võtma samuti igati esileküündiva ajasõna *pärast*. Teatud kõhklustega võiks neile lisada veel ka *eest* ja *vastu*.

3.2. Katsed verbidega

3.2.1. Kognitiivse esiletuleku katse

Kaassõnade loetelukatsega samades tingimustes viidi läbi loetelukatse ka verbide esilduvuse selgitamiseks. Vastajatel paluti 3 minuti jooksul kirjutada tegusõnu. Verbide osas olid arvestatavad kõik 77

vastustelehte, millel oli kokku 2772 analüüsitavat ühikut, keskmiselt oli üks vastaja kirjutanud 36 verbi, kõige rohkem oli 57 ja kõige vähem 12 verbi. Erinevaid tegusõnu esines 801. Meid huvitavad katse tulemused kahes plaanis. Esiteks vaatleme, milliseid verbe nimetati kõige sagedamini ning seejärel teeme kindlaks, kas eesti keele tuumverbideks pakutud verbid (Tragel 2001) loeteludes esinevad ning millised on nende kognitiivse esiletuleku indeksid.

Tabelis 2 on 21 kõige sagedamini esinenud verbi, nende keskmised esinemispositsioonid ning kognitiivse esiletuleku indeksid.

Tabel 2. Verbide loetelukatse tulemused

A – astak; S – sagedus; KA – keskmine astak, indeks – kognitiivse esiletuleku indeks

A	Verb	S	A	Verb	KA	A	Verb	Indeks
1	kirjutama	57	1	jooksma	7,69	1	jooksma	0,0929
2	jooksma	55	2	tegema	7,85	2	laulma	0,068
3	sööma	47	3	olema	8,11	3	tegema	0,0662
4	magama	45	4	laulma	8,21	4	kirjutama	0,0610
5	vaatama	43	5	sööma	10,7	5	sööma	0,057
	laulma	43	6	magama	11,6	6	magama	0,0503
7	tegema	40	7	kirjutama	12,1	7	olema	0,0449
8	naerma	39	8	joonistama	12,17	8	vaatama	0,0389
9	lugema	36	9	õppima	12,32	9	istuma	0,0366
10	istuma	35	10	istuma	12,43	10	lugema	0,0357
11	rääkima	32	11	jooma	12,75	11	jooma	0,0326
	jooma	32	12	lugema	13,11	12	naerma	0,0317
	hüppama	32	13	tantsima	13,77	13	joonistama	0,0309
14	tantsima	31	14	rääkima	13,78	14	rääkima	0,0302
15	joonistama	29	15	armastama	13,85	15	õppima	0,0295
16	tulema	28	16	vaatama	14,37	16	tantsima	0,0292
	minema	28	17	hüppama	15,78	17	hüppama	0,0263
	õppima	28	18	naerma	16	18	armastama	0,0244
	olema	28	19	tulema	16,18	19	tulema	0,0225
20	armastama	26	20	minema	16,32	20	minema	0,0223
	pesema	26	21	pesema	25,88	21	pesema	0,013

Lisaks neile verbidele esinesid üle kümne korra veel järgmised verbid: *nutma, kõndima* (25 korda), *mängima* (24), *nägema* (23), *kuulama, mõtlema* (21), *seisma, ujuma* (19), *jalutama* (16), *karjuma,*

koristama, ostma (15), *õmblema, kuduma, maalima* (14), *saama, elama, leidma* (13), *ehitama, surema* (12), *kallistama, keetma, lõikama, müüma, puhastama* (11).

Situatsioonist tingitult on ootuspäraselt sagedased katse ajal toimuvate tegevuste nimetamine: katseisikud istusid ja kirjutasid. Katseisikute sotsiaalne staatus (üliõpilane) on tõenäoliselt põhjustanud verbide *õppima* ja *lugema* rohke esinemise (*õppima* esines 28 korda, *töötama* näiteks ainult 12 korda). Ootuspärane on ka nn eluliste tegevuste – söömise, joomise, magamise, pesemise sage nimetamine. Teatud mõttes eluline "tegevus" on ka armastamine, eriti kui arvestada katseisikute ealisi ja soolisi näitajaid (valdavalt oli tegemist noorte naisterahvastega). Prototüüpse tegevusega võiks ilmselt põhjendada verbide *jooksma, tantsima* ja *hüppama* sagedast nimetamist. Võimalik, et paarikutena on meenunud *laulma* (*tantsima*) ja *joonistama* (*kirjutama*, vrd keskmist esinemispositsiooni). Kognitiivverbidest esineb *vaatama*, nägemistaju eelistus teistele tajudele on igati põhjendatud. Verbi *laulma* sage esinemine (43 korda) ja eriti kõrge kognitiivne esiletulek (indeks 0,068, *jooksma* järel teisel kohal) on raskesti seletatav, eespool pakutud paarikulisus *tantsima*ga ei saa siiski olla ainus põhjus – need verbid ei esinenud sugugi alati kõrvuti (seda osutab ka erinevus keskmise positsiooni osas: *laulma* keskmiselt 8. ja *tantsima* 14. kohal). Paarikulisus on aga selgest märgatav verbide *naerma* ja *nutma* puhul – loetelude esimeses pooles esinesid nad enamasti koos, *naerma* esines lisaks üksi veel 14 korda, tavaliselt loetelu lõpuosas. Naermine, nutmine ja armastamine esindavad niisiis selle katse põhjal eestlase emotsionaalsete tegevuste esiletulevamat osa. Samale tulemusele on eesti keele emotsioonisõnavara uurides jõudnud E. Vainik (Vainik 2001: 115), tema katse põhjal on sagedamini nimetatud emotsioone väljendavad verbid *naerma* (kognitiivse esiletuleku indeks 0,104), *nutma* (0,081) ja *armastama* (0,101) (samas, 123).

Tuumverbidest esinesid 21 sagedasema tegusõna hulgas *tegema, tulema, minema, olema*. Kognitiiv- ja suhtlusverbidest *vaatama* ja *rääkima*, asendiverbidest *istuma* (*seisma* esines 19 korda).

Sagedusloendite tipus olevad verbid on enamasti väga polüseemilised (vt näiteks www.cl.ut.ee/ee/tulemusi/sag_lem_1000.kogu) ning olulisimaid sageduse põhjusi ongi see, et neid kasutatakse väga paljudes eri tähendustes. Siinse loetelukatse tulemused näikse osutavat, et nimetamisel ei ole polüseemsusel erilist tähtsust, sama

selgus ka kaassõnade katsest. Kognitiivselt esilduvamad on pigem ühemõttelisemad ning katsesituatsioonis esil olevad tegevused, mida kajastavad vastavad verbid.

3.2.2. Tuumverbid jt põhisõnavara operaatorverbid loetelukatses

Vastavalt tuumverbi määratlusele (Tragel 2001: 169) nimetame tuumverbideks tegusõnu, mida kasutatakse grammatilistes funktsioonides ja/või mis väljendavad üldisi mõisteid. Loetelukatset on eesti keele kohta kasutatud suhteliselt selgepiiriliste valdkondade (nagu värvid või maitsed) põhisõnade väljaselgitamiseks (vt nt Sutrop 1995, 1998). Tegusõnade puhul on loetelukatse kasutamise puhul tegemist eksperimendiga, mille tulemuste põhjal on võimalik loetleda eesti keele tegusõnade valdkonna kõige esiletulevamad sõnad. Tegusõnad ei ole keelekasutaja jaoks kategooria, eraldi määratletav valdkond. Kategooriana on mõeldav näiteks TEGEVUS, kuid selline määratlus välistaks ilmselt üldse skemaatilise sisuga verbide nagu *ajama* või *olema* kaasamise. Niisiis tuleb käesoleva katse tulemuste esitamisse suhtuda eelkõige kui eksperimentaalsetel kaalutlustel tehtud temaatilisse kõrvalepõikesse.

Tabelis 3 on kognitiivse esilduvuse testimiseks korraldatud katse tulemused tuumverbide osas.

Tabel 3. Tuumverbid loetelukatses

A – astak; S – sagedus; KA – keskmine astak; indeks – kognitiivse esiletuleku indeks

A	Verb	S	A	Verb	KA	A	Verb	Indeks
1	tegema	40	1	tooma	6,36	1	tegema	0,0662
2	minema	28	2	tegema	7,85	2	olema	0,0448
	olema	28	3	olema	8,11	3	pidama	0,0354
	tulema	28	4	pidama	11	4	tulema	0,0225
5	saama	13	5	viima	12,2	5	minema	0,0223
6	käima	10	6	saama	15,23	6	tooma	0,0163
	panema	10	7	panema	15,4	7	saama	0,0111
8	andma	9	8	võtma	15,57	8	panema	0,0084
9	tooma	8	9	tulema	16,18	9	käima	0,0073
10	hakkama	7	10	minema	16,32	10	võtma	0,00583
	võtma	7	11	käima	17,9	11	andma	0,00581

A	Verb	S	A	Verb	KA	A	Verb	Indeks
12	laskma	5	12	andma	20,11	12	viima	0,0053
	viima	5	13	ajama	22,3	13	hakkama	0,0037
14	ajama	3	14	hakkama	24,71	14	jääma	0,0033
	pidama	3	15	jääma	27,5	15	laskma	0,0023
16	jääma	2	16	laskma	28,6	16	ajama	0,0017
17	võima	–	17	võima	–	17	võima	–

Esimesena hakkab silma verbi *võima* puudumine. 77 katseisikust ei ole seda mitte keegi mitte kordagi nimetanud (tuletame meelde, et kokku esines 801 eri verbi!). Ilmselt on seletus *võima* suhteliselt puhtas modaaltähenduses – modaalsus ei ole niisiis ilmselt iseseisva tegevusena tajutav ja seetõttu tegusõnade nimetamise katses puhtmodaalse tähendusega abiverb ei kajastugi. *võima* on keelekasutaja jaoks sedavõrd "sisutühi" või üldise tähendusega (vrd kaassõnadest *kohta*, vt eespool), et iseseisva tegusõnana mainimist ei leiagi.

Eespool juba tõdesime, et loetelukatses tulevad esile konkreetset ja katsesituatsioonis aktuaalset tegevust tähistavad verbid (*jooksma, kirjutama, istuma*). Tuumverbide osas on siiski üllatav see, kui vähe verbe, millel konkreetne tähendus täiesti olemas, siiski mainiti. Näiteks *panema* esines 10 korda, s.o ainult 13%-l vastajaist, *võtma* 7 korda, s.o 9%-l. Võib oletada, et üks põhjus on selles, et tuumverbide prototüüpsete tähendustega märgitavad tegevused on nii igapäevased, tavalised ja rutiinsed, et nende keeleline avaldamine ei tundugi vajalikuna. Seda võib võrrelda laiemas plaanis keeles kajastuvaga: meie harilikud asendid, staatused ja olekud omavad palju vähem keelelisi väljendusvahendeid kui ebaharilikud (nt *ma olen inimene* või *ma seisan* ei ole ilmselt sugugi nii igapäevased kui nt *ma olen väsinud* või *ma pikutan*).

3.2.3. Tuumverbide kriteeriumide paikapidavuse test

Tuumverbide määratlemise jaoks välja töötatud kriteeriumide testimiseks korraldati katse. Selleks tutvustati osavõtjatele eelnevalt nimetatud kriteeriume, nad said käsilehe sisuliste ja formaalsete kriteeriumite loetelu ja näidetega (vt lisa ja Tragel 2001: 152–156). Seejärel paluti neil viie minuti jooksul kirjutada lehele 10–20 verbi, mis nende arvates võiksid tuumverbide hulka kuuluda. Katseisikutelt küsiti veel nende staatust ning seotust verbide uurimise ja tuum-

verbide teemaga. Seosed nende näitajatega jäävad seekord analüüsimata. Katse toimus 9. 6. 2001 Kütiorus TÜ ja TPÜ magistrantide ja doktorantide ühisseminaril. Katseisikuid oli 34, neilt saadi 30 arvestatavalt täidetud vastust. Keskmiselt oli üks vastaja kirjutanud 11 verbi, kõige rohkem oli 20 ja kõige vähem 7 verbi (kõik kokku 342 vastusühikut), kokku esines 67 eri verbi. Vastustest registreeriti verb ja selle positsioon (astak), tulemuste esitamisel tabelis 4 kasutatakse vähemalt kuus korda esinenud verbe.

Tabel 4. Tuumverbide katse

A – astak, S – sagedus, KA – keskmine astak.

A	Verb	S	A	Verb	KA
1.	tulema	26	1.	olema	2
	andma	26	2.	saama	4
3.	minema	25	3.	andma	4,7
	saama	25	4.	tegema	5,2
5.	olema	24	5.	panema	5,3
6.	tegema	20	6.	tulema	5,7
7.	võtma	18	7.	hakkama	6
8.	hakkama	13		minema	6
9.	panema	12	9.	võtma	6,2
	tooma	12	10.	tahtma	7
11.	ütlema	10		käima	7
12.	võima	9	12.	võima	7,1
	viima	9	13.	ütlema	7,8
	vaatama	9	14.	tooma	8,75
15.	nägema	8	15.	pidama	8,8
16.	pidama	6	16.	vaatama	9
	tahtma	6	17.	mõtlema	9,3
	mõtlema	6	18.	nägema	9,5
	käima	6	19.	viima	9,7

Vähemalt kaks korda esinesid veel verbid *sööma* ja *jooma* (5 korda); *jääma* ja *elama* (4); *rääkima, ajama, küsima, vastama, kuulama* ja *töötama* (3) ning *laskma, tundma, lööma, seisma, jätma, laulma, kirjutama, leidma* ja *muutuma* (2).

Katse tulemused osutavad, et tuumverbide kriteeriumide lühitutvustuse põhjal on katseisikud jõudnud samade järeldusteni, mida

uurija kriteeriume välja töötades eeldas. Loetletud verbide hulgas esinesid kõik praegused tuumverbide kandidaadid, ootuspäraselt kõige vähem *laskma* (2 korda), *ajama* (3) ja *jääma* (4). Väärtuslik materjal edasimõtlemiseks on "uued kandidaadid", nt *ütlema* (esines 10 korda), *vaatama* (9), *nägema* (8), *tahtma* ja *mõtlema* (6), aga ka *rääkima, küsima* ja *vastama* (3). Näiteks verbi *vaatama* grammatiline funktsioon (1. sisuline kriteerium) avaldub kasutamises suulises kõnes diskursuspartiklina (*vata tead kus seal on see `Kuum `hind see Emil `Rutiku `teeb, sa pead seal `välja astuma*). Need verbid on kaalutlemisel põhisõnavara operaatorite hulka kaasamisel kognitiiv- ja suhtlusverbide rühmas (vt ka loetelukatse tulemused).

Kokkuvõte

Käesolevas artiklis üritasime määratleda eesti keele põhisõnavara operaatorid, nii nagu meie neid mõistame. See tähendab, et tegemist pole harilikult põhisõnavaraks loetud täistähenduslike nimisõnadega, vaid nende sõnadega, milleta keel/kommunikatsioon inimeste vahel ei toimiks. Sellised sõnad võivad tihti olla leksikaalsete ja grammatiliste vahendite piirimail ning nende tähendus on skemaatiline.

Verbidest on juba varem välja toodud need, mis võiks kuuluda eesti keele tuumverbide hulka: *olema, tulema, minema, saama, andma, võtma, panema, võima, pidama, tegema, käima, jääma, viima, tooma, ajama, hakkama, laskma*. Sageduse ja kognitiivse esilduvuse järgi võiks neile lisada veel liikumis- ja asendiverbidest *seisma* ja *istuma*, kognitiivverbidest *vaatama* (ja võibolla ka *teadma, nägema, kuulama, tahtma, mõtlema, tundma, arvama*), suhtlusverbidest *rääkima* (ja võibolla ka *ütlema, küsima, vastama*). Läbiviidud loetelukatse osalt kinnitas varasemaid andmeid, osalt lisas uusi vaatlusväärseid verbe ja osalt näitas, nagu arvata oli, et verbi kategooria on liiga lai loetelukatse täisväärtuslikuks arvessevõtmiseks.

Pronoomenite hulgast on põhisõnavara operaatoritena sõelale jäänud *mina, sina, tema, see, nii, selline, oma, ise, miski, keegi, mingi, sama*. Nende olulisust kinnitab peamiselt sagedus ja lai kasutusala. Kuid muuhulgas ka vastavate sõnade esinemine A. Wierzbicka primitiivide hulgas ja Ogdeni loetletud inglise keele operaatorite seas. Loetelukatset nii grammatiliste sõnadega läbi viia ei õnnestu. Sama võib öelda ka **kogusesõnade** kohta, millest välja on valitud *kõik, veel, enam, ainult, pool, palju, mõni, väga, kogu, iga, üks, teine, kaks, esimene*.

Aja-, ruumi ja muid **suhteid väljendavatest sõnadest** võib põhisõnavara operaatorite hulka lugeda *siin, seal, siis, nüüd, praegu, juba, pärast, alla, peale, läbi, üle* ja võibolla ka *eest* ja *vastu*. Ka nende puhul võib osaliselt toetuda loetelukatse tulemustele, ent arvesse tulevad eelkõige siiski polüseemsus, sagedus ja varane omandamine.

Kirjandus

Berlin, B.; Kay, P. 1969. Basic Color Terms: Their Universality and Evolution. Berkeley: University of California Press.

Erelt, Mati 2001. Some notes on the grammaticalization of the verb *pidama* in Estonian. – Estonian Typological Studies V. Publications of the Department of Estonian of the University of Tartu 18. Ed. by M. Erelt. Tartu. 7–25.

Hennoste, Tiit; Muischnek, Kadri 2000. Eesti kirjakeele korpuse tekstide valiku ja märgendamise põhimõtted ning kahe allkeele võrdluse katse. – Arvutuslingvistikalt inimesele. Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 1. Toim. T. Hennoste. Tartu. 183–218.

Kasik, Reet 2001. Analytic causatives in Estonian. – Estonian Typological Studies V. Publications of the Department of Estonian of the University of Tartu 18. Ed. by M. Erelt. Tartu. 77–122..

Kasik, Reet 1999. Ajab segadusse: eesti ja soome keele leksikaalsetest erinevustest. – Lähivertailuja 10. Tampere: Tamperen yliopiston suomen ja yleisen kielitieden laitos.

Lindström *et al* 2000 = Lindström, Liina; Lonn, Varje; Mets, Mari; Pajusalu, Karl; Teras, Pire; Veismann, Ann; Velsker, Eva; Viikberg, Jüri 2000. Eesti murrete korpus ja kolme murde sagedasema sõnavara võrdlus. – Keele kannul. Pühendusteos Mati Erelti 60. sünnipäevaks. Tartu Ülikooli eesti keele õppetooli toimetised 17. Toim. R. Kasik. Tartu. 186–211.

Metslang, Helle 1994. Temporal relations in the predicate and the grammatical system of Estonian and Finnish. Dissertation. Oulun yliopiston suomen ja saamen kielen laitoksen tutkimusraportteja 39.

Ogden, C. K. 1933. Basic English. A General Introduction with Rules and Grammar. Psyche Miniatures. General Series No. 29. London: Kegan Paul, Trench, Trubner & Co.

Pajusalu, Renate 2001. The polysemy of *seisma* 'to stand': multiple motivations for multiple meanings. – Papers in Estonian Cognitive Linguistics. Publications of the Department of General Linguistics 2. University of Tartu. Ed. by I. Tragel. Tartu. 170–191.

Sutrop, Urmas 1995. Eesti keele põhivärvinimed. – Keel ja Kirjandus 12, 797–808.

Sutrop, Urmas 1998. Basic temperature terms and subjective temperature scale. – Lexicology 4:1, 61–104.

Sutrop, Urmas 2000. Basic terms and basic vocabulary. – Estonian Typological Studies IV. Publications of the Department of Estonian of the University of Tartu 14. Ed. by M. Erelt. Tartu. 118–145.

Sutrop, Urmas 2001. List task and a Cognitive Salience Index. – Field Methods 13:3, 263–276.

Sutrop, Urmas ilmumas. Loetelukatse ja kognitiivse esiletuleku indeks. – Teoreetiline keeleteadus Eestis 2001. Tartu Ülikooli üldkeeleteaduse õppetooli toimetised.

Sweetser, Eve 1988. Grammaticalization and semantic bleaching. – Berkeley Linguistics Society, Proceedings of the 14th Annual Meeting. 389–405.

Tragel, Ilona 2001. On Estonian core verbs. – Papers in Estonian Cognitive Linguistics. Publications of the Department of General Linguistics 2. University of Tartu. Ed. by I. Tragel. Tartu. 145–169.

Vainik, Ene 2001. Eestlaste emotsioonisõnavara. Magistritöö. Tallinn: Eesti Keele Instituut.

Vider, Kadri 1995. 2–3-aastaste eesti laste sõnavara. Tartu. Bakalaureusetöö. Käsikiri eesti keele õppetoolis.

Vija, Maigi 2000. Ühe eesti lapse keeleline areng vanuses 1,5–2,0. Tartu. Bakalaureusetöö. Käsikiri eesti keele õppetoolis.

Wierzbicka, Anna 2000. Universal semantic primitives as a key to lexical semantics (The case of emotions). – Presentation at the conference Das Wort: Strukturen und Konzepte. Kommunikation 2000. Philipps-Universität Marburg 14.2.2000.

Õim, Haldur 1965. *tulema, saama* ja *pidama* tähenduste strukturaalne analüüs. – Keel ja struktuur. Töid strukturaalse ja matemaatilise lingvistika alalt. Tartu Riiklik Ülikool (Eesti keele kateeder). Tartu. 27–45.

Õim, Haldur 1990. Kognitiivse lähenemise võimalusi keeleteaduses. – Akadeemia 9, 1818–1838.

Õim, Haldur 2001. Is there a folk theory of Self. The case of Estonian ise and enda~enese. – Papers in Estonian Cognitive Linguistics. Publications of the Department of General Linguistics 2. University of Tartu. Ed. by I. Tragel. Tartu. 7–21.

Lisa: Tuumverbide määratlemise kriteeriumid

I SISULISED
1. Grammatiline funktsioon (formaalselt avaldub mh süntaktilise iseseisvuse vähenemises või kadumises)
2. Üldmõistelisus (testib mh nt asendatuvus: nendega saab asendada paljusid hierarhiliselt madalamate tasandite verbe, vrd nt *Ulata soola!* - *Anna soola!*)
3. Konstruktsiooniskeem, millest võib kujuneda verbist lähtuv tähendusega konstruktsiooniskeem, kuhu sobivad teised sarnase tähendusega verbid (vrd *nt Mari annab Jürile raamatu. – Mare faksib KK-le artikli.* Konstruktsioon: [S + V + N:ALL + O])
4. Skemaatiline tähendus (vrd nt nn andmisskeem, millesse kuuluvad ANDJA, ANTAV ja SAAJA)
5. Polüseemilisus

II FORMAALSED
1. Korpuste andmete põhjal sagedusloendite tipus
2. Lühike, lihtne omasõna või väga vana laensõna
3. Lapsed omandavad esimeste verbide hulgas

III UNIVERSAALSUS
Need verbid on maailma keeltes suhteliselt universaalselt nn funktsioonisõnad (väljendavad grammatilist tähendust) ning esineb arvukalt näiteid nende grammatikaliseerumisest.

Expressive synonyms: some implications for bilingual lexicography

Enn Veldi
University of Tartu

1. Introduction

Language is a heterogenous sign system; it incorporates iconic, indexical, and symbolic signs. Iconic words or expressive words, such as onomatopoeic and sound-symbolic words, reveal direct linkage between sound and sense. Therefore, it is not surprising that the study of the expressive subsystem of a language has to proceed from different premises than the study of 'ordinary' language. Hinton *et al.* write that "meaning and sound can never be fully separated, and linguistic theory must accommodate itself to that increasingly obvious fact" (1994: 12).

The terms onomatopoeia and sound symbolism are used differently by various authors. Hinton *et al.* (1994: 2–6), for example, distinguish between four types of sound symbolism: 1) corporeal sound symbolism (e.g. *aaugh!, achoo!*); 2) imitative sound symbolism (onomatopoeic words); 3) synesthetic sound symbolism (e.g. size and shape symbolism); 4) conventional sound symbolism (phonesthemic and phonesthetic associations). An onomatopoeic word is a sound-imitative word (e.g. Est *põmm* 'bam', *sumisema* 'buzz', *piiksuma* 'beep'); that is, an onomatopoeic word denotes an acoustic phenomenon. The lexicography of onomatopoeic words was discussed in three previous articles by the author of this article (Veldi 1994; Veldi 1999; Veldi 2001). A sound-symbolic word, on the other hand, denotes a non-acoustic phenomenon. Pejorative meaning, for example, can be expressed by labio-velarity (Wescott 1980: 362–377; see also Voronin 1982: 94–98). Examples of phonosemantically motivated pejorative words include the Estonian *öökima*, *pakkima,* and *ropsima*; all of them are informal synonyms for 'vomit'. Their English counterparts are *puke* and *barf*; the latter word is more characteristic of American English. The basic descriptive unit for phonosemantic studies is not the phoneme but the phonemotype (Voronin 1987).

As is known, "the search for translation equivalence is crucial and difficult, requiring a high degree of bilingual competence and depending on how similar or different the respective languages are" (Hartmann 2001: 141).

The article claims that the awareness of phonosemantic alternations and patterns helps a lexicographer to establish expressive synonyms in two languages and to improve their treatment in bilingual dictionaries. Phonosemantic awareness is comparable to metaphor awareness in cognitive semantics in that enhanced awareness can yield fruitful results in such areas as bilingual lexicography and second language acquisition (see Õim 1999; Boers 2000).

1.1. *Ühe raksuga* 'at a stroke, at one go' and *ühe ropsuga* 'at a stroke, at one go'

Let us begin our study with an expressive alternation to be found in Estonian. A number of important iconic alternations in English were established by Roger Williams Wescott, see especially his article "Consonantal apophony in English" (Wescott 1980: 336–361).

For example, the Estonian phrases *ühe raksuga* 'at a stroke, at one go' and *ühe ropsuga* 'at a stroke, at one go' reveal an expressive alternation between *-ks* and *-ps*. These two phrases are synonymous. It seems, however, that lexicographers do not fully understand the potential and implications of such alternations.

Let us examine some dictionary evidence. "The Explanatory Dictionary of Written Estonian" (*Eesti kirjakeele seletussõnaraamat*, henceforth *EKSS*) treats the phrases *ühe raksuga* and *ühe ropsuga* as follows:

> *EKSS* (IV, 4, 819)
> (1) **ühe raksuga** järsku, hooga, ühe ropsuga. *Vihaga võtab tüdruk kõige jämedama propsi, tõstab selle ühe raksuga pukile.* V. Saar. .. *märkas poolikut viinapudelit, kallas selle ühe raksuga kurku.* E. Maasik.

> *EKSS* (V, 1, 176)
> (2) **ühe ropsuga** 1. järsku, hooga. Tõmbab ukse ühe ropsuga ristseliti lahti. • *August, püksid rebadel, väänas ründaja käe ühe ropsuga küünarliigesest välja.* J. Peegel. 2. korrapealt, hoobilt. *Põgenikul oli olukord ühe ropsuga selge.* • *Ühe ropsuga käsutati kõik pidulauast üles.* A. Kivikas. 3. ühekorraga, peatust v. vahet tegemata. *Haaras laudlinal otsast kinni ja tõmbas kõik ühe ropsuga maha.* • *Omal ajal käisime kolmkümmend [kilomeetrit] ja isegi rohkem ühe ropsuga.* H. Angervaks.

One can see that the treatment of these two phrases is different. It is interesting that *EKSS* points out the connection between *ühe raksuga* and *ühe ropsuga* only in the entry *ühe raksuga*. One can also notice that the treatment of *ühe ropsuga* is much more thorough in that it has three numbered senses whereas the phrase *ühe raksuga* is covered by one sense.

Next let us examine the treatment of the same phrases in the "Estonian–English Dictionary" by Paul F. Saagpakk (henceforth SAAGPAKK):

> SAAGPAKK
> (3) *ühe raksuga* at one go (or pull)
>
> (4) *ühe ropsuga* in a jiffy, all at once, in a twinkling (*or* flash, trice); at one go, with a leap, at one blow, at one swoop, at one stroke

It is obvious that the lexicographer was unaware of the alternation between *-ks* and *-ps* and thus treated these synonyms in an unsystematic way.

2. Informal and slang synonyms for the head, mouth, and testicles

The paper focuses on the Estonian and English informal and slang synonyms for the head, mouth, and testicles. The terms *pea* 'head', *suu,* 'mouth', and *munandid* 'testicles' are neutral, their synonyms *nupp* 'nob, nut', *(suu)mulk* 'gob, trap' and *munad* 'balls, nuts' carry the label *informal* or *slang*. The presence of labial phonemes in these words suggests sound symbolism as rounded shape is often symbolized by means of labials (both vowels and consonants, see Voronin 1982: 98–102). One can also see that many of these words include not one but two labial phonemes, for example Est *nupp, muna, mulk* and Eng *nob, gob*. It should be noted that the etymology of sound-symbolic words is often marked in etymological dictionaries as uncertain, obscure or unknown.

2.1. Head

The informal synonyms for *pea* 'head' include *nupp, nutt,* and *kolu.*

> SAAGPAKK
> (5) **nupp** 1. button; (*kaba, kepil, uksel* jne); knob; push button; (*mõõga käepideme*) pommel; (er. *kilbi keskel*) boss; (*lauamängus* jne.) piece 2. *(lille-)* bud

> 3. *(pää, aru) (E. fam)* sconce, noddle, pate
> *tal on ~ otsas e. tal on ~u (E. fam)* he has a good head on his shoulders, *(fam.)* he has his head screwed on the right way
> *nupust nikastanud (E. sl)* to have a screw loose, a bit cracked in the noddle, *(sl)* to have bats in one's belfry

The treatment of the first two senses of *nupp* is adequate. However, the expressive senses *pää* 'head' and *aru* 'reason, common sense', which are marked as informal *(E. fam)*, are treated inadequately. Are the provided equivalents *sconce, noddle*, and *pate* trustworthy? I decided to re-check their reliability and came up with surprising findings. Although all three words do denote the head, the problem is that all of them are either archaic or outdated. According the "Oxford English Dictionary" (henceforth *OED*) *sconce* is labelled as *archaic*; the 'most recent' quotation of *sconce* dates from 1888. The "Collins English Dictionary" and the "Encarta World English Dictionary" label *sconce* as archaic; "The New Oxford Dictionary of English" (henceforth *NODE*) does not list this entry at all. The second equivalent *noddle* is not a good choice either. *NODE* and the *Collins English Dictionary* label the word as *informal* and *dated, OED* defines the word as "the head as the seat of the mind or thought. (Colloq.), and usually with playful or contemptuous suggestion of dullness or emptiness". Thus, this equivalent has to be discarded for semantic reasons because the Estonian *nupp* does not carry a pejorative meaning. The Estonian equivalent for *noddle* is *peakolu*; this meaning is, by the way, recorded by the "English–Estonian Dictionary" by Johannes Silvet (henceforth *SILVET 3*).

> SILVET 3
> (6) **noddle** *fam.* nutt, peakolu

The third equivalent in *SAAGPAKK* is *pate*. It is not a good equivalent either. *NODE, Encarta*, and *SILVET 3* point out that *pate* is archaic.

> SILVET 3
> (7) **pate** *arh., nalj* pea, peakolu; pealagi

It appears that *sconce, noddle*, and *pate*, provided by *SAAGPAKK* as equivalents for the Estonian word *nupp* in the sense 'head; reason', are archaic or dated. Any criticism, however, should be constructive. Thus, we must find some current informal words that denote the head. After some work, the following equivalents could be found.

NODE
(8) **nob**
noun *informal* a person's head.

NODE
(9) **nut** *informal* a person's head

NODE
(10) **napper**
noun Brit. *informal* a person's head: *a couple of shaven nappers.*
ORIGIN late 18th cent.: from thieves' slang, of unknown origin.

NODE
(11) **noggin**
noun *informal* **1** a person's head.
ORIGIN mid 17th cent. (in the sense small drinking cup): of unknown origin.

NODE
(12) **bonce**
noun Brit. *informal* a person's head.
ORIGIN mid 19th cent. (denoting a large marble): of unknown origin.

NODE
(13) **noodle** *informal* a person's head; of unknown origin

As can be seen below, some of these words are covered by the *SILVET 3*. Interestingly enough, dictionary evidence shows that Silvet did realize that *nutt* and *nupp* are synonyms in the senses 'head' and 'reason'. Silvet lists only *nutt*, *kolu*, and *pea*. It shows once again that if a lexicographer is unaware of expressive alternations, he or she may fail to recognize the necessary synonyms.

SILVET 3
(14) **nob** *si* nutt, kolu, pea

SILVET 3
(15) **noodle** *Am.* (pea)kolu, pea

SILVET
(16) **nut** *sl.* (pea)nutt, (pea)kolu

SILVET 3
(17) **noggin** *Am. fam.* nutt, pea

Next let us examine the phrase *nupust nikastanud* 'off one's onion, off one's head'. The Estonian phrase reveals sound repetition of the initial sound. Once we realize that sound repetition plays a remarkable role in Estonian informal phrases (cf. also *nupp nokib* '(he) has brains'), we can establish a connection with the phrase

peast põrunud, which means the same. It should be pointed out that sound repetition is a common feature of Estonian informal compounds as well (e.g. *kurikael* 'bad guy, lit. cruel neck', *kiimakott* 'lecherous man', *tibatilluke* 'teeny-weeny, tiny').

SAAGPAKK provides three English equivalents for *nupust nikastanud*: 'to have a screw loose', 'a bit cracked in the noddle', and 'to have bats in one's belfry'. Of these the third equivalent *to have bats in one's belfry* is especially suitable because in this case the alliteration found in the Estonian phrase is rendered in English, too. One is tempted to ask, however, whether it is possible to find some other possible equivalents. It appears that the list of possible candidates is rather long: *off one's onion, off one's nut, out of one's gourd* (Am), *off one's pannikin* (Aus), *off one's scone, off one's head, out of one's head, off one's trolley* (Br), *wanting in the upper storey, nutso* (Am).

However, when we compare the treatment of *nupust nikastanud* and *peast põrunud* in *SAAGPAKK*, we can see that the provided equivalents for these two expressions are surprisingly different. It is also interesting to point out that *SAAGPAKK* lists the phrase *peast põrunud* both under *pääst* and *põrunud* and here the treatment is not identical either. It shows that Saagpakk's dictionary reveals a remarkable degree of unsystematicity.

> SAAGPAKK
> **nupp**
> (18) *nupust nikastanud (E. sl)* to have a screw loose, a bit cracked in the noddle, *(sl)* to have bats in one's belfry
>
> **pääst**
> (19) *pääst põrunud (E. fam.: hull)* crackbrained *(a.)*, crazy, moonstruck, *(fam.: ogar)* dotty, *(sl.)* nuts, *(Am. sl.)* nutty; *(veider)* cranky; *pääst põrunud olema* to be off one's nut *(sl.)*, to be wrong in the upper storey.
>
> **põrunud**
> (20) *(piltl.) (pääst)* cracked, *(er.)* cracky, crackbrained, *(Am. sl.)* wacky, gaga; *ta on pääst veidi põrunud* he is crazy *(Am. sl.* a little wacky); *põrunud (pää e. pea)* (a p.) queer in the head, a crazy p., *(sl.)* a nut

Thus, Saagpakk did not fully realize that the expressions *nupust nikastanud* and *peast põrunud* are synonymous. This finding is supported by Saagpakk's own dictionary of synonyms, which also fails to establish a connection between these two expressions (Saagpakk 1992: 182), cf.

nupp
(21) *(sl.)* nupust nikastanud (tal mõni kruvi logiseb)

2.2. Mouth

The mouth has also a number of informal synonyms. The best-known Estonian informal synonym is *suumulk*, which is a derogatory compound consisting of 'mouth + hole'.

According to *SAAGPAKK* and *SILVET 3*, *molu* is another derogatory synonym for the mouth. It seems, however, that the meaning of this word has changed and currently *molu* is a pejorative term for the face or a stupid person. According to *EKSS*, it may mean the mouth only marginally.

SAAGPAKK
(22) **suumulk** *(E. vulg.)* mouth
(23) **molu** *(E.sl.) (suu)* potato box, potato trap *(sl.)*
pea molu (E. vulg.) shut your rag-pox! *(sl.)* keep your trap shut *(Am. sl.)*

We can see that the treatment is rather uneven. In the first case the lexicographer provided only the neutral term *mouth* and forgot about the informal synonyms altogether. He also forgot about the phrase *pane suumulk kinni* 'shut your trap, shut your gob'. In the second case the informal equivalents are there, but the problem is whether they are well selected. I found the compound *potato trap* 'mouth' in the third edition of "The Concise Oxford Dictionary" (1934) but not in the tenth edition of the same dictionary published in 1999. Thus, with the exception of *keep your trap shut* the treatment of *molu* in *SAAGPAKK* is outdated both from the point of view of Estonian and English. Actually, there are some informal terms for the mouth in English.

NODE
(24) **gob**[1]
noun informal, chiefly Brit. a person's mouth: Jean told him to shut his big gob.
ORIGIN mid 16th cent.: perhaps from Scottish Gaelic *gob* 'beak, mouth'.

NODE
(25) **cakehole**
noun Brit. *informal* a person's mouth.

NODE
(26) **bazoo** US *informal* a person's mouth; of unknown origin

Phonosemantically, the word *gob* is an excellent equivalent for *suumulk*. Both the English word and the Estonian word reveal velarity and labiality. Roger Williams Wescott established four typical syllable structures for derogatory terms (Wescott 1980: 365):
1) labial onset, velar coda (wog, fuck, puke, punk, fink);
2) velar onset, velar coda (kike, cock, cack, kook, crook);
3) velar onset, labial coda (wop, quiff, crap, creep, guff);
4) labial onset, labial coda (pimp, boob, poop, flub, fop).

It appears that the English *gob* represents the third pattern (velar onset, labial coda) in Wescott's classification. The Estonian word *mulk*, however, represents the first pattern (labial onset, velar coda). In both words labiality is reinforced by the labial vowel, which makes them polylabial words.

The following examples from *SILVET 3* show that, apart from the fact that the semantics of *molu* is outdated and thus misleading, Silvet was unable to provide the most typical informal equivalent *suumulk*.

SILVET 3
(27) **gob** molu, suu

SILVET 3
(28) **trap** *sl. er. Am.* "molu", suu

2.3. Testicles

Now let us explore the treatment of informal synonyms for the testicles in Estonian–English and English–Estonian dictionaries:

SAAGPAKK
(29) **muna** *(anat.)* testicle
munad maha võtma (E. vulg.) to castrate; *(piltl.)* to emasculate, to make a p. powerless *(or* impotent)

(30) **kera**
(E.vulg.) kerad (pl.) (man's) testicles, balls (vulg.)

SAAGPAKK provides the English scientific and neutral term as an equivalent for an informal word and forgets about the informal equivalents. It concerns also the expression *munad maha võtma* 'cut off sb's balls', where an informal English equivalent would be preferable. The entry *kerad* is better in that it provides *balls* in addition to *testicles*. However, there are of course many more

informal synonyms in English for testicles. The following list was compiled from *NODE*.

NODE
(31) **balls** *vulgar slang* testicles

(32) **bollocks** (also **ballocks**) Brit. *vulgar slang*
plural noun **1** the testicles
ORIGIN mid 18th cent.: plural of *bollock*, variant of earlier *ballock*, of Germanic origin; related to ball[1].

(33) **cobblers** Brit. *informal* a man's testicles. from rhyming slang *cobbler's awls* balls.
ORIGIN Middle English: of unknown origin.

(34) **goolie** (also **gooly**)
noun (pl. **-ies**) **1** (usu. **goolies**) Brit. *vulgar slang* a testicle. 1930s: apparently of Indian origin;
compare with Hindi *golī* 'bullet, ball, pill'

(35) **cojones**
plural noun *informal*, chiefly US a man's testicles.
ORIGIN Spanish.

(36) **rocks** *vulgar slang* a man's testicles

(37) **nuts** *vulgar slang* a man's testicles

(38) **knackers** Brit. *vulgar slang* a man's testicles

It is interesting to note that *SILVET 3* does not list any of the above-mentioned informal words meaning 'testicles'. *SILVET 3* was published in 1989–1990, and at that time informal words denoting sex organs were simply omitted in general dictionaries published in this country.

3. Conclusion

The analysis of the *Estonian–English Dictionary* by Paul Saagpakk and the *English–Estonian Dictionary* by Johannes Silvet shows that the treatment of informal synonyms for the head, mouth, and testicles reveals a number of drawbacks. The lexicographers failed to establish connections between synonymous informal words and expressions. As a consequence, important synonyms are neglected. If they are covered by the dictionary, their treatment is unsystematic, which lowers the degree of trustworthiness. On the whole, the treatment of informal synonyms in Estonian–English and English–Estonian dictionaries is unsystematic and inadequate. Enhanced

awareness of phonosemantic alternations and patterns would help the lexicographer to improve the quality of informal equivalents in bilingual dictionaries.

References

Boers Frank 2000. Metaphor awareness and vocabulary retention. – Applied Linguistics 21: 4, 553–571.

Collins English Dictionary. Fourth edition. Glasgow: HarperCollins, 1998.

EKSS = Eesti kirjakeele seletussõnaraamat. (1988–2001). Tallinn: Eesti Keele Instituut.

Encarta World English Dictionary. Ed. by. Anne H. Soukhanov. NY: St. Martin's Press, 1999.

Hartmann, R. R. K. 2001. Teaching and Researching Lexicography. Harlow, England: Longman.

Hinton, Leanne; Nichols, Johanna; Ohala, John 1994. Introduction: Sound-symbolic processes. – Sound Symbolism. Ed. by Leanne Hinton, Johanna Nichols, John J. Ohala. Cambridge UP. 1–12.

NODE = The New Oxford Dictionary of English. Ed. Judy Pearsall. Oxford UP, 1998.

OED = Oxford English Dictionary. Second edition. CD-ROM, Version 2.0. Oxford UP.

SAAGPAKK = Saagpakk, Paul F. Estonian–English Dictionary. Tallinn: Koolibri, 1992.

Saagpakk, Paul F. 1992. Sünonüümisõnastik. Brampton ON: Maarjamaa.

SILVET 3 = Silvet, J. Inglise–eesti sõnaraamat I–II. Tallinn: Valgus, 1989–1990.

Veldi, Enn 1994. Onomatopoeic words in bilingual dictionaries (with focus on English–Estonian and Estonian–English). – Dictionaries. Journal of The Dictionary Society of North America 15, 74–85.

Veldi, Enn 1999. Lexicography of onomatopoetic words. – Estonian: Typological Studies 3. Ed. by Mati Erelt. Publications of the Department of Estonian of the University of Tartu 11. Tartu. 204–230.

Veldi, Enn 2001. Estonian–English dictionary of onomatopoeic words. – Itämerensuomalaista ekspressiivisanaston tutkimusta. Toim. Juha Leskinen. Suomen Kielen Laitoksen Julkaisuja 42. Jyväskylän Yliopisto. 137–144.

Voronin, Stanislav V. 1982. Osnovy fonosemantiki. Leningrad: Izdatel'stvo Leningradskogo Universiteta.

Voronin, Stanislav V. 1987. The phonemotype: A new linguistic notion (Implications for typological phonosemantics). – Proceedings of the Eleventh International Congress of Phonetic Sciences. Vol. 4. Tallinn. 197–200.

Wescott, Roger Williams 1980. Sound and Sense. Linguistic Essays on Phonosemic Subjects. Edward Sapir Monograph Series in Language, Culture, and Cognition. Vol. 8. Lake Bluff, Illinois: Jupiter Press.

Õim, Haldur 1999. How to portray emotions? – Estonian: Typological Studies 3. Publications of the Department of Estonian of the University of Tartu 11. Ed. by Mati Erelt. Tartu. 231–252.

Politeness debate continued – notes on some key controversial issues in Brown and Levinson's theory

Krista Vogelberg
University of Tartu

1. Introduction

The present article attempts to tackle a few of the key controversial issues related to Brown and Levinson's politeness theory (Brown and Levinson 1978/1987), exploring both its general validity and the extent to which and areas where it needs to be modified. Brown and Levinson's theory is certainly only one of the competing models that set out to explain linguistic politeness phenomena. Meanwhile, the very fact that most of the latest work in the area chooses this theory as its object of criticism, effectively bypassing other models, and, in the final analysis, the alternatives recently offered constitute modifications of its basic framework, testifies to its continuing relevance and vitality.

In my view, this vitality is not simply due to the fact that this theory "most clearly claims its pancultural validity" (O'Driscoll 1996: 2) or just happens to currently have the highest profile (ibid), but rather owing to a basic strength that its rivals – notably the maxim approaches propounded by Leech (1983) and Lakoff (1973, 1990) – lack. As has justifiably been pointed out by Brown and Levinson themselves, the problem with the maxim approaches is that they treat politeness on a par with cooperation in the Gricean sense. However, while cooperativeness is equally charateristic of any communication, being a presumptive framework in which efficient communication takes place, politeness is a matter of degree: it is not the case that the same modicum of politeness is owed or accorded to all partipants in a communicative event on all occasions. We tend to be more polite to our superiors, to strangers, etc., i.e. "the distribution of politeness (who has to be polite to whom) is socially controlled" (Brown, Levinson 1987: 4).

The maxim approaches, by failing to recognize the socially contingent nature of politeness or at least to account for it in a

principled way, posit cultural differences in linguistic politeness behaviour but leave these largely unexplained. Thus, according to Leech, the Maxim of Modesty prevails over the Maxim of Agreement in Japanese culture while the opposite is true for Anglo-Saxon culture, as evidenced in situations where the two clash, e.g. in responses to compliments (Leech 1983: 137). However, we are given no explanation why this should be so. Politeness thus tends to be studied in an immanentist way as a self-contained phenomenon unrelated to the wider social context in which it operates. Meanwhile, Brown and Levinson account for politeness behaviour on the basis of extralinguistic variables, which allows one to link up politeness norms with more general factors of socio-cultural and, in the final analysis, historical and economic nature. Indeed, the theory is informed with the fundamental belief that it is in interaction that "most profound interrelations between society and language are to be found" (Brown, Levinson 1987: 280).

It is this basic strength that has made and continues to make it worthwhile to subject Brown and Levinson's concrete articulation of their theory to closer scrutiny. The present article takes up only a few of the wide range of moot issues, selecting, in the main, those that have emerged in the course of analysing the data collected in the empirical work of the author and her colleagues.

2. Politeness and face-threat

Brown and Levinson define politeness as a way of dealing with potential aggression inherent in much of human contact. The central notion of their theory is "face", derived from Goffman (1967), for which they give various definitions (e.g. "the public self-image that every member /of a group – K.V./ wants to claim for himself", Brown and Levinson: 61), of which the clearest seems to be "self-esteem" (Brown, Levinson 1987: 2) with the qualification that it should be recognized by others, as in the English expressions "save/keep/lose face". Brown and Levinson further distinguish between negative and positive face, defining the first as "the basic claim to territories, personal preserves, rights to non-distraction – i.e. freedom of action and freedom from imposition" (Brown, Levinson 1987: 61) and the latter as "the positive consistent self-image (crucially including the desire that this self-image be appreciated and approved of)" (ibid). The latter definition, however, seems to largely

coincide with the definition of face in general. It appears that a better formulation, which is actually best in tune with Brown and Levinson's own approach as realised in the bulk of their work, would focus on the individual's need for inclusion in the group (e.g., Gudykunst and Ting-Toomey 1988: 86). Further, positive and negative politeness are defined as addressed to positive and negative face, respectively.

The link Brown and Levinson claim between politeness and threat to face by conceiving of politeness as a way of mitigating the threat, of redressing face, is an aspect of their theory that has attracted abundant criticism. Thus, Nwoye (1992: 311) claims that human interaction that revolves round "continuous mutual monitoring of potential threats" is robbed of all elements of pleasure and Fraser (1990: 235), in fact echoing Riley (1981), describes it as "a zero-sum game". Schmidt (quoted in O'Driscoll 1996) goes as far as to characterise Brown and Levinson's view of communication as "paranoid".

Meanwhile, Janney and Arndt (1992) see the function of politeness in Brown and Levinson's sense exactly as a uniquely human way of conflict avoidance (and, one might add, conflict resolution), as against the typical animal responses of fighting or fleeing[1], and conflicts themselves as biologically pre-programmed, innate, and universal.

It should be noted that Janney and Arndt make a distinction between social politeness and tact and apply the function of conflict avoidance to the latter alone. Their distiction essentially coincides with the difference between "volitional"/"instrumental" politeness and "discernment" as suggested by Ide (1989). I believe that Brown and Levinson's politeness actually covers both – a point of view for which O'Driscoll (1996) has made a convincing case. The issue is worth further discussion which, however, is beyond the scope of the present article. Note, however, that Janney and Arndt leave politeness in the narrow sense only the role of ensuring a smooth organization of interaction and account neither for its biological origins nor indeed for the very necessity of special linguistic routines to ensure smoothness of interaction.

---

[1] One could actually claim that at least gregarious animals have forms of conflict resolution analogous to human politeness, e.g. the submission rituals among wolves and primates, etc.

Brown and Levinson themselves in their Introduction to the 1987 re-issue of their seminal work concede that their theory, proceeding as it does from Goffman's concept of "virtual offence", might overexploit the metaphor of threat. They note, however, that politeness is precisely designed to turn a zero-sum game into a cooperative non-zero-sum game (Brown, Levinson 1987: 52).

Along these lines, one can reason that politeness may be motivated not only by an actual face threatening act but also by a wish to forestall **potential** threat. This would tally with Scollon and Scollon's (1981) observation that positive politeness is relevant to all aspects of a person's positive face and, unlike negative politeness, not specific to a concrete face-threatening act. The use of positive politeness strategies (e.g. small talk, joking, etc.) in the absence of immediate face threat may be interpreted as constant bolstering of ingroup feelings against the potential threat of their crumbling in a hypothetical yet possible conflict situation.

In fact, Brown and Levinson themselves do not treat the notion of face threat in an unequivocal way. On the one hand, face-threatening acts do constitute for them only a subset of all acts (see, e.g., Brown, Levinson 1987: 60) and they repeatedly refer to acts in the case of which face threat is missing. Also, they at times link face threat directly to the presence of a concrete impositive speech act, i.e. to the positive value of R (rate of imposition) in their formula (thus, they at least once equate a linguistic form's being "FTA-sensitive" to its choice being influenced by the R-factor, op. cit.: 18).

On the other hand, however, their general formula computes the weightiness of a face-threatening act (W) as a simple sum of **three** variables: power of Hearer over Speaker (P), distance between Hearer and Speaker (D), and rate of imposition (R). Indeed, the simple summative basis of assigning values to W is specifically emphasised by the authors as "working surprisingly well" (op. cit. 76) though they admit that some more complex composition of values might be involved. In fact, one can suggest that if FTA-sensitivity genuinely equalled R-sensitivity, a more logical formula would be something like $(P + D) \times R$, in the case of which $R = 0$ would entail $W = 0$.

However, it does not seem to have been noticed that the formula as it stands actually accounts for situations where politeness is used though a concrete impositive act is absent. The most famous of these is Matsumoto's (1987/1989) example of Japanese where

with an "obviously non-FTA utterance" such as "Today is Saturday" the speaker has to choose between a plain form (*da*), the adressee honorific (*desu*) and the super-polite addressee-honorific (*de gozaimasu*). Levinson himself is perfectly aware of this when he points out that in some languages, notably South-East Asian, including Korean, Japanese and Javanese, it is "almost impossible to say anything at all which is not sociolinguistically marked as appropriate for certain kinds of addressees only" (Levinson 1983: 90). Now, since Brown and Levinson also admit that honorifics are a means of expressing politeness, this fact would indeed be seriously damaging to their theory were it not that W also has a positive value when R equals zero but P and D have positive values. In the case of Matsumoto's example, it is telling that there does exist a plain form – presumably to be used when P and D also approximate zero.

Thus, it turns out that Brown and Levinson's own formula dissociates face threat in general, as represented by W, from a concrete impositive speech act such as a request as represented by R. A mere addressing of a higher-ranking or distant person constitutes a face threat that has to be mitigated, be it with means pertaining to "normative" or "volitional" politeness.

The presence of face threat in merely addressing the Hearer is, in fact, not specific to stable stratified societies where more of the necessary mitigating devices tend to be grammaticalised but also to those characterised by greater mobility and egalitarianism where the Speaker's latitude in the choice of pertinent linguistic expressions is larger. For instance, bringing a piece of neutral information to the attention of one's superior often calls for fairly extensive facework also in Western societies. Also, Anglo-Saxon academic literature abounds in hedges addressed to the Gricean maxims of quality, quantity, and relevance such as "As is well known", "It seems / appears that ...", modal verbs "may", "might", "could" etc. that qualify as hedges precisely because they are not required by the Hallidayan ideational function of the text but rather serve the interpersonal function. In other words, they are used to project the author's definition of the relationship between him or her and the reader where the reader rather than the author is seen as occupying the position of authority.

In this connection, Brown and Levinson's (1987: 165) analysis of "As is well known" as a negative politeness strategy employed to

"disclaim the assumption that the point of S's assertion is to inform H" is particularly significant as it in fact admits that simply **informing** the Hearer can constitute a face threat, though informing as a speech act is not included in their list of face-threatening acts (cf. Brown, Levinson 1987: 68). The discrepancy can again be well explained in terms of the W/R distinction suggested above.

Similarly, Kivik and Vogelberg (in press) found that in a discussion situation framed by its American participants as competitive, negative politeness strategies accompanied what was essentially personal story-telling: "That's something that happened last year in Estonia/ **really anyone/ I mean// I don't know if this is true/ and if anyone else has experienced this/ but** I would come into a group situation/ and ...", "and **maybe it's just from the male perspective/ I don't know but/ again** ...me in the US/ **maybe I'm just strange".** Again, it can be argued that simple imparting of information (R = 0) was seen by interactants as face-threatening (W > 0)[2].

One might of course also follow the line taken by O'Driscoll (1996) and drop the notion of face threat in favour of that of the simple need for face to be attended to, i.e. the need for a symbolic recognition by others of one's desires for merging, on the one hand, and independence, on the other. In this case, however politeness strategies would also be present where the value of W equals zero. Meanwhile, numerous studies (including Ide 1989 on Japanese and American requests, and work conducted under the supervision of the present author on Estonian, Russian and Anglo-American requests, see, e.g. Aas 1999, Konovalov 2001) confirm that both "normative" and "volitional" politeness is dropped among interactants who take the small values of P and D for granted and do not feel the need to buttress in-group feelings.

---

[2] I have elsewhere (Vogelberg, in press a) ventured the suggestion that where status-equals engage in competition as serious as that lying at the basis of American society, we would have to introduce the notion of a **potential** high value of P. Where everyone struggles to enhance his or her position (i.e. P), face-threat is computed not on the basis of the actual value of P but on the basis of the difference sought after over the long-term. In other words, P is not what it is at the moment, nor even what it is hoped to be by the end of the speech event as a result of negotiating the values of the parameters but what participants unconsciously hope – or fear – it might be eventually, as the result of a long competitive race.

It thus seems expedient to retain the notion of face threat – albeit sometimes potential and long-term in nature – as a motivating force for politeness. Also, Brown and Levinson's formula in its simple summative form has proved to capture the independence of face threat in the broad sense from a concrete impositive act better than is realized by its authors.

Meanwhile, the fact that W is computed on the basis of a sum of three variables does not mean that the use of particular politeness strategies may not be linked to only one of the variables. Thus, in "tu/vous" languages the choice of one of the alternatives is relatively immune to the value of R. However, the total use of strategies does seem to depend on the value of W. For instance, in "tu/vous" languages one does not switch to the "vous" form with a friend even when presenting a large request, but one has to make up for this "rigidity" of honorifics by copious use of other politeness strategies[3].

3. Negative and positive face: hierarchy or symmetry?

If we accept Brown and Levinson's basic formula – and as I hope to have demonstrated in the above, the formula's potential seems to be even greater than realised by its authors – then we have to think of politeness as a matter of degree[4]. Whether degree is further fraught with **kind** is, however, a separate question. For Brown and Levinson, superstrategies (i.e., in essence, kinds) of politeness are ranked in the descending order from off-record to negative politeness to positive politeness. However, critics have pointed out that, e.g., off-record strategies are often perceived by subjects as less polite than those of negative politeness (Blum-Kulka 1985) or that on-record and positive politeness strategies are rated as more polite than those of negative politeness (cf., e.g. Wierzbicka 1991).

---

[3] This mutually compensatory nature of "normative" and "volitional" politeness, which is also borne out by Aas's (1999) and Konovalov's (2001) Estonian-English and Russian-English contrastive data, once more indicates the absence of an essential difference between the two.

[4] Degrees of politeness are, of course, also recognised in other approaches (cf., e.g., Leech's fine-grained presupposition-based analysis of various forms of indirect requests, Leech 1983) yet not linked to social parameters of the context.

The strongest argument in favour of Brown and Levinson's ranking is the fact that across cultures in asymmetrical power configurations the superior uses positive politeness strategies to the subordinate and the subordinate negative politeness strategies to the superior (cf. Scollon, Scollon 1995: 45–47). As will be shown below, the superior can use his or her power to redefine the situation and, e.g., insist on positive politeness strategies on the part of the subordinate (this power constitutes metapower in relation to Brown and Levinson's P, Vogelberg 1997), but the prototypical situation is as described above. Also, Brown and Levinson refer to Durkheim 1915 who describes negative rites – those of avoidance – as addressed to higher deities, and positive rites – those of approach – as reserved to lesser deities (Durkheim 1915). Note that, contrary to Goffman, Brown and Levinson see the sacred rites as an extension or projection of the primordial interpersonal ones.

On this view, however, face dualism itself must be asymmetrical – an issue that continues to be hotly debated. O'Driscoll (1996: 9, 10) in his perspicuous defence of the universality of face and face dualism traces it back to existential wants that human beings share with animals: the needs to do some things together with others and some things alone. His examples concerning primates (nurturing versus defecating), however, make it clear that the wants are seen as symmetrical. Human as well as primate activities are mapped on "an axis which defines the sheer amount of interaction" with others. Tannen's famous metaphor of humans as porcupines in winter who are drawn together by the need to keep warm and pushed apart by the prickles expresses the same approach.

Further support for a symmetry of the two faces comes from authors who contend that their relative importance varies with the culture. Thus, Wierzbicka (1991: 37) claims that "in Anglo-Saxon culture, distance is a positive cultural value which is linked to respect for the individual's autonomy", whereas "in Polish culture it is associated with hostility and alienation". Similar observations have been made about Israeli culture (Blum-Kulka *et al* 1985).

Wierzbicka (1991) describes Anglo-American culture as one where every individual wishes and is entitled to have a little wall around him or her. Interestingly, however, the same wall-metaphor has been used by Americans with respect to Estonians: Kivik in her 1996 study quotes an American student (who clearly had no

knowledge of the works of Wierzbicka): "With Estonians, when you are talking with them, there's a kind of **distance** ... its sort of as if there was a **wall** there".

Meanwhile, the fact that prototypical values of D (or, for that matter, P and D) vary with the culture does not in itself invalidate face asymmetry. Brown and Levinson are perfectly aware of the existence of what they term "ethos" of a culture, understood by them as its "general interactional quality" (Brown, Levinson 1987: 243). They bring as typical examples the "friendly, backslapping cultures as in the Western US", and the "stand-offish" cultures like the British – in the eyes of the Americans – and the Japanese – in the eyes of the British, adding, however, that generalisations of this kind involve "immense crudity". However, what is interesting is that the use of negative politeness strategies where positive politeness is expected is perceived as hostile, alienating, uncooperative.

This impression, implied in the use of the "wall" metaphor at least in the data of Kivik, surfaced in unmistakable terms in the present author's experiment where videotaped role-play was used to study Estonian politeness behaviour in interlanguage communication with a native speaker of English (Vogelberg 1997). In the experiment, 10 Estonian students of English with varying degrees of Estonian pragmatic transfer in their interlanguage were asked to participate in a mock scholarship interview with an American as an interviewer. The interviewees were rated both by the American interviewer and 3 other Americans as well as a group of Estonian teachers of English. The ratings of the American group, on the one hand, and the Estonian teachers, on the other, showed high negative correlation (-0.78). Analysis of the videotapes revealed that Estonian interviewees who were more highly appreciated by Americans used predominantly positive politeness strategies while those whose performance was more acceptable to Estonian observers transferred negative politeness strategies from Estonian. Essentially, the modesty that appealed to Estonian observers did not resonate with the Americans: giving the interviewer an 'out', instead of endearing the interviewee to him and thereby promoting a positive decision, worked against the interviewee.

This conclusion was confirmed by follow-up interviews with all participants. The result particularly significant here is that the self-effacement of the Estonians who employed negative politeness

strategies was not only taken at face value by the Americans (as in the interviews discussed by Scollon and Scollon 1983) but in their description of these interviewees words such as "uncooperative", "conceited" and "hostile" were actually used.

Thus, we seem to have cogent arguments against the hierarchy of politeness strategies as propounded by Brown and Levinson and the face asymmetry that follows logically from it (though is not explicitly formulated by the authors themselves). However, the arguments in favour of the asymmetry are also there and should not be precipitately discarded.

To examine the question, let us turn back to the primates invoked by O'Driscoll. It is true that the wants of association and dissociation as exemplified by him are symmetrical and equally satisfied for all members of the group. However, if we remember that negative face does not just embrace desire for freedom from **contact** but also, crucially, desires for freedom from **impositions** and freedom of **action**, and "a basic claim to territories"[5], then these are clearly satisfied to a greater extent for those members whose position in the hierarchy of the group is higher. Thus, while all members are included in the group (their positive face wants are met to an equal extent), only a select few have full freedom of action. For these few, however, satisfaction of negative face wants automatically **entails** satisfaction of positive face wants and need not be separately signalled. This basic asymmetry would also account for the default option of asymmetrical use of negative and positive politeness strategies in hierarchical politeness systems.

The situation changes, however, when, e.g. the leader/superior wishes to use his or her meta-power to redefine the value of P and/or D in the basic formula. For instance, when the subordinate is invited to play the game of "being pals" and refuses[6], this is naturally interpreted as an expression of antagonism.

---

[5] Note how Brown and Levinson, who in general do not link their theory to animal behviour, here use an expression that actually points to the link.

[6] An example of this kind of negotiation is provided by O'Driscoll (1996: 17) where a boss asks his secretary to switch to a first-name basis and the secretary persists in title-last-name usage to "keep him at arm's length".

The formula, after all is interactional in nature, i.e., it describes values of P, D and R that guide the choice of interactional strategies which, as I have argued elsewhere (Vogelberg 2001, in press a, in press b), may but need not coincide with the "real values" of the parameters that regulate other behaviour.

In American culture, in particular, where the myth of equality of opportunity combines with status-uncertainty to yield the paradoxical result of excessive status-consciousness, while the myth of equality of respect counteracts the expression of the status-consciousness, relations between the value of P and real power can become fascinatingly complex (Vogelberg in press a).

Also, in a very individualistic society where the "real", behavioural value of D is great and no individual can ever feel quite certain of his or her membership in any group, interactional strategies can in many ways be seen to perform a compensatory function by outwardly signalling a low value of D. This argument is in tune with Giles *et al* (1992) claim that the American fear of silence is linked to American individualism, and would also explain the meaning of the expression "superficial friendliness" that one so often hears in characterisations of Americans.

Thus, the situation also changes as soon as at least one of the interactants is not sure of his or her inclusion in the in-group. This is especially common in intercultural situations where inclusion in the in-group is, at least initially, ruled out by definition. It is, therefore, particularly in this situation that the power of interaction to **construct** or **define** the situation rather than merely reflect it becomes apparent.

Brown and Levinson treat the causes of inappropriateness of "too much" politeness only in passing. In the light of the foregoing, it could be argued, however, that negative politeness is truly more polite only in stable in-group situations where granting the addressee independence does not preclude his or her belonging to the group. In these situations attending to the addressee's negative face automatically comprises also attending to his or her positive face. In a situation, however, where the interactant with more "real" power wishes to redefine the interactional situation towards lower values of the variables, as well as in an intercultural situation where interactants belong to different groups, negative politeness can also be

interpreted as indicative of the opposite of inclusion, i.e., rejection, often with the purpose of protecting the Speaker's own face[7].

These two considerations together would explain the results of the interview-experiment, where the "real" power of the American interviewer was significant and further reinforced in the intercultural zone by centre–periphery relations, yet the rules of the game (dictated by the real power as meta-power) provided for a low value of P and D. Also, rejection, naturally, is least tolerable in a gate-keeping situation such as an interview which in many ways can be regarded as a ritual where the applicant has to demonstrate his or her eagerness to become member of the institution involved, to treat it as ingroup.

4. Conclusion

Both empirical work and theoretical analysis indicate that Brown and Levinson's basic approach, though in need of modifications, is still the best one available to account for linguistic politeness phenomena as contingent on factors related to the social context of interaction. The present article has attempted to draw attention to the potential of the summative nature of their basic formula in describing face threat as a phenomenon broader than the imposition involved in a concrete face-threatening act. However, not all politeness strategies may not be equally linked to all variables. The hierarchy of politeness superstrategies – and the consequent face asymmetry – in the main appears to withstand critical scrutiny and can actually be grounded in animal behaviour. However, this applies only to stable in-group situations. In the context of intra- or intercultural negotiation of the values of P, D and R – which are interactional in nature and may

[7] Brown and Levinson devote relatively little attention to Speaker's face as against Hearer's – a fact that has prompted Chen (2001) to claim that in the bulk of their work they do not deal with self-politeness at all, which in fact is not quite correct. Specially, they do seem to imply defence of the self face in their analysis of Apache attitudes towards white American positive politeness when they claim that positive politeness from semi-strangers is "irredeemably invasive of their /the Apaches' – K.V./ person" (14). The Apaches themselves, of course, employed negative politeness and were perceived as hostile by the whites.

thus differ from "real" power and distance – negative politeness can be shown to lose its status of being intrinsically "more polite" than positive politeness.

References

Aas, Annika 1999. Requests: A Cross-Cultural Study. Unpublished Bachelor's Thesis. University of Tartu.

Blum-Kulka, Shoshana 1985. Indirectness and politeness in requests: same or different? – Proceedings of the International Conference at Viareggio, Sept. 1985. Ed. by M. Papi, J. Verschueren. Amsterdam. 35–52.

Blum-Kulka, Shoshana; Danet, Brenda; Gherson, Rimona 1985. The language of requesting in Israeli society. – Language and Social Situations. Ed. by J. P. Forgas. New York: Springer. 113–139.

Brown, Penelope; Levinson, Stephen C. 1978/87. Politeness: Some Universals in Language Use. Cambridge: Cambridge University Press.

Chen, R. 2001. Self-politeness: a proposal. – Journal of Pragmatics 33, 87–106.

Durkheim, Emil 1915. The Elementary Forms of Religious Life. London.

Fraser, Bruce 1990. Perspectives on politeness. – Journal of Pragmatics 14, 219–236.

Giles, Howard; Coupland, Nikolas; Wiemann, John M. 1992. 'Talk is cheap...' but 'My word is my bond': beliefs about talk. – Sociolinguistics Today: International Perspectives. Ed. by K. Bolton, H. Kwok. London, New York: Routledge. 218–243.

Goffman, Erving 1967. Interactional Ritual: Essays in Face-to-Face Behavior. New York: Garden City.

Gudykunst, William B.; Ting-Toomey, Stella 1988. Culture and Interpersonal Communication. Newbury Park, London, New Dehli: Sage.

Ide, Sachiko 1989. Formal forms and discernment: two neglected aspects of universals of linguistic politeness. – Multilingua 8, 223–248.

Janney, Richard W.; Arndt, Horst 1992. Intracultural tact versus intercultural tact. – Politeness in Language. Studies in History, Theory and Practice. Berlin, New York: Mouton de Gruyter. 21–43.

Kivik, Piibi-Kai 1996. Beliefs about and attitudes towards silence: A cross-cultural study. Unpublished MA Thesis. University of Tartu.

Kivik, Piibi-Kai; Vogelberg, Krista. In press. Contrasts between contrasters: What discussion groups can tell us about discourse

pragmatics. – Meaning through Language Contrast. Ed. by K. M. Jaszcolt, K. Turner. John Benjamins Publishers.

Konovalov, Vyatscheslav 2001. Indirectness and Politeness in English and Russian Requests. Unpublished MA Thesis. University of Tartu.

Lakoff, Robin 1973. The logic of politeness, of minding your p's and q's. – Papers from the Ninth Regional Meetings of the Chicago Linguistics Society. 292–305.

Lakoff, Robin 1990. Talking Power. New York: Basic Books.

Leech, Geoffrey 1983. Principles of Pragmatics. London, New York: Longman.

Levinson, Stephen C. 1983. Pragmatics. Cambridge: Cambridge University Press.

Matsumoto, Yoshiko 1987. Politeness and conversational universals – observations from Japanese. Paper presented at the 1987 International Pragmatics Conference. Antwerp, Belgium.

Matsumoto, Yoshiko 1989. Politeness and conversational universals – observations from Japanese. – Multilingua 8, 207–221.

Nwoye, Onuigbo G. 1992. Linguistic politeness and socio-cultural variation of face. – Journal of Pragmatics 18, 309–328.

O'Driscoll, Jim 1996. About face: A defense and elaboration of universal dualism. – Journal of Pragmatics 25, 1–32.

Riley, Peter 1981. Strategy: collaboration or conflict? Paper presented at the 1981 BAAL Conference.

Scollon, Ron; Scollon, Suzanne Wong 1981. Narrative, Litercay and Face in Interethnic Communication. New Jersey: Norwood.

Scollon, Ron; Scollon, Suzanne Wong 1983. Face in interethnic communication. – Language and Communication. Ed. by I. C. Richards, R. W. Schmidt. London: Longman. 156–188.

Scollon, Ron; Scollon, Suzanne Wong 1995. Intercultural Communication: A Discourse Approach. Oxford, Cambridge: Blackwell.

Vogelberg, Krista 1997. Message-construction strategies in the interlanguage of Estonian learners of English: The example of gatekeeping encounters. – Views on the Acquisition and Use of the Second Language. (EUROSLA'7 Proceedings). Ed. by L. Diaz, C. Perez. Barcelona: Pompeu University Press. 469–475.

Vogelberg, Krista 2001. Salient issues in intercultural communication. – EATE Newsletter 20, 5–10.

Vogelberg, Krista. In press a. "Anglo-American linguaculture?" Notes on British and American politeness behaviour. – Britain in the New Millennium. The Challenge of the Grassroots. Ed. by P. Rajamäe, K. Vogelberg. Tartu: Tartu University Press.

Vogelberg, Krista. In press b. Eestlaste viisakuskäitumisest vahekeelses suhtluses ameeriklastega. – Kultuuridevahelise suhtlemise uurimusi Eestis. Toim. A. Valk.

Wierzbicka, A. 1991. Cross-Cultural Pragmatics: The Semantics of Human Interaction. Berlin, NY: Mouton de Gruyter.

Using the web as corpus for linguistic research

Martin Volk
University of Zurich

1. Introduction

In the last decade the working methods in Computational Linguistics have changed drastically. Fifteen years back, most research focused on selected example sentences. Nowadays the access to and exploitation of large text corpora is commonplace. This shift is reflected in a renaissance of work in Corpus Linguistics and documented in a number of pertinent books in recent years, e.g. the introductions by Biber *et al* (1998) and Kennedy (1998) and the more methodologically oriented works on statistics and programming in Corpus Linguistics by Oakes (1998) and Mason (2000).

The shift to corpus-based approaches has entailed a focus on naturally occurring language. While most research in the old tradition was based on constructed example sentences and self-inspection, the new paradigm uses sentences from machine-readable corpora. In parallel the empirical approach requires a quantitative evaluation of every method derived and every rule proposed.

Corpus Linguistics, in the sense of using natural language samples for linguistics, is much older than computer science. The dictionary makers of the 19th century can be considered Corpus Linguistics pioneers (e.g. James Murray for the Oxford English Dictionary or the Grimm brothers for the Deutsches Wörterbuch). But the advent of computers has changed the field completely.

Linguists started compiling collections of raw text for ease of searching. In a next step, the texts were semi-automatically annotated with lemmas and recently with syntactic structures. First, corpora were considered large when they exceeded one million words. Nowadays, large corpora comprise more than 100 million words. And the World Wide Web (WWW) can be seen as the largest corpus ever with more than one billion documents.

Professor Õim and the Computational Linguistics group at the University of Tartu have participated in this field early on. They have compiled a Corpus of Estonian Written Texts and a Corpus of Old

Estonian Texts. Later they have been active in the EuroWordNet project and contributed the Estonian side to this thesaurus project.

In addition to written text corpora some researchers have compiled spoken language corpora with audiotapes and transcriptions. There are also corpora with parallel videos and texts that allow the analysis of gestures. In this paper we will focus on text corpora for written language.

The current use of corpora falls into two large classes. On the one hand, they serve as the basis for intellectual analysis, as a repository of natural language data for the linguistic expert. On the other hand, they are used as training material for computational systems. A program may compute statistical tendencies from the data and derive or rank rules which can be applied to process and to structure new data.

Corpus Linguistics methods are actively used for lexicography, terminology, translation and language teaching. It is evident that these fields will profit from annotated corpora (rather than raw text corpora). Today's corpora can automatically be part-of-speech tagged and annotated with phrase information (NPs, PPs) and some semantic tags (e.g. local and temporal expressions). This requires standard tag sets and rules for annotation. As Geoffrey Sampson has put it: "The single most important property of any database for purposes of computer-assisted research is that similar instances be encoded in predictably similar ways."[1]

However, Corpora distributed on tape or CD-ROM have some disadvantages. They are limited in size, their availability is restricted by their means of distribution and they no longer represent the current language by the time they are published. The use of the web as corpus avoids these problems. It is ubiquitously available and due to its dynamic nature represents up-to-date language use.

The aim of this paper is to show a number of examples of how the web has been used as a corpus for linguistic research and their applications in natural language processing. Based on this, we list the advantages and limits of the web as corpus. We conclude with a proposal for a linguistic search engine to query the web.

---

[1] Quote from a talk at Corpus Linguistics 2001 in Lancaster: "Thoughts on Twenty Years of Drawing Trees".

2. Using the web for Corpus Linguistics

Using the web for Corpus Linguistics is a very recent trend. The number of approaches that are relevant to Computational Linguistics is still rather small – and yet the web has already been tried for tasks on various linguistic levels: lexicography, syntax, semantics and translation.

2.1. Lexicography

Almost from its start the web has been a place for the dissemination of lexical resources (word lists in various languages). There are a large number of interfaces to online dictionaries that are more or less carefully administered. But the more interesting aspect from a computational perspective comes from discovering and classifying new lexical material from the wealth of texts in the web. This includes finding and classifying new words or expressions and gathering additional information such as typical collocations, sub-categorization requirements, or definitions.

Jacquemin and Bush (2000) describe an approach to learning and classifying proper names from the web. This is a worthwhile undertaking since proper names are an open word class with new names being continually invented. Their system works in three steps. First, a harvester downloads web pages retrieved by a search engine following a query to a keyword pattern (e.g *universities such as; the following list of politicians*). Second, shallow parsers are used to extract candidate names from enumerations, lists, tables and anchors. Third, a filtering module cleans the candidates from leading determiners or trailing unrelated words.

The example sentence (1) will lead to the acquisition of the university names *University of Science and Technology* and *University of East Anglia* with acronyms for both names and the location of the former.

> (1) While some universities, such as the University of Science and Technology at Manchester (UMIST) and the University of East Anglia (UEA), already charge students using the internet ...

Names are collected for organizations (companies, universities, financial institutions etc.), persons (politicians, actors, athletes etc.), and locations (countries, cities, mountains etc.). The precision of this

acquisition and classification process is reported as 73.6% if the names are only sorted into wide semantic classes. For fine-grained classes it is 62.8%. These are remarkable results that demonstrate the principal feasibility of the approach.

2.2. Syntax

There are a number of web sites that allow the parsing of sentences at different levels of sophistication (see www.ifi.unizh.ch/cl/volk/-InteractiveTools.html for a collection of links to such systems for English and German). Most are demos that work only for single sentences or on short texts. Again, the more interesting question is: How can the vast textual resources of the web be exploited to improve parsing? In this section we summarize some of our own research on using the web (i.e. frequencies obtained from search engines) to resolve PP attachment ambiguities (see also Volk 2000, 2001).

2.2.1. Disambiguating PP attachment

Any computer system for natural language processing has to struggle with the problem of ambiguities. If the system is meant to extract precise information from a text, these ambiguities must be resolved. One of the most frequent ambiguities arises from the attachment of prepositional phrases (PPs).

An English sentence consisting of the sequence verb + NP + PP is a priori ambiguous. (The same holds true for any German sentence.) The PP in example sentence (2) is a noun attribute and needs to be attached to the noun, but the PP in (3) is an adverbial and thus part of the verb phrase.

(2) Peter reads a book about computers.
(3) Peter reads a book in the subway.

If the subcategorization requirements of the verb or of the competing noun are known, the ambiguity can sometimes be resolved. Often, however, there are no clear requirements. Therefore, there has been a growing interest in using statistical methods that reflect attachment tendencies. The first idea was to compare the cooccurrence frequencies of the pair verb + preposition and of the pair noun + preposition.

However, subsequent research has shown that it is advantageous to include the core noun of the PP in the cooccurrence counts. This means that the cooccurrence frequency is computed over the triples V+P+N2 and N1+P+N2 with N1 being the possible reference noun, P being the head and N2 the core noun of the PP. Of course, the frequencies need to be seen in relation to the overall frequencies of the verb and the noun occurring independently of the preposition. For example sentence (2) we would need the triple frequencies for (*read, about, computer*) and (*book, about, computer*) as well as the unigram frequencies for *read* and *book*.

Obviously it is very difficult to obtain reliable frequency counts for such triples. We therefore used the largest corpus available, the WWW. With the help of a WWW search engine we obtained frequency values ('number of pages found') and used them to compute cooccurrence values. Based on the WWW frequencies we computed the cooccurrence values, using the following formula (X may be either the verb V or the head noun N1):

cooc(X,P,N2) = freq(X,P,N2) / freq(X)

For example, if some noun N1 occurs 100 times in a corpus and this noun co-occurs with the PP (defined by P and N2) 20 times, then the cooccurrence value cooc(N1,P,N2) will be 20 / 100 = 0.2. The value cooc(V,P,N2) is computed in the same way. The PP attachment decision is then based on the higher cooccurrence value. When using the web, the challenge lies in finding the best query formulation with standard search engines

2.2.2. Query formulation

WWW search engines are not intended for linguistic queries, but for general knowledge queries. For the PP disambiguation task we need cooccurrence frequencies for full verbs + PPs as well as for nouns + PPs. From a linguistic point of view we will have to query for

1) a noun N1 occurring in the same phrase as the PP that contains the preposition P and the core noun N2. The immediate sequence of N1 and the PP is the typical case for a PP attached to a noun but there are numerous variations with intervening genitive attributes or other PPs.
2) a full verb V occurring in the same clause as the PP that contains the preposition P and the core noun N2. Unlike in English, the

German verb may occur in front of the PP or behind the PP depending on the type of clause.

Since we cannot query with these linguistic operators ('in the same phrase', 'in the same clause'), we have approximated these cooccurrence constraints with the available operators. We used the NEAR operator (V NEAR P NEAR N2). In AltaVista this operator restricts the search to documents in which its argument words co-occur within 10 words.

Let us illustrate the method with example (4) which we took from a computer magazine text. It contains the PP *unter dem Dach* following the noun *Aktivitäten*. The PP could be attached to this noun or to the verb *bündeln*.

(4) Die deutsche Hewlett-Packard wird mit Beginn des neuen Geschäftsjahres ihre Aktivitäten **unter dem Dach** einer Holding bündeln.[Hewlett-Packard Germany will bundle its activities under the roof of a holding at the beginning of the new business year.]

We queried AltaVista for the following triple and unigram frequencies, which led to the cooccurrence values in column 4. Since the value for the verb triple is higher than for the noun triple, the method will correctly predict verb attachment for the PP.

(X,P,N2)	freq(X NEAR P NEAR N2)	freq(X)	cooc(X,P,N2)
(*Aktivitäten, unter, Dach*)	58	246,238	0.00024
(*bündeln, unter, Dach*)	47	14,674	0.00320

For the evaluation of the method we have extracted 4383 test cases with an ambiguously positioned PP from a German treebank. 61% of these test cases have been manually judged as noun attachments and 39% as verb attachments. With the triple cooccurrence values computed from web queries we were able to decide 63% of the test cases, and we observed an attachment accuracy (percentage of correctly disambiguated cases) of 75% over the decidable cases.

This result is based on using the word forms as they appear in the sentences, i.e. possibly inflected verbs and nouns, some of which are rare forms and lead to zero frequencies for the triples. Since a coverage of 63% is rather low, we experimented with additional queries for base forms. We combined the frequency counts for base

forms and surface forms and in this way increased the coverage to 71% (the accuracy stayed at 75%).

The good news is that such simple queries to standard search engines can be used to disambiguate PP attachments. The result of 75% correct attachments is 14% better than pure guessing (of noun attachments), which would lead to 61% correct attachments. The negative side is that these accuracy and coverage values are significantly lower than our results from using a medium size (6 million words) local accessible corpus with shallow corpus annotation. Frequencies derived from that corpus led to 82% accuracy for a coverage of 90%. Obviously, using the NEAR operator introduces too much noise into the frequency counts.

Many of the unresolved test cases involved proper names (person names, company names, product names) as either N1 or N2. Triples with names are likely to result in zero frequencies for WWW queries. One way of avoiding this bottleneck is proper name classification and querying for well-known (i.e. frequently used) representatives of the classes. As an example, we might replace the (N1,P,N2) triple *Computer von Robertson Stephens & Co.* by *Computer von IBM*, query for the latter and use the cooccurrence value for the former. Of course, it would be even better if we could query for *Computer von <company>*, which would match any company name. But such operators are currently not available in web search engines.

One way to escape this dilemma is the implementation of a linguistic search engine that would index the web in the same manner as AltaVista or Google but offer linguistic operators for query formulation. Obviously, any constraint to increase the query precision will reduce the frequency counts and may thus lead to sparse data. The linguistic search engine will therefore have to allow for semantic word classes to counterbalance this problem. We will get back to this in section 3.

Another option is to automatically process (a number of) the web pages that are retrieved by querying a standard WWW search engine. For the purpose of PP attachment, one could think of the following procedure.

1. One queries the search engine for all German documents that contain the noun N1 (or the verb V), possibly restricted to a subject domain.

2. A fixed number of the retrieved pages are automatically loaded. Let us assume the thousand top-ranked pages are loaded via the URLs provided by the search engine.
3. From these documents all sentences that contain the search word are extracted (which requires sentence boundary recognition).
4. The extracted sentences are compiled and subjected to corpus processing (with proper name recognition, part-of-speech tagging, lemmatization etc.) leading to an annotated corpus.
5. The annotated corpus can then be used for the computation of unigram, bigram and triple frequencies.

2.3. Semantics

While gathering semantic information from the web is an ambitious task, it is also the most rewarding with regard to practical applications. If the semantic units in a web site can be classified, retrieval will be much more versatile and precise. In some sense the proper name classification described in section 2.1 can be seen as basic work in semantics.

Agirre *et al* (2000) go beyond this and describe an approach to enriching the WordNet ontology using the WWW. They show that it is possible to automatically create lists of words that are topically related to a WordNet concept. If a word has multiple senses in WordNet, it will be accompanied by synonyms and other related words for every sense. Agirre *et al* query the web for documents exemplifying every sense by using these co-words. The query is composed by using the disjunction of the word in question and its co-words and by the exclusion of all co-words of the competing senses (via the NOT operator). The documents thus retrieved are locally processed and searched for terms that appear more frequently than expected, using the X^2 function. They evaluated the resulting topic signatures (lists of related words) by successfully employing them in word sense disambiguation.

Moreover, they use the topic signatures to cluster word senses. For example, they were able to determine that some WordNet senses of *boy* are closer to each other than others. Their method could be used to reduce WordNet senses to a desired grain size.

2.4. Translation

Translators have found the web a most useful tool for looking up how a certain word or phrase is used. Since queries to standard search engines allow for restrictions to a particular language and, via the URL domain, also to a particular country, it has become easy to obtain usage information which was buried in books and papers (or local databases at best) prior to the advent of the web. In addition to simply browsing through usage examples, one may exploit the frequency information. We will summarize two other examples of how a translator may profit from the web:

2.4.1. Translating compound nouns

Grefenstette (1999) has shown that WWW frequencies can be used to find the correct translation of German compounds if the possible translations of their parts are known. He extracted German compounds from a machine-readable German–English dictionary. Every compound had to be decomposable into two German words found in the dictionary and its English translation had to consist of two words. Based on the compound segments more than one translation was possible. For example, the German noun *Aktienkurs* (*share price*) can be segmented into *Aktie* (*share, stock*) and *Kurs* (*course, price, rate*) both of which have multiple possible translations. By generating all possible translations (*share course, share price, share rate, stock course, ...*) and submitting them to AltaVista queries, Grefenstette obtains WWW frequencies for all possible translations. He tested the hypothesis that the most frequent translation is the correct one.

He extracted 724 German compounds according to the above criteria and found that his method predicted the correct translation for 631 of these compounds (87%). This is an impressive result given the simplicity of the method.

2.4.2. Parallel texts in the web

Translation memory systems have become an indispensable tool for translators in recent years. They store parallel texts in a database and can retrieve a unit (typically a sentence) with its previous translation equivalents when it needs to be translated again. Such systems come

to their full use when a database of the correct subject domain and text type is already stored. They are of no help when few or no entries have been made.

But often previous translations exist and are published in the web. The task is to find these text pairs, judge their translation quality, download and align them, and store them into a translation memory. Furthermore, parallel texts can be used for statistical machine translation.

Resnik (1999) therefore developed a method to automatically find parallel texts in the web. As a first step he used a query to AltaVista by asking for parent pages containing the string "English" within a fixed distance of "German" in anchor text. This generated many good pairs of pages such as those reading "Click here for English version" and "Click here for German version", but of course also many bad pairs.

Therefore he added a filtering step that compares the structural properties of the candidate documents. He exploited the fact that web pages in parallel translations are very similarly structured in terms of HTML mark-up and length of text. A statistical language identification system determines whether the documents found are in the suspected language. 179 automatically found pairs were subjected to human judgement. Resnik reports that 92% of the pairs considered as good by his system were also judged good by the two human experts.

In a second experiment he increased the recall by not only looking for parent pages but also for sibling pages, i.e. pages with a link to their translated counterpart. For English–French he thus obtained more than 16,000 pairs. Further research is required to see how many of these pairs are suitable for entries in a translation memory database.

2.5. Diachronic change

In addition to accessing today's language, the web may also be used to observe language change over time. Some search engines allow restricting a query to documents of a certain time span. And although the web is young and old documents are often removed, there are first examples in which language change is documented.

Let us look at the two competing words for *mobile phone* in Swiss German. When the Swiss telephone company Swisscom first

launched mobile phones, they called them *Natel*. At about the same time mobile phones in Germany were introduced as *Handy*. Ever since, these two words are competing for prominence in Switzerland. And our hypothesis was that *Handy* has become more frequently used because of the open telecom market.

We therefore checked the frequency of occurrence in the web before and after January 1st 2000. We used the Hotbot search engine since it allows this kind of time span search. We queried for all inflected forms of the competing words and restricted the search to German documents in the .ch domain. It turned out that before the year 2000 the number of documents found for *Natel* was about twice the number for *Handy*. For the period after January 2000 the frequency for both is about the same. This is clear evidence that the use of *Handy* is catching up, and it will be interesting to follow whether it will completely wipe out *Natel* in the future.

3. Towards a corpus query tool for the web

Current access to the web is limited in that we can only retrieve information through the bottleneck of search engines. We thus have to live with the operators and options they offer. But these search engines are not tuned to the needs of linguists. For instance, it is not possible to query for documents that contain the English word *can* as a noun. Therefore we call for the development of a linguistic search engine that is designed after the most powerful corpus query tools. Of course their power depends on what kind of linguistic processing they can apply.

We checked the query languages of the Cosmas query tool at the "Institut für deutsche Sprache"[2] and the corpus query language for the British National Corpus (BNC). The Cosmas query tool knows Boolean operators, distance operators, wildcards and form operators ("all inflected forms" and "ignore capitalization"). In particular the distance operators for words, sentences and paragraphs are positive. In comparison however, the query language for the BNC is much more powerful. Since the texts are part-of-speech tagged and marked up with SGML, this information can be accessed through the queries.

---

[2] See http://corpora.ids-mannheim.de/~cosmas/ to query the corpora at the "Institut für deutsche Sprache".

Let us go through the requirements for an ideal corpus query tool and the operators that it should comprise. The operators typically work on words but sometimes it is desirable to access smaller units (letters, syllables, morphemes) or larger units (phrases, clauses, sentences).

1. Boolean operators (AND, OR, NOT): combining search units by logical operators. These operators are the most basic and they are available in most search engines (on the word level). Combining Boolean operators with some of the other operators may slow down retrieval time and is therefore often restricted.
2. Regular expressions: widening the search by substituting single search units (letters, words) or groups by special symbols. The most common are the Kleene star (*) substituting a sequence of units and the question mark (?) substituting exactly one unit.
3. Distance operators: restricting the search to a specific distance between search units (word, phrase, clause, sentence, paragraph, chapter, document; e.g. find *bank* within 2 sentences of *money*) often combined with a direction (to the left or right; e.g. find *with* within 2 words to the left of *love*).
4. Environment operators: restricting the search to a specific section of a document (e.g. header, abstract, quote, list item, bibliographic list; e.g. find *Bill Gates* in headers).
5. Form operators: widening the search by abstracting from a specific form. They include the capitalization operator (ignore upper and lower case) and the inflection operator (use all inflected forms).
6. Syntactic class operators: restricting the search to a certain part-of-speech (e.g. *with* followed by any noun), or phrase (e.g. an accusative NP followed by a PP).
7. Subject domain operators: restricting the search to a specific subject domain (e.g. linguistics, computer science, chemistry etc.).
8. Text type operators: restricting the search to a specific text type (e.g. newspaper articles, technical manuals, research reports).
9. Semantic class operators: restricting the search to e.g. proper name classes (person, location, organization, product), temporal and local phrases; causal relations, or certain word senses (e.g. find *bank* in the sense of *financial institution*); synonym and hyperonym searches; definitory contexts.

A search engine stores information about a document at the time of indexing (e.g. its location, date and language). This speeds up online retrieval but requires offline processing of the documents. And some of the operators rely on information that is difficult and costly to compute over large amounts of text.

Therefore it might be advisable to use offline processing as suggested by (Corley *et al* 2001). They describe the Gsearch system which allows finding syntactic structure in unparsed corpora. The idea is that the Gsearch user provides the system with context free grammar rules that describe constituents for the syntactic phenomenon under investigation. Gsearch's bottom up chart parser processes every sentence and checks whether it can apply the grammar rules and whether they lead to the search goal. For example, a grammar might specify rules for NPs and PPs. Then the search goal might be to find sentences which contain a sequence of a verb, an NP and a PP for the investigation of PP attachments. Gsearch is not intended for accurate unsupervised parsing but as a tool for corpus linguists. Since its grammar is exchangeable, it can be adapted to the specific needs of the linguist.

4. Conclusion

We have shown that the web offers a multitude of opportunities for corpus research. Access to this corpus is ubiquitous and its size exceeds all previous corpora. Currently the access is limited by search engines that are not tuned to linguistic needs. We therefore propose to use these search engines only for the preselection of documents and add linguistic postprocessing. A better but much more complex solution is the implementation of a special purpose linguistic search engine.

Kilgarriff (2001) has sketched an intermediate solution of organizing a controlled corpus distributed over various web servers. This would escape some of the problems that we currently encounter in web-based studies. The web is a very heterogeneous collection of documents. Many documents do not contain running text but rather lists, indexes or tables. If these are not filtered out, they might spoil collocations or other types of cooccurrence statistics. In addition, the web is not a balanced corpus, neither with respect to text types nor with respect to subject areas. Agirre *et al* (2000) found that sex

related web pages strongly influenced their word sense disambiguation experiments for the word *boy*.

When working with low occurrence frequencies, one also has to beware of negative examples. For our study on German prepositions we checked which of them have corresponding pronominal adverbs. This test is used for determining whether the preposition can introduce a prepositional object, since pronominal adverbs are proforms for such objects. It is taken for granted that the preposition *ohne* (without) is an exception to this rule. It can introduce a prepositional object, but it does not form a pronominal adverb. But a Google query for the questionable pronominal adverb form *darohne* results in 14 hits. At first sight this seems to contradict the linguistic assumptions. But a closer look reveals that many of these hits lead to literary texts written more than a century ago while others discuss the fact that the form *darohne* does not exist anymore.

We believe that harvesting the web for linguistic purposes has only begun and will develop into an important branch for Computational Linguistics. We envision that future NLP systems will query the web whenever they need information that is not locally accessible. We think of parsers that will query the web for the resolution of structural or sense ambiguities. Or of MT systems that perform automatic lexical acquisition over the web to fill their lexicon gaps before translating a text.

One of the main arguments against using the web is its constant change. A result computed today may not be exactly reproducible tomorrow. But, as Kilgarriff (2001) notes, this is same for the water in any river and nobody will conclude that investigating water molecules is therefore senseless. We will all have to learn to fish in the waters of the web.

References

Agirre, E.; Olatz, A.; Hovy, E.; Martinez, D. 2000. Enriching Very Large Ontologies Using the WWW. ECAI 2000, Workshop on Ontology Learning. Berlin.

Biber, D.; Conrad, S.; Reppen, R. 1998. Corpus Linguistics. Investigating Language Structure and Use. Cambridge University Press.

Corley, S.; Corley, M.; Keller, F.; Crocker, M.W.; Trewin, S. 2001. Finding syntactic structure in unparsed corpora. – Computers and the Humanities. 35:2, 81–94.

Grefenstette, Gregory 1999. The World Wide Web as a resource for example-based machine translation tasks. – Proceedings of Aslib Conference on Translating and the Computer 21. London.

Jacquemin, C.; Bush C. 2000. Combining lexical and formatting clues for named entity acquisition from the web. – Proceedings of Joint SIGDAT Conference on Empirical Methods in NLP and Very Large Corpora. Hongkong. 181–189.

Kennedy, G. 1998. An Introduction to Corpus Linguistics. London: Addison Wesley Longman..

Kilgarriff, Adam 2001. The web as corpus. – Proceedings of Corpus Linguistics 2001. Lancaster.

Mason, O. 2000. Java Programming for Corpus Linguistics. Edinburgh Textbooks in Empirical Linguistics. Edinburgh University Press.

Oakes, M. 1998. Statistics for Corpus Linguistics. Edinburgh Textbooks in Empirical Linguistics. Edinburgh University Press.

Resnik, Philip 1999. Mining the web for bilingual text. – Proceedings of 37th Meeting of the ACL. Maryland. 527–534.

Volk, Martin 2000. Scaling up. Using the WWW to resolve PP attachment ambiguities. – Proceedings of Konvens-2000. Sprachkommunikation. Ilmenau.

Volk, Martin 2001. Exploiting the WWW as a corpus to resolve PP attachment ambiguities. – Proceedings of Corpus Linguistics 2001. Lancaster.

Haldur Õimu tööde bibliograafia 1964–2001

1964

Keeleteadus ja matemaatika. – TRÜ, 1964, 14. märts.

1965

Eesti keele verbide formaalsest semantilisest analüüsist. Diplomitöö. Tartu, 1965, 125 lk. [Käsikiri Tartu Ülikooli eesti keele õppetoolis.]

tulema, saama ja *pidama* tähenduste strukturaalne analüüs. – Keel ja struktuur 1. Töid strukturaalse ja matemaatilise lingvistika alalt. Tartu: TRÜ, 1965, lk. 27–45.

1966

Polüseemiliste sõnade tähenduste eristamisest. – Keele modelleerimise probleeme I. Tartu: TRÜ, 1966, lk. 197–219.

1967

О селекционных ограничениях внутри именной фразы. – Межвузовская конференция по порождающим грамматикам. Кяэрику, 15–25 сентября 1967. Тезисы докладов. Тарту, 1967, с. 136.

1969

Semantic Theory and the Category of Predication. – Generatiivse grammatika grupi aastakoosoleku teesid. Annual Meeting of the Research Group for Generative Grammar 1969. Abstracts. Tartu: TRÜ, 1969, pp. 51–57.

On the Semantic Representations of Predicates. – Generatiivse grammatika grupi aastakoosoleku teesid. Annual Meeting of the Research Group for Generative Grammar 1969. Abstracts. Tartu: TRÜ, 1969, pp. 58–71.

Kas inimkeel on päritav? – Noorus 1969, nr. 4, lk. 74–75.

Kuidas on keel ehitatud inimese ajus? – Noorus 1969, nr. 8, lk. 74–75.

О лексических категориях в глубинной структуре. – Keele modelleerimise probleeme 3.2. Tartu Riikliku Ülikooli toimetised, vihik 228. Tartu: TRÜ, 1969, с. 197–208.

Erelt, M., Kaplinski, J., Rätsep, H., Uuspõld, E., Õim, H. Lähtugem keeleteoreetilistest seisukohtadest! – Keel ja Kirjandus 1969, nr. 2, lk. 105–106.

Ivič, M. Keeleteaduse põhisuunad. [Inglise keelest tõlkinud Erelt, M. Hint, M., Kaplinski, J., Remmel, M., Rätsep, H., Uuspõld, E., Viitso, T.-R., Õim, H.] Tartu: TRÜ, 1969. 302 lk.

1970

Isiku mõistega seotud sõnarühmade semantiline struktuur eesti keeles. Väitekiri filoloogiakandidaadi kraadi taotlemiseks. Tartu, 1970. 263 lk. [Käsikiri Tartu Ülikooli eesti keele õppetoolis.]

Eesmärk, taotlema, saavutama, tulemus semantiline analüüs. – Keel ja struktuur 4. Töid strukturaalse ja matemaatilise lingvistika alalt. Tartu: TRÜ, 1970, lk. 71–108.

Semantikast. (Keele väljendite vahekordade uurimisest). – Noorus 1970, nr. 11, lk. 71–73.

Erelt, M. ja Õim, H. Раянди Х. Синтаксис эстонского имперсонала и пассива. H. Rajandi, Eesti impersonaali ja passiivi süntaks. Väitekiri filoloogiakandidaadi kraadi taotlemiseks. Tallinn 1968. – Советское финно-угроведение 1970, № 1, с. 73–78.

On Pronominalization and Semantic Treatment of Sentences. – Generatiivse grammatika grupi aastakoosoleku teesid. Annual Meeting of the Research Group for Generative Grammar 1970. Abstracts. Tartu: TRÜ, 1970, pp. 58–63.

Semantic Analysis of Estonian Abstract Nouns and the "Underlying Situations". – Congressus Tertius Internationalis Fenno-Ugristarum. Tallinn, 17.–23. VIII 1970. Teesid, lk. 88.

On the relationship between syntactic and semantic representations (with special reference to the semantic analysis of some Estonian words). – Советское финно-угроведение 1970, № 1, с. 25–36.

К истолкованию понятия предикации в рамках порождающей семантики. – Структурно-математические методы моделирования языка. Тезисы докладов и сообщений весоюзной научной конференции. Часть 2. Киев, 1970, с. 148–149.

Кулль И., Ыйм Х. О некоторых проблемах составления больших тезаурусов на основе текстов естественного языка. – Структурно-математические методы моделирования языка.Тезисы докладов и сообщений весоюзной научной конференции. Часть 2. Киев, 1970, с. 73.

Семантическая структура словесных груп, связанных с понятием лица в эстонском языке. 10.661 – Языки народов СССР (эстонский язык). Автореферат диссертации на соискание ученой степени кандитата филологических наук. [Isiku mõistega seotud sõnarühmade semantiline struktuur eesti keeles. Kandidaadiväitekirja autoreferaat.] Тарту, 1970. 26 с.

1971

Isiku mõistega seotud sõnarühmade semantiline struktuur eesti keeles. – Keele modelleerimise probleeme 4. Tartu, TRÜ Toimetised, vihik 278, 1971, lk. 5–260.

Predicates, Nouns and Referring Structures. – Generatiivse grammatika grupi aastakoosoleku teesid. Annual Meeting of the Research Group for Generative Grammar 1971. Abstracts. Tartu: TRÜ, 1971, pp. 62–73.

Eesmärk 'Ziel', *taotlema* 'anstreben', *saavutama* 'erreichen', *tulemus* 'Ergebnis'. Semantiche Analyse. – ASG, 1971, Bericht Nr. 10. 42 s. (Deutche Akademie der Wissenschaften zu Berlin. Zentralinstitut für Sprachwissenschaft. Arbeitsgruppe Strukturelle Grammatik).

1972

Semantics and Functional Description of Sentences. – Generatiivse grammatika grupi aastakoosoleku teesid. Annual Meeting of the Research Group for Generative Grammar 1972. Abstracts. Tartu: TRÜ, 1972, pp. 42–50.

G. Lakoffi raamatust "Lingvistika ja loomulik loogika." – Keel ja struktuur 6. Töid strukturaalse ja matemaatilise lingvistika alalt. Tartu: TRÜ, 1972, lk. 139–150.

Lauseiden funktionaalisesta analyysista semanttiikassa. – Lauseopin ja semanttiikan päivät. Tampere, 23.–24.9.1972. Preprint.

1973

On the Semantic Treatment of Predicative Expressions. – Generative Grammar in Europe. Ed. by N. Ruwet and F. Kiefer. Dordrecht, Holland: Reidel, 1973, pp. 360–386.

Presuppositions and the Ordering of Messages. – Trends in Soviet Theoretical Linguistics. Foundations of Language: Supplemantary Series, Vol. 18. Ed. by F. Kiefer. Dordrecht, Boston, 1973, pp. 123–134.

Кулль И., Сильдмяэ И., Хелемяэ А., Ыйм Х. О разработке тезауруса юридических терминов для информационно-поисковой системы. – Правовая кибернетика. Москва: Наука, 1973, с. 54–62.

Keel, keeleteadus ja pragmaatika. – Keel ja struktuur 8. Töid strukturaalse ja matemaatilise lingvistika alalt. Tartu: TRÜ, 1973, lk. 107–147.

1974

Semantika. Tallinn: Valgus, 1974. 167 lk. (Mosaiik 7)

Keele automaattöötlus ja automatiseeritud infosüsteemid. – Keel ja Kirjandus 1974, nr. 8, lk. 469–480; nr. 9, lk. 545–55.

1975

Сильдмяэ И., Хелемяэ А., Ыйм Х. О лингвистическом аспекте описания естественного языка как языка программирования. – Труды IV Международной обьединенной конференции по искуственному интеллекту – Тбилиси 1975. Дополнительные материалы. Москва, 1975, с. 207–216.

Семантический анализ эстонских абстрактных слов и структура "семантических полей". – Congressus Tertius Internationalis Fenno-Ugristarum. Tallinae habitus, 17–23. VIII 1970. Pars I: Acta Linguistica. Tallinn: Valgus, 1975, lk. 310–313.

Väitekiri parömioloogia alalt. (Folklorist A. Krikmanni väitekirja kaitsmise puhul). – Keel ja Kirjandus 1975, nr. 9, lk. 571–572.

Сильдмяэ И., Ыйм Х. Семантический анализ нормативных актов. – Актуальные проблемы теории и практики применения математических методов и ЭВМ в деятельности органов юстиции. Тезисы докладов всесоюзной конференции по проблемам правовой кибернетики 2. Москва, 1975, с. 131–134.

1976

Elementaartähendused keele semantilises struktuuris. – Keel ja Kirjandus 1976, nr. 10, lk. 598–605; nr. 11, lk. 675–682.

Keelelise interaktsiooni sõnarühma semantiline analüüs. – Keel ja struktuur 9. Tartu: TRÜ, 1976, lk. 16–68.

Valge, J., Õim, H. Teooria praktikasse: automatiseeritud infosüsteemid. – Keel, mida me uurime. Koost. M. Mäger. Tallinn: Valgus, 1976, lk. 18–22.

Kuidas on keel ehitatud inimese ajus? [Varem avaldatud ajakirjas Noorus 1969, nr. 8, lk. 74–75.] – Keel, mida me uurime. Koost. M. Mäger. Tallinn: Valgus, 1976, lk. 34–37.

Kas inimkeel on päritav? [Varem avaldatud ajakirjas Noorus 1969, nr 4, lk. 74–75.] – Keel, mida me uurime. Koost. M. Mäger. Tallinn: Valgus, 1976, lk. 158–161.

Semantikast. – Keel, mida me harime. Koost. M. Mäger. Tallinn: Valgus, 1976, lk. 125–129.

Сильдмяэ, И., Хелемяэ, А., Ыйм Х. Формализация языка для общения с ЭВМ. – Ученые записки ТГУ выпуск 376. Тарту: ТГУ, 1976, с. 81–91.

Сильдмяэ, И., Ыйм Х. О структурном анализе понятий действий и о правовом регулирований действий. – Ученые записки ТГУ выпуск 376. Тарту: ТГУ, 1976, с. 64–80.

1977

Towards a theory of linguistic pragmatics. – Journal of Pragmatics, Vol. 1, 1977, No 3, pp. 251–268.

Messages and memory structures in the pragmatic description of sentences. – SMIL: Journal of Linguistic Calculus 1977, No 2, pp. 4–43.

Sildmäe, I., Õim, H., Ääremaa, K. Understanding normative texts. – First International Congress on Legal Science. The Hague (Netherlands), 1977, pp. 47–48.

Сильдмяэ, И., Ыйм Х. Оценочные суждения и структура языковой интеракции. – Семантические вопросы искусственного интеллекта. Киев, 1977, с. 48–50.

Салувеэр М., Сильдмяэ, И., Ыйм Х. Оценочные суждения и структура языковой интерактции. – Семантические вопросы искусственного интеллекта. Киев, общество "Знание" Укр. ССР, 1977, с. 14.

Сильдмяэ, И., Ыйм Х., Ээремаа К. Об идеологии юридических автоматизированных системах. Автоматизированные системы правовой информации. Братислава, 1977, с. 146–157.

Сильдмяэ, И., Ыйм Х. Автоматизация семантической обработки нормативных текстов. – Вопросы кибернетиики 1977, № 40, с. 81–85.

1978

Tähelepanekuid teoreetilisest keeleteadusest Lääne-Euroopas. – Keel ja Kirjandus 1978, nr. 4, lk. 223–226.

Helemäe, A., Õim, H. Üleliiduline kompuuterlingvistika seminar Otepääl. – Keel ja Kirjandus 1978, nr. 6, lk. 382–383.

Решения, деистви я и язык. – Проблемы моделирования языковой интеракции. Труды по искусственному интеллекту I. Ученые записки ТГУ выпуск 472.Тарту: ТГУ, 1978, с. 18–37.

Хелемяэ, А., Ыйм Х. Два замечания о семантическом компоненты модели "смысль<=>текст". – Проблемы моделирования языковой интеракции. Труды по искусственному интеллекту I. Ученые записки ТГУ выпуск 472.Тарту: ТГУ, 1978, с. 93–100.

Ыйм Х., Салувеэр М. Фреймы и понимание языка. – Проблемы моделирования языковой интеракции. Труды по искусственному интеллекту I. Ученые записки ТГУ выпуск 472. Тарту: ТГУ, 1978, с. 101–113.

Об организации информации в семантической структуры слов выражающие действия. – Linguistica X. Ученые записки ТГУ выпуск 453. Тарту: ТГУ, 1978, с. 147–158.

Сильдмяэ, И., Ыйм Х., Ээремаа К. Автоматизированная система поиска нормативной информации. – Ученые записки ТГУ выпуск 444. Тарту: ТГУ, 1978, с. 90–103.

Некоторые проблемы описания семантической структуры слов обозначающих целенаправленные действия. – Всесоюзная научная конференция по проблемам семантики. Рига 16–18 марта 1978. Рига, 1978.

1979

Проблемы понимания связного текста. – Взаимодействие с ЭВМ на естественном языке 2. Новосибирск, 1979, с. 180–193.

Проблемы понимания связного текста. – Синтаксический и семантический компонент лингвистического обеспечения. Сборник научных трудов. ВЦ СО АН СССР, Новосибирск, 1979, с. 19–32.

К описанию структуры диалога. – Проблемы общего и прикладного языкознания. Linguistica XI. Ученые записки ТГУ выпуск 502. Тарту: ТГУ 1979, с. 157–168.

1980

Elektronarvutid ja eesti keel. – Taas emakeele lätteil. Tallinn, 1980, lk. 56–62.

Эписоды в структуре дискурса. – Представление знаний и моделирования процесса понимания. ВЦ СО АН СССР, Новосибирск, 1980, с. 79–96.

Language, meaning and human knowledge. – Grammar and Semantics. Academy of Sciences of the Estonian SSR, Institute of Language and Literature. Tallinn: Valgus, 1980, pp. 5–49.

Raamat eesti keele struktuurist. [Rets.: H. Rätsep. Eesti keele lihtlausete tüübid. Tln. 1978.] – Sirp ja Vasar 1980, 1. mai.

Язык, значения и знания. – Семантика и представление знаний. Труды по искусственному интеллекту II. Ученые записки ТГУ выпуск 519. Тарту: ТГУ, 1980, с. 117–129.

Литвак С., Роосмаа Т., Салувеэр М., Ыйм Х. Об автоматическом морфологическом анализе ограниченного естественного языка. – Логико-семантические вопросы искусственного интеллекта. Труды по искусственному интеллекту III. Ученые записки ТГУ выпуск 551. Тарту: ТГУ, 1980, с. 82–86.

Литвак С., Роосмаа Т., Салувеэр М., Ыйм Х. Подсистема автоматического синтаксического анализа для эксперементальной ВОС. – Логико-семантические вопросы искусственного интеллекта. Труды по искусственному интеллекту III. Ученые записки ТГУ выпуск 519. Тарту: ТГУ, 1980, с. 87–91.

1981

Language, Meaning and Human Knowledge. – Nordic Journal of Linguistics, Vol. 4, 1981, No 2, pp. 67–90.

A Constrative Model of Text Comprehension. – AILA 81. Proceedings 1. Sections and workshops. Abstracts. Lund 1981, pp. 410–411.

Tööjuhend ja programm kaasaegsest eesti keelest TRÜ filoloogiateaduskonna eesti filoloogia osakonna IV–V kursuse kaugüliõpilastele. [Koost. R. Kasik, H. Õim. Tartu 1981, 6 lk.]

Teoreetilise keeleteaduse vanast ja uuest paradigmast. – Keel ja Kirjandus 1981, nr. 7, lk. 385–391; nr. 8, lk. 456–464.

Tekst ja selle mõistmine. Eesti keele grammatika probleeme. – Töid eesti filoloogia alalt VIII. TRÜ Toimetised, vihik 574. Tartu, 1981, lk. 106–122.

Литвак С., Роосмаа Т., Салувеэр М., Ыйм Х. О распознавании гиперсобытий в системе понимания связного текста. – Диалоговые системы и представление знаний. Труды по искусственному интеллекту IV. Ученые записки ТГУ выпуск 594. Тарту: ТГУ, 1981, с. 56–70.

Литвак С., Роосмаа Т., Салувеэр М., Ыйм Х. Работа с базой знаний в системе понимания связного текста. – IX всесоюзный симпозиум по кибернетики. Тезисы симпозиума Сухуми. Представление знании. Тезисы, т. 1. Москва, 1981, с. 169–170.

Литвак С., Роосмаа Т., Салувеэр М., Ыйм Х. Лингвистический процессор системы TARLUS. – Диалог в автоматизированных системах. Материалы семинара. Москва: МДТНП, 1981, с. 72–76.

Литвак С., Роосмаа Т., Салувеэр М., Ыйм Х. О проблеме функционирования базы знаний в системе понимания связного текста. – Тезисы докдадов 2-го всесоюзного симпозиума "Диалоговые и фактографические системы информационного обеспечения". Суздаль, 1981, с. 112–114.

1982

Tekstisemantiikan vaikutuksista sanasemantiikkaan. – Psykolingvistisia Kirjoituksia III. Jyväskylä: AFinLA, 1982, s. 101–107.

Litvak, S., Roosmaa, T., Saluveer, M., Õim, H. On the Interaction of Knowledge Representation and Reasoning Mechanism in Discourse Comprehension. – 1982 European Conference on Artificial Intelligence – ECAI-82. Conference Proceedings. Orsay (France), 1982, pp. 125–126.

Research at Tartu State University. – SIGART Newsletter, 1982, 79, pp. 138–140.

1983

Семантика и теория понимания языка. Анализ лексики и текстов директивного общения эстонского языка. [Semantika ja keele mõistmise teooria. Eesti keele direktiivse suhtluse leksikoni ja tekstide analüüs.] 10. 02. 07. Финно-угорские языки, 10. 02. Общее языкознание. Автореферат диссертации на соискание ученой степени доктора филологических наук. Тартуский государственный университет. Тарту, 1983. 40 с.

Inimene, keel ja arvuti ehk kompuuterlingvistika. Tallinn: Valgus, 1983. 144 lk. (Mosaiik 33).

Ühest põhimõttelisest süntaksikäsitlusest. [Rets: Erelt, M. Eesti adjektiivide süntaks. Gradatsioon. Tallinn, 1977; Erelt, M. Eesti lihtlause probleeme. Tallinn, 1979; Erelt, M. Predikatiivne adjektiiv (lausemallid). Tallinn, 1979; Erelt, M. Adjektiiv-atribuut eesti keeles. Tallinn, 1980.] – Keel ja Kirjandus 1983, nr. 10, lk. 583–588.

Sanankäyttövirheistä sekä viron ja suomen sanastoeroista. – Folia Fennistica & Linguistica. Suomalais–virolainen virheanalyysiseminaari Summassaaressa 18. ja 19. toukokuutta 1983. Tampereen yliopiston Suomen kielen ja yleisen kielitieteen laitoksen julkaisuja 10. Tampere, 1983, ss. 25–32. [Ka eesti keeles: Sõnastusvigu ja eesti–soome leksikaalseid erinevusi. lk. 51–58.]

Ыйм Х., Салувеэр М. Использование локальных структур в процессе понимания дискурса. – Материалы школы-семинара "Семиотические аспекты формализации интеллектуальной деятельности" (Телави). Тезисы докладов и сообщений. Москва, 1983, с. 150–152.

Ыйм Х., Койт М., Литвак С., Роосмаа Т., Салувеэр М. Система с полным циклом понимания ОЕЯ TARLUS: текущее состаяние. – Международный семинар по машинному переводу. Тезисы докладов. Москва, 1983, с. 249–252.

Koit, M., Litvak, S., Roosmaa, T., Saluveer, M., Õim, H. Using frames in causal reasoning. – Механизмы вывода и обработки знаний в системах понимания текста. Труды по искусственному интеллекту. Ученые записки ТГУ выпуск 621. Тарту: ТГУ, 1983, с. 41–55.

Ыйм Х., Койт М., Литвак С., Роосмаа Т., Салувеэр М. Построение и текущее пополнение динамической модели предметной области. – Международный симпозиум по искусственному интеллекту. Ленинград, 1983. 11 с.

Koit, M., Litvak, S., Roosmaa, T., Saluveer, M., Õim, H. Local and global structures in discourse understanding. – Abstracts of the First Conference of the European Chapter of the American Association for Computational Linguistics. Stanford, 1983, pp.152–154.

1984

Õim, H., Koit, M., Litvak, S., Roosmaa, T., Saluveer, M. Reasoning and discourse: experts as a link between high and low level inferences. – Принципиальные вопросы теорий знаний. Труды по искусственному интеллекту. Ученые записки ТГУ выпуск 688. Тарту: ТГУ, 1984, с. 176–190.

Õim, H., Koit, M., Litvak, S., Roosmaa, T., Saluveer, M. Constructing and updating a dynamic model of a problem domain. – Artificial Intelligence on Automatic Control (IFAC). Symposium Leningrad 1983. 1984, pp. 93–100.

Õim, H., Koit, M., Litvak, S., Saluveer, M., Roosmaa. Tasking of combinatorial explosion: experts as a link between high and low level inferences. – Proceedings of ECAI-1984. European Conference on Artificial Intelligence. Appendix, 1984.

Viron lauseoppia englanniksi. [Rets: Valter Tauli. Standard Estonian grammar. Part II. Syntax. Acta Universitatis Upsaliensis. Studia Uralica et Altaica. Upsaliensia 14. Uppsala, 1983. 359 s.] – Virittäja 1984, s. 121–124.

Ыйм Х., Койт М., Литвак С., Роосмаа Т., Салувеэр М. Некоторые принципиальные вопросы построения естественно-языкого диалога в системах искусственного интеллекта. – Проблемы искусственного интеллекта и распознавания образов. Научная конференция с участием ученых из социалистических стран. Секция I: Искусственный интеллект: Тезисы докладов и сообщений. Киев, 1984, с. 194–197.

Койт М., Ыйм Х. Естественный диалог при взаимодействии с экспертными. системами. – Тезисы докладов республиканской научно-технической конференции "Интерактивные системы и их практическое применение". Кишинев, 1984, с. 14–18.

Keel ja suhtlemine. – Looming 1984, nr. 2, lk. 282–287.

1985

Saluveer, M., Õim, H. Frames in linguistic descriptions. – Quaderni di Semantica Vol. 6, 1985, No 2, pp. 282 –292.

Elektronarvutid ja eesti keel. – Kodumurre 17. Tallinn, 1985, lk. 22–29.

Койт М., Ыйм Х. Общение человека с ЭВМ: знания о диалоге. – Проблемы построения проблемно-ориентированных диалоговых систем. Труды республиканской конференции Батуми. Секция 2. Тбилиси, 1985, с. 214–223.

Ыйм Х., Койт М., Литвак С., Роосмаа Т., Салувеэр М. Построение прагматической интерпретации реплик в человеко-машинном диалоге. – IV Всесоюзная конференция "Диалог человека с ЭВМ". Тезисы докладов. Киев, 1985, с. 160–161.

Ыйм Х., Койт М., Литвак С., Роосмаа Т., Салувеэр М. Построение прагматической интерпретации реплик в человеко-машинном диалоге. – IV Всесоюзная конференция "Диалог человека с ЭВМ". Доклады. Киев, 1985, с. 159–166.

Койт М., Ыйм Х. Построение формальной модели диалога. – Семиотические аспекты формализации интеллектуальной деятельности. Школа-семинар "Кутаиси-85". Тезисы докладов и сообщений. Москва, 1985, с. 401–404.

1986

Pragmaatika ja keelelise suhtlemise teooria. – Keel ja Kirjandus 1986, nr. 5, lk. 257–269.

Saluveer, M., Õim, H. Rules and Reasoning in Text Comprehension. – New Approaches in Maschine Translation. Sprache und Information; Bd. 13. Ed. by I. Batori, H. J. Weber. Tübingen: Niemeyer, 1986, pp. 139–163.

Where do communicative acts come from. – Pragmatics and Linguistics: Festschrift for Jacob L. Mey. Odense University Studies in Linguistics, Vol. 15. Odense: Odense University Press, 1986, pp. 129–135.

Tauli Valter: Standard Estonian grammar. Part II. Syntax. Uppsala 1983. Acta Universitatis Upsaliensis. Studia Uralica et Altaica. Upsaliensia 14. Uppsala, 1983. 359 s. [Rets.] – Ural-Altaische Jahrbücher. Neue Folge, 1986, s. 249–253.

Emakeele olümpiaadi tulemused. – Nõukogude Õpetaja 1986, 26. aprill.

Emakeele olümpiaad keskkooliõpilastele. – Edasi 1986, 16. mai.

1987

Communicative strategies and related concepts in a model of dialogue. – XIV International Congress of Linguistics. Section papers. Berlin, 1987, pp. 370–375.

Keel ja keele mõistmine tehisintellektis. – Ars Grammatica 1986. Eesti NSV Teaduste Akadeemia Keele ja Kirjanduse Instituut. Tallinn, 1987, lk. 107–125.

Разработка модели пользования в системах общения человека с ЭВМ. – Тезисы докладов II всесоюзной конференции по проблемам информатики и вычислительной техники. Ереван, 1987, с. 392–393.

Прагматика речевого общения. – Интеллектуальные процессы и их моделирование. Москва: Наука, 1987, с. 196–207.

Реализация модели коммуниканта в семантической структуре текста. – Семантика целого текста. Тезисы выступлений на совещание Одесса. 1987. Москва, 1987, с. 166

1988

Койт М., Ыйм Х. Понятие коммуникативной стратегии в модели общения. – Психологические проблемы познания действительности. Труды по искусственному интеллекту. Ученые записки ТГУ выпуск 793. Тарту: ТГУ, 1988, с. 97–111.

Моделирование естественных рассуждений в интеллектуальных системах: необходимость, возможности, средства. – Тезисы докладов, г. Переславль-замский. Москва, 1988, с. 563–568.

Лингвистические проблемы разработки модели в системах общения человека с ЭВМ. – Прикладная лингвистика и автоматический анализ текста. Тезисы докладов научной конференции. Тарту, 1988, с. 97–98.

1989

Language Understanding and Problem Solving: on the Relation between Computational Linguistics and Artificial Intelligence. – Computational Linguistics. An International Handbook on Computer Oriented Language Research and Applications. Ed. by I. Batory, W. Lenders and W. Putschke. Berlin, New-York: Walter de Gruyter, 1989, pp. 277–283.

Kaks teed keele juurde. – Akadeemia 1989, nr. 3, lk. 559–584.

Roosmaa, T., Õim, H. Loomuliku keele kasutusvõimalusi tehisintellekti süsteemides. – Rakenduslik tehisintellekt ja infosüsteemid. Vabariikliku teaduslik-praktilise seminari teeside kogumik. Tallinn 1989, lk. 13–14.

Принципы естественного рассуждения в модели взаимодействия. – Формальные и неформальные рассуждения. Труды по искусственному интеллекту. Ученые записки ТГУ выпуск 840. Тарту: ТГУ, 1989, с. 172–184.

Huno Rätsep 60: Piirjooni ühele arengujärgule eesti keeleteaduses. – Emakeele Seltsi aastaraamat 1987. Tallinn, 1989, lk. 5–13.

Hennoste, T., Rätsep, H., Õim, H. Eesti keel ja kultuur: ühest programmist. [TRÜ eesti keele uurimise labori tööst.] – Edasi 1989, 12. aprill.

The Estonian language – eesti keel. – Estonia & Tartu. [These esseys have been prepared for the USSR–USA children's summer camp "Angel Bridge" in Tartu, Estonia, August 1989.] Tartu, 1989, pp. 6–7.

1990

Koit, M., Õim, H. An Approach to the Organization of Natural Reasoning. – AIMSA'90 Proceedings. Ed. by P. Jorrand, S. Sgurev. Sofia. 1990, pp. 93–101.

Kognitiivse lähenemise võimalusi keeleteaduses. – Akadeemia 1990, nr. 9, lk. 1818–1838.

Väitekiri keeleteooria alalt. [NSVL TA Keeleteaduse Instituudis kaitses 28. 12. 1989. a. kandidaadiväitekirja Madis Saluveer.] – Edasi 1990, 9. jaanuar.

Suhtlusstrateegiad. – Generatiivse grammatika grupi juubelikoosolek (27. 12. 1990). Teesid. Tartu, 1990, lk. 19.

Возможности моделирования методов конверсационного анализа на ЭВМ. – Актуальные проблемы компьютерной лингвистики. Тезисы докладов всесоюзной конференции. Тарту, 1990, с. 150.

Моделирование естественно-языкового рассуждения (коммуникативный подход). – Когнитивные исследования за рубежом. Методы искусственного интеллекта в моделирования политическог мышления. Москва: Институт США и Канады, 1990, с. 62–87.

Койт М., Ыйм Х. Об одном подходе к моделированию процесса естественного рассуждения. – Исследования по когнитивным аспектам языка. Труды по искусственному интеллекту. Ученые записки ТГУ выпуск 903. Тарту: ТГУ, 1990, с. 91–101.

Модель дискурса и исследование этноспецифического аспекта речевого общения. – Прагматика этноспецифического дискурса. Материалы всесоюзного симпозиума. Беьлцы, 1990, с. 57–59.

1991

Lingvistilised korpused keeleuurimises. – Keel ja Kirjandus 1991, nr. 5, lk. 257–267.

Рассуждения и порождения реплик диалога. – Текст в коммуникации. Москва: Институт языкознания АН СССР, 1991, с. 101–108.

Mati Ereltil on tähtpäev. [Eesti TA KKI grammatikasektori juhataja 50. sünnipäevaks] – Keel ja Kirjandus 1991, nr. 3, lk. 177–178.

The Baltic States. Tallinn, Riga, Vilnius. 1991, p. 76.

Toomas Help. Eesti regulaarne ja irregulaarne verb. Väitekiri filoloogia-kandidaadi teadusliku kraadi taotlemiseks, Tartu 1990. [Rets.] – Linguistica Uralica 1991, nr. 4, lk. 297–299.

1993

Koit, M., Õim, H. Modelling communicative strategies. – Proceedings of the Third Symposium on Programming Languages and Software Tools. University of Tartu, 1993, pp. 73–82.

Koit, M., Õim, H. A Formal Model of Communicative Strategy. Proceedings of the Scandinavian Conference on Artificial Intelligence '93. Stockholm, 1993, pp. 226–231.

Koit, M., Õim, H. Rahvusvahelisel arvutuslingvistika konverentsil. [Arvutuslingvistika Assotsiatsiooni Euroopa Osakonna konverents: Utrecht, Holland] – Keel ja Kirjandus 1993, nr.10, lk. 639–640.

1994

Õim, H., Koit, M. Mõtlemine ja selle mõjutamine tavakujutluses. Kommunikatiivsed strateegiad. – Akadeemia 1994, nr. 1 lk. 25–45.

Metodoloogilis-keeleteaduslikke arutlusi ("Filoloogia teelahkmel"). – Keel ja Kirjandus 1994, nr. 3, lk. 129–132.

Keel ja keeletehnoloogia Euroopa Liidu vaatenurgast. – Postimees, 1994, 16. märts.

Keeletehnoloogia on Euroopa Liidu ametliku strateegilise programmi osa: (Balti riikide ühisseminarist "Keel ja tehnoloogia Euroopas aastal 2000: Balti perspektiiv": Riia). – Postimees, 1994, 30. november.

Arvutuslingvistika ja keeletehnoloogia. [II tekstikorpuse päev.] – Universitas Tartuensis, 1994, 30. september.

1995

Naïve theories: a conception of cognitive semantics. – Cognitive Linguistics, Vol. 6, 1995, No 1, pp. 63–88.

Reet Kasiku doktoriväitekirjast "Verbid ja verbaalsubstantiivid tänapäeva eesti keeles. Tuletusprotsess ja tähendus". – Linguistica Uralica 1995, nr. 3, lk. 230–232.

Eesti keeles peitub maailmapilt. – Eesti Päevaleht, 1995, 2.oktoober.

Õim, H., Roosmaa, T. COPERNICUS'e keeletehnoloogiaprojektid. – Universitas Tartuensis, 1995, 17. veebruar.

1996

Koit, M., Roosmaa, T., Õim, H. Teaching computational linguistics: one vision. – Estonian in the Changing World. Ed. by H. Õim. Tartu University, 1996, pp. 115–122.

The need for a theory of folk theories in cognitive semantics: a review and a discussion. – Estonian in the Changing World. Ed. by H. Õim. Tartu University, 1996, pp. 193–210.

Naïve Theories and Communicative Competence: Reasoning in Communication. – Estonian in the Changing World. Ed. by H. Õim. Tartu University, 1996, pp. 211–231.

Teoreetiline keeleteadus ja integreeritud keeleteooria. – Keel ja Kirjandus 1996, nr. 11, lk. 731–744.

Sissejuhatavalt eesti keele akadeemilisest grammatikast. [Rets.: Eesti keele grammatika II. Süntaks. Lisa: kiri. Tallinn, 1993.] – Keel ja Kirjandus 1996, nr. 12, lk. 852–854.

1997

Linguistic semantics and naïve theories: a view of language competence. – Proceedings of the 16th International Congress of Linguistics. Paris, July 1997, pp. 97.

Eesti keele mentaalse maailmapildi allikaid ja piirjooni. – Pühendusteos Huno Rätsepale. Tartu Ülikooli eesti keele õppetooli toimetised 7. Toim. M. Erelt, M. Sedrik, E. Uuspõld. Tartu, 1997, lk. 255–268.

1998

Koit, M., Õim, H. Arvutuslingvistika mujal ja meil. – Keel ja Kirjandus 1998, nr. 1, lk. 1–7.

Koit, M., Õim, H. The Concept of Communicative Strategies: A Theoretical Model and an Implementation. – Proceedings of International Conference "Cognitive Strategies for Language Communication – CSC'98". Partenit, 1998, pp. 23–24.

Koit, M., Õim, H. Developing a Model of Dialogue Strategy. – Proceedings of the First Workshop "Text, Speech, Dialogue – TSD'98". Brno, 1998, pp. 387–390.

Koit, M., Õim, H. Uus eriala – arvutuslingvistika. – Universitas Tartuensis, 1998, 6. veebruar.

1999

Койт М., Ыйм Х. Диалог с компютером на естественном языке. – Труды по русской и славянской филологии. Лингвистика. Новая серия II. Прагматический аспект исследования языка. Тарту, 1999, с. 58–67.

Teoreettisen kielitieteen synty Virossa ja Suomessa: eroavuuksia ja yhtäläisyyksiä. – 75 vuotta viroa Helsingin yliopistossa. Viron kielen ja kultuurin opettaminen Suomessa – seminaari 23.11.1998. esitelmät. Castrenianumin toimitteita 56. Toim. R. Kasik ja Huima. Helsinki 1999, ss. 72–79.

Koit, M., Õim, H. Communicative strategies in human–computer interaction: a model that involves natural reasoning. – 23. Deutsche Jahrestagung für Künstliche Intelligenz, Bonn 1999.
Abstract 2 pp: http://www.ikp.uni-bonn.de/NDS99/Abstracts/8.ps
Paper 14 pp.: http://www.ikp.uni-bonn.de/NDS99/Finals/1_2.ps .

How to portray emotions. – Estonian: Typological Studies III. Publications of the Department of Estonian of the University of Tartu 11. Ed. by M. Erelt. Tartu, 1999, pp. 231–252.

Kaalep, H.-J., Õim, H. Eesti keeletehnoloogia sihtprogrammi arengust. – Infotehnoloogia haldusjuhtimises. Aastaraamat 1999. Tallinn, 1999, lk. 47–51.

Õim, H., Koit, M. Arvutuslingvistika kogub tuure. – Universitas Tartuensis, 1999, 12. veebruar.

Doktoriväitekiri keeletehnoloogiast. [Heiki-Jaan Kaalep. Eesti keele ressursside loomine ja kasutamine keeletehnoloogilises uurimistöös.] – Universitas Tartuensis, 1999, 16. aprill.

2000

Klaus, V., Mikone, E., Nurk, A., Oja, V., Parve, M., Päll, P., Õim, H. Congressus Nonus Internationalis Fenno-Ugristarum Tartu, 7.–13.8. 2000. – Keel ja Kirjandus 2000, nr. 10, lk. 760–767.

Kahusk, N., Orav, H., Paldre, L., Vider, K., Õim, H. Eesti keele tesaurus. – Arvutuslingvistikalt inimesele. Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 1. Toim. T. Hennoste. Tartu, 2000, lk. 127–152.

Koit, M., Õim, H. Konversatsiooniagendi modelleerimine. – Arvutuslingvistikalt inimesele. Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 1. Toim. T. Hennoste. Tartu, 2000, lk. 285–307.

Koit, M., Õim, H. Dialogue Management in the Agreement Negotiation Process: A Model that Involves Natural Reasoning. – Proceedings of the 1st SIGdial Workshop on Discourse and Dialogue. 2000, pp. 102–111.

Koit, M., Õim, H. Reasoning in Interaction: A Model of Dialoge. –TALN 2000 7e conference annuelle sur le Tratement Automatique des Langues Naturelles. Lausanne, 16–18 Octobre 2000. Ed. by M. Rajman, E. Wehrli. Lausanne, 2000, pp. 217–224.

Koit, M., Õim, H. Developing a model of natural dialogue. – From Spoken Dialogue to Full Natural Interactive Dialogue – Theory, Empirical Analysis and Evaluation. LREC2000 Workshop Proceedings. Ed. by L. Dybkjær. Ateena, 2000, pp. 18–21.

Tragel, I., Veismann, A., Õim, H. Kognitiivse keeleteaduse konverentsist ja teoreetilise keeleteaduse arengust selle taustal. – Keel ja Kirjandus 2000, nr. 4, lk. 260–168.

Otse, sirge ja *õige*: a Domain of Metaphoric Extension in Estonian. – Estonian: Typological Studies IV. Publications of the Department of Estonian of the University of Tartu 14. Ed. by M. Erelt. Tartu, 2000, pp. 198–220.

Eesti keele arvutitugi. - Humanitaarteadused ühiskonnale – ideoloogia ja tehnoloogia. Teadus ühiskonnale. Eesti Teaduste Akadeemia seminari materjalid. Tallinn, 2000, lk. 19–25.

Keeleuurimine ja keeleteooria läbi aegade. – Oma Keel 2000, nr. 1, lk. 7–17.

2001

Meister, E., Õim, H. Estonian Language Technology – Where Do We Stand? – Rahvusvaheline keelte arengu konverents "Eesti keel Euroopas", Tallinn, 12.–14.03.2001. Teesid, lk. 43–48.

Kahusk, N., Orav, H., Õim, H. Sensiting inflectionality: Estonian task for SENSEVAL 2. – SENSEVAL-2: Second International Workshop on Evaluating Word Sense Disambiguation Systems. Proceedings of the ACL-EACL 2001 Workshop SENSEVAL-2, Toulouse, France.

Keeletehnoloogiast ja eesti keelest. – Keel ja Kirjandus 2001, nr. 7, lk. 499–501.

Is there a folk theory of Self. The case of Estonian *ise* and *enese-enda*. – Papers in Estonian Cognitive Linguistics. Publications of the Department of General Linguistics 2. Ed. by Ilona Tragel. University of Tartu, 2001, pp. 7–21.

Straight in Estonian. – Fourth International Conference on Reasearching and Applying Metaphor (RAAM IV) Metaphor, Cognition and Culture (Tunis, Manouba, 5–7 April 2001). Conference Book. Tunis, Manouba: University of Manouba, 2001, pp. 39.

Keeletehnoloogia tagab keele säilimise infoühiskonnas. – Postimees, 2001, 12. september.

Koostanud Urve Talvik