TARTU ÜLIKOOL
Filosoofiateaduskond
Üldkeeleteaduse õppetool

Andres Loopmann

SÕNASTIKE HALDUSSÜSTEEM EELex

Magistritöö

Juhendaja: prof Haldur Õim

Tallinn 2007

Tartu Ülikool

Teaduskond Filosoofiateaduskond	**Õppetool** Üldkeeleteaduse õppetool	
Töö pealkiri Sõnastike haldussüsteem EELex		
Teadusvaldkond Üldkeeleteadus		
Taotletav kraad teadusmagister	**Kuu ja aasta** mai 2007	**Lehekülgede arv** 43

Referaat

Eesti Keele Instituudis on välja töötatud ja kasutusele võetud sõnastike haldussüsteem EELex. Antud töö kirjeldab sõnastikusüsteeme ning EELexi ülesehitust. EELex kui leksikograafi töövahend on ette nähtud sõnastike koostamiseks, toimetamiseks ja küljendamiseks. Sõnaraamatu artiklid vastavad etteantud struktuurile ning toimetamise käigus kontrollitakse artikli sisu vastavust struktuurile, toimetamisprotsessi käigus lubatavad tegevused on kontekstipõhised. Sõnastiku küljendatud kuju esitatakse MS Word tekstitöötlusprogrammis.

Võtmesõnad:

leksikoloogia, leksikograafia, sõnastike haldussüsteem, ükskeelne sõnaraamat, kakskeelne sõnaraamat, sõnaraamatu küljendamine, ÕS 2006, tekstikorpus, multimeediakorpus, AJAX, XML, XSD, XSLT, XPath, LibXML, LibXSLT, Perl, XMLHttpRequest, MSXML4, regulaaravaldis, GUID, skeem, valideerimine, globaliseerimine, lokaliseerimine

Säilitamise koht:	
Töö autor: Andres Loopmann	**allkiri**:
Kaitsmisele lubatud: **Juhendaja**: prof Haldur Õim	 **allkiri**:

Faculty	**Department**
Faculty of Philosophy	Dep. of General Linguistics

Title

The Dictionary Management System EELex

Field of Research

General linguistics

Degree	**Date**	**Number of Pages**
Master's Degree (MA)	May 2007	43

Abstract

The Dictionary Management System EELex, created at the Institute of the Estonian Language, is designed to provide tools for lexicographers to compile, edit and layout dictionaries. The independent components of the System are: (a) dictionary databases in XML format, and (b) software for dictionary management. Current document describes different dictionary management systems and EELex framework. Dictionary entries are checked against schema and all editing operations are context sensitive. Dictionary layout is presented in MS Word.

Keywords:

lexicology, lexicography, dictionary management systems, monolingual dictionary, bilingual dictionary, dictionary layout, Dictionary of Correct Usage 2006, text corpus, multimedia corpus, AJAX, XML, XSD, XSLT, XPath, LibXML, LibXSLT, Perl, XMLHttpRequest, MSXML4, regular expression, GUID, schema, validation, globalization, localization

Author: Andres Loopmann	**Signature**:
Supervisor: Haldur Õim, professor	**Signature**:

SISSEJUHATUS

Eesti Keele Instituudis on välja töötatud ja kasutusele võetud sõnastike haldussüsteem EELex. EELex kui leksikograafi töövahend on ette nähtud sõnastike koostamiseks, toimetamiseks ja küljendamiseks. Sõnaraamatu artiklid vastavad etteantud struktuurile ning toimetamise käigus kontrollitakse artikli sisu vastavust struktuurile, toimetamisprotsessi käigus lubatavad tegevused on kontekstipõhised. Sõnastiku küljendatud kuju esitatakse MS Word tekstitöötlusprogrammis.

Töö eesmärgiks on kirjeldada võimalikke sõnastikusüsteeme ja näidata väljatöötatud EELexi süsteemi tööpõhimõtet.

EELexi süsteemi tarkvara ja tööpõhimõtte autor on antud töö kirjutaja. Süsteemi funktsionaalsust on pidevalt täiendatud ja edasi arendatud koostöös Eesti Keele Instituudi (EKI) leksikoloogide ja sõnastike toimetajatega. Süsteemis kasutatavad sõnastikud – nii süsteemi imporditud kui ka süsteemiga loodud – on EKI omand. Sõnastike impordil süsteemi on kasutatud olemasolevaid EKI elektroonilisi andmefaile. Kogu töö sõnastike haldussüsteemiga EELex on seotud riikliku programmiga “Eesti keele keeletehnoloogiline tugi (2006–2010)”. Haldussüsteemi ettevalmistustöid rahastas osaliselt riiklik sihtprogramm “Keeletehnoloogia ja EKI sõnaraamatud”.

Töö esimeses peatükis vaadeldakse sõnastikusüsteeme kui selliseid üldiselt: milline funktsionaalsus peaks tänapäeval sõnastikusüsteemidel olema ja millised on mõned olemasolevad sõnastikusüsteemid. Teises peatükis kirjeldatakse EELexi süsteemi ennast: milline on tema ülesehitus, millised sõnastikud on EELexiga toodetud, millised koostamisel ning milline on EELexi tööpõhimõte.

Olen tänulik Margit Langemetsale ja Ülle Viksile, kes aitasid mind antud töö kirjutamisel kasulike märkuste ja ettepanekutega.

1. Sõnastikusüsteemid

Sõnastike haldussüsteemi (edaspidi: sõnastikusüsteem) kui leksikograafi töövahendi ülesandeks üldiselt on toetada sõnaraamatute koostamist, toimetamist ja küljendamist, kasutades tänapäevast IT-infrastruktuuri. Sõnastikusüsteem eeldab sõnastikuandmete ühtse standardi loomist, vanemate elektrooniliste sõnastike puhul ka andmete üleviimist uuele kujule. Lisaks leksikograafilistele põhifunktsioonidele peab tänapäevane sõnastikusüsteem võimaldama kontrollida sisestatavate andmete õigsust, sõnaartiklite vastavust struktuurile, pakkuma toimetamise käigus kontekstipõhiseid valikuid, olema kasutajasõbralik ning peaks ideaalis võimaldama hallata ka arvutigraafikat, heli ja väliseid andmeallikaid nagu teksti- ja multimeediakorpused ning internet. Sõnastikusüsteemide sissejuhatavas osas püütakse anda ülevaade sellest, milliste ideaalide poole sõnastikusüsteemid pürivad ning kuidas need ideaalid ühes või teises süsteemis kajastamist leidnud on. Leksikograafi ideaalide kirjeldamisel on põhilise allikana kasutatud G. M. de Schryveri artiklit „Lexicographers' Dreams in the Electronic-Dictionary Age", mis on ilmunud ajakirjas *International Journal of Lexicography* (De Schryver 2003).

1.1. Sõnastikusüsteemide omadused

1.1.1. Sõnastike tüübid

Laias laastus võib sõnastikud jagada üks- ja mitmekeelseteks. Eraldi võib ära märkida ka terminoloogiasõnastikud. Nende sõnastikutüüpide korral võivad olulisel määral erineda nii märksõnade hulk kui ka artikli struktuur. Olulise ükskeelse sõnaraamatuna eesti keele jaoks tuleb nimetada õigekeelsussõnaraamatuid, vastandina inglise-ameerika traditsioonile, kus oluliseks sõnaraamatuks keele jaoks on seletavat tüüpi sõnaraamatud (Erelt 2007: 5). Seletavat tüüpi sõnaraamatuid iseloomustab suur märksõnade hulk ning see, et kõik tavapärased infoüksused on esindatud (Langemets 2003). Seda tüüpi sõnaraamatud on kõige mahukamad.

Sõnaraamatu prototüübi valikul on oluline, milline märksõnade loend aluseks võtta: sellest sõltub sõnaraamatu maht. Sõnastikusüsteemi seisukohast on oluline, et sõnastike prototüübid oleksid omavahel ühilduvad: st nad kas on koostatud kasutusel oleva sõnastikusüsteemiga või on võimalik prototüüp sõnastikusüsteemi importida. Samuti peab mitmekeelsete sõnaraamatute korral olema korrektselt märgendatud infoüksuse keel. Selle abil on võimalik teostada korrektselt õigekirjakontrolli ning soovi korral esitada kellaaja, kuupäeva jm andmeid vajalikus vormingus. Sõnaraamatu mahukus – nagu allpool sõnastikuandmete esituses näeme – ei ole tänapäevaste arvutite võimaluste juures enam oluline. Oluline on siiski, et sõnaartikli struktuur oleks „mõistlikult" lihtne: keerulisema struktuuriga töötades tuleb sõnaraamatu toimetajaid kauem koolitada ning ka vead on tõenäolisemad. Teisest küljest annab üksikasjalikum struktuur rohkem võimalusi eri tüüpi päringute sooritamiseks.

1.1.2. Sõnastikuandmete esitus

XML on tänapäeval põhiline elektroonilise andmevahetuse standard. Sõnastikusüsteemide kirjeldustest võib tuua väga palju näited XML vormingu ühe või teise kuju kasutamisest, kuid leidub näiteid ka relatsiooniliste andmebaaside rakendamisest. Seetõttu kerkib küsimus,

kuidas on parem sõnastikuandmeid hoida: kas relatsioonilises andmebaasis või XML vormingus.

Relatsiooniliste andmebaaside eelisteks on andmete kirjepõhine (sõnaraamatute korral siis sõnaartiklid) juurdepääs ja andmeväljade indekseerimise võimalus.

XML vormingu eeliseks on eelkõige tema tekstipõhisus, mis hõlbustab dokumenteerimist ja teeb andmed inimesele loetavaks. Oluline on ka see, et andmeid saab XML vormingus esitada hierarhiliselt, st on lihtne kirjeldada sõnaartikli struktuuri. Nt võib artikli plokis sisalduda päise plokk, mille sees omakorda paiknevad märksõnagrupid jne. Sellise sõnaartikli struktuuri saab kirjeldada nn XML skeemiga ning skeemi on edaspidi võimalik kasutada artikli struktuuri kontrollimisel. Mida keerulisem on skeem, seda raskem on aga luua vastavust hierarhilise ja relatsioonilise mudeli vahel ning mingil tasemel osutub mõistlikuks loobuda kõikide andmeväljade järgi päringu tegemise võimalusest. Samas võib öelda, et praegu välja antavad paberkandjal sõnaraamatud nagu nt ÕS 2006 – ca 1200 lk ja 50 000 sõnaartiklit – on parasjagu nii suured, et tänapäevased arvutid tulevad nende haldamisega suurepäraselt toime, isegi kui sõnaraamat on esitatud XML vormingus. ÕS 2006 suuruseks XML vormingus on ca 20 MB, mis vastab 150–200 MB-le dokumendimudeli (DOM) esitusele arvuti töömälus. XML vormingu teiseks suureks eeliseks on XSLT teisenduste kasutamise võimalus, st andmeid saab lihtsalt esitada vajalikul visuaalsel kujul. Relatsioonilise mudeli eelised tulevad esile alles siis, kui andmebaasil on sadu ja rohkem samaaegseid kasutajaid. Seega võib öelda, et hierarhilise mudeli ja XML vormingu eelised on ilmsed ning selle vormingu kasutamine sõnastikuandmete hoidmiseks on õigustatud.

1.1.3. Multimeedia sõnastikes

Multimeedia andmeid nagu jooniseid, illustratsioone, fotosid, kaarte, graafikuid ja heli on sõnastikes kasutatud abistava materjalina juba pikka aega. Multimeediat kui illustreerivat materjali on kasutatud nt Oxford Advanced Learner's CD-ROM Dictionary (2000) elektroonilises versioonis keerulisemate tegusõnade tähenduse seletamiseks (De Schryver 2003). Aja jooksul on multimeedia kasutamine muutunud ka sisulisemaks: nt on võimalik

animatsiooni abil hieroglüüfiliste tähtede (hiina, jaapani keeled) kirjutamist õppida (Chinese Character Bible vms) või võrrelda enda salvestatud võõrkeelse sõna häälduse helilaine visuaalset kuju sõnastikus leiduvaga.

Oluline koht multimeedia andmetes on helil: võimalik on kuulata märksõnade ja tervete seletuste heliesitust, kasutades automaatset kõnesünteesi. Potentsiaalselt sobiks selline omadus nägemispuudega inimeste keeleõppeks vms. Siiski võib öelda, et heliesitus on jäänud pidama „sõna tasemele" ning hetkel ollakse kaugel nt lauserõhkude ja intonatsiooni väljendamisest (De Schryver 2003). Realiseerimata on ka näitelausete – kas siis salvestatud kujul või kõnesünteesi abil esitatud näitelaused – kuuldeline esitus sõnastikes. Vähemalt üks suur sõnaraamat – Elektronische Grote van Dale (2000) – on võimeline difoonipõhise kõnesünteesi abiga esitama sõnaraamatu sisu, kuid seda siiski ainult märksõnade piires. Seega: seni, kuni lauserõhku ja intonatsiooni pole võimalik adekvaatselt kõnesünteesi abil esitada, peavad näitealaused jm materjal olema sisseloetud. Siinkohal võib märkida, et Eesti Keele Instituudis on praegu loomisel eesti emotsionaalse kõne korpus eestikeelse tekst-kõne sünteesi tarbeks ning seega avarduvad ka eestikeelse sünteeskõne võimalused. Ja multimeedia osa lõpetuseks: heli kasutamine sõnaraamatutes peaks ideaalis jõudma selleni, et sõnastikule esitatud suulise kõne käsklustele reageerib sõnaraamat ka vastava informatsiooniga heli kujul (nt Crystal 1986).

1.1.4. Korpused

Korpuseid kasutatakse sõnastike tegemise juures nt järgmistel eesmärkidel: kasutaja võiks väljaantavate sõnastike juures otse saada kasutada puhast keelematerjali, leida korpuse tsitaate otse märksõna tähenduse juurest; kollokatsioonide otsingul peaks olema võimalik lehitseda korpuste näiteid. Olukord on praegu aga selline, et on väga raske leida sõnastikku, mis oleks ühendatud tekstikorpustega. Näitena võib tuua ainult Collins Cobuild on CD ROM 1995 ja 2001 (De Schryver 2003).

Samas ei pea korpused olema mitte ainult tekstikorpused. Multimeedia korpused võivad sisaldada nii arvutigraafikat kui ka kõnet. Näidetena võib tuua filmide ja/või uudiste

andmebaase, milles oleks võimalik teha otsinguid tegeliku sõnakasutuse sisu järgi (nt Hovy et al 1999). Samuti peaks olema võimalik sooritada otsinguid kõnekorpustes.

1.1.5. Sõnastikuga suhtlemine

Sisestamine klaviatuurilt, kopeerimine/kleepimine ja hiireklõpsud on tänapäeval leksikograafi töös kõige sagedasemad tegevused. On olemas aga juba tehnoloogiad, mis võimaldavad sisendina kasutada käsitsi kirjutatavat kirja puutetundliku ekraani ja optilise tuvastuse kaudu. Kasutajal peaks ka olema võimalik määrata, millisest muust andmebaasist ta lisainfot saada soovib, kas ükskeelsest sõnaraamatust või mitmekeelsetest jne. Päringuid peab olema võimalik teha kõigi sõnastikus sisalduvate infoväljade järgi: ei pea olema tingimata vajalik otsida märksõnu ja alles nende kaudu pääseda ligi ülejäänud informatsioonile. Loomulik on otsingul kasutada ka filtreid, loogilisi operaatoreid, kasutada otsitavate sõnade või järjendite loendeid jms.

1.2. Eksisteerivaid sõnastikusüsteeme

Tänapäeva vajadustele vastavalt on peaaegu kõik haldussüsteemid nn klient-server süsteemid. Klient-server arhitektuuri korral on võimalik andmeid keskselt hallata; klientidele saab esitada samu andmeid eri kujul, mida saab kasutada paljude rakendustega; andmete esituse (küljenduse) muutumisel saab muudatused korraga viia kõikide klientideni; vigade ilmnemisel saab muudatused tsentraalselt läbi viia, ilma et klienditarkvaras midagi muuta oleks vaja jne (nt Loopmann et al 2006). Eksisteerivate sõnastikusüsteemide kohta vt ka nt DWS 2006 materjale (De Schryver (ed by) 2006) ning Keeles ja Kirjanduses dets 2006 ilmunud ülevaadet (Langemets et al 2006b).

Küsimusele, milline on eelistatavam, kas universaalne (kommertsiaalne?) või omaloodud haldussüsteem, võib vastata, et. üldiselt vastab kõige paremini vajadustele ikkagi omaloodud haldussüsteem. Seda nii ülesannete erinevate püstituste kui ka spetsiifiliste vajaduste tõttu. Sisu, eesmärkide, funktsionaalsuse ja kasutajasõbralikkuse poolest on haldussüsteemid muidugi erinevad (Langemets et al 2006b).

Ilmselt tuntuima kommertsiaalse haldussüsteemina võib märkida TshwaneLex tarkvara sõnastike koostamiseks (TshwaneLex 2002–2006). Rakendusse on sisse ehitatud automaatne viidete jälgimise ja uuendamise süsteem, projektihalduse abivahendid, kasutaja poolt määratavad sorteerimistingimused ning mitmekeelsete sõnastike jaoks ka sõnastike pööramise funktsioonid. Versioonis 2.0 (ilmunud 3. juulil 2006a) on lisandunud meeskonnatöö võimalus (mitme kasutaja versioon keskse Oracle, Microsoft SQL või PostgreSQL andmebaasiga), XML vormingus andmete importimise võimalus (andmete eksport XML vormingusse oli võimalik juba versioonis 1.0), andmete eksport komadega eraldatud tekstifaili, üksuste peitmise võimalus (andmete maskid, mis võimaldavad „toota mitut sõnastikku ühest andmebaasist"). Täiustatud on ka sõnastike võrdlemise/liitmise funktsiooni: enne sõnastike liitmist on võimalik tehtavad muudatused üle vaadata. Haldussüsteemi puudusena tuleb märkida suletud firmapärast andmevormingut.

Iga haldussüsteemi väljatöötamine algab konkreetsest vajadusest. Nt Baskimaa ülikoolis oli selleks Kuuba koolisõnastik *Cuban Diccionario Básico Escolar* (Alegria et al 2006). Keskseks andmebaasiks valiti Berkeley DB XML, mis on vabavaraline XML-andmebaaside haldussüsteem ning võimaldab ka XPath ja XQuery juurdepääsu XML dokumentidele. Klienditarkvara seevastu käsitleb andmeid ainult XML vormingus: serveri tarkvara pakub kliendile artiklite XML kuju ning kliendi poolt muudetud artiklid saadetakse serverisse ja salvestatakse andmebaasi. Keskse andmebaasi kasutamine annab võimaluse indekseerida olulised andmed ning sooritada päringuid korraga mitmest sõnastikust kiiresti ja mugavalt. Klienditarkvara osas on tähelepanu pööratud kasutajasõbralikkusele, leksikograafilised põhitegevused on automatiseeritud. Artiklit saab parandada toimetamisalas ning samaaegselt on võimalik jälgida artikli küljendust (vaadet), mis vastavalt toimetamisalas tehtud muudatusele automaatselt muutub. Toimetamisala ja vaate üksused on omavahel ühendatud: elemendi klõpsamine ühes toob ta esile teises. Kogu süsteemi puuduseks võib lugeda seda, et klienditarkvara ei ole standardtarkvara nagu nt veebibrauser, seega klienditarkvara ei ole võimalik lihtsalt uuendada ning tülikas on jälgida kõikide klientide versiooniuuendusi.

Savoie ülikoolis väljatöötatud Jibiki platvormi (Mangeot 2006) keskmeks on relatsiooniline andmebaas: Postgres. Kogu haldussüsteem on välja töötatud vabavaralise tarkvara abil Java platvormil. Käesoleval ajal kasutab Jibiki platvormi 3 projekti:

- Papillon projekt. On ette nähtud mitmekeelsete andmebaaside jaoks ning hõlmab järgmisi keeli: hiina, inglise, prantsuse, saksa, jaapani, lao, malai, tai, vietnami. Projekti tulemused on vabalt kättesaadavad ja kasutatavad.
- GDEF projekt (Suur eesti-prantsuse sõnaraamat). Eesmärgiks kakskeelne 80 000 märksõnaga sõnaraamat. Käesoleval ajal (2006 sügis) on valmis ca mõni tuhat artiklit.
- LexALP projekt. Papillon projektil põhinev terminoloogiasõnastik, mis hõlmab avalik-õiguslikke haldustermineid neljas Alpi ala keeles: prantsuse, saksa, itaalia ja sloveeni keeles.

Klienditarkvarana on kasutusel veebibrauser. Toimetamisala genereeritakse artiklit kirjeldava XML skeemi alusel. Päringu funktsionaalsus on siiski piiratud: ei ole võimalik sooritada päringut artikli kõikide elementide järgi. Päringu vormistus ei ole ka kasutajasõbralik: puuduvad nt loendid otsitava elemendi sisestamiseks (sõnaliik jt) jne.

DEBII platvorm Masaryk'i ülikoolis (Pala, Horák 2006) on loodud eelkõige sõnastikuandmete haldamiseks. Platvormi loomise peamiseks motiiviks oli tšehhi keele sõnavara andmebaasi – Czech Lexical Database – loomine ning hilisem sõnaraamatu avaldamine. Samuti on silmas peetud WordNet taoliste andmebaaside lehitsemist ja parandamist, kuna semantilised andmevõrgud on keeleressursina omandanud suure populaarsuse. Nt EuroWordNet jaoks on välja töötatud lehitsemise ja parandamise programmid Polaris ja Periskope (kasutusel ka Eesti WordNetis). BalkaNet jaoks on välja töötatud VisDic tarkvara. BalkaNet projekti käigus (2001–2004) on välja töötatud 13 keele wordnet'id: inglise, hollandi, itaalia, hispaania, prantsuse, saksa, tšehhi, eesti, bulgaaria, kreeka, rumeenia, serbia ja türgi.

DEBII platvormil on välja arendatud järgmised rakendused:

- DEBDict: üldine sõnastike lehitseja, kasutusel on tšehhi kirjakeele sõnaraamat, võõrsõnade sõnaraamat, sünonüümide sõnaraamat, fraseologismide ning väljendite sõnaraamat ning Diderot entsüklopeedia; võimaldab ka ühendust luua morfoloogilise analüsaatoriga, väliste veebiserveritega (Google, Answers.com) ning geoinfosüsteemiga.
- DEBVisDic: WordNet editor.
- PRALED: tšehhi keele sõnavara andmebaasi tarkvara. Eesmärgiks on ca 100 000 artikliga kaasaegse Tšehhi keele andmebaasi loomine. PRALED on praegu loomisel tšehhi keele instituudi juures.
- DEB CPA: korpuse analüsaator (Corpus Pattern Analysis).
- DEB TEDI: tarkvara tšehhi terminoloogiasõnaraamatu loomiseks.

Keskseks andmebaasiks on valitud Berkeley DB XML-andmebaas ning serveri tarkvara koosneb paljudest moodulitest. Klienditarkvara loomiseks on kasutatud Mozilla arendusplatvormi (Mozilla Development Platform). Klienditarkvara graafilise kasutajaliidese peamine programmeerimiskeel on XUL (XML User-interface Language), mis võimaldab kasutada paljusid standardseid tehnoloogiaid nagu CSS, JavaScript, DOM, XSLT, XPath, DTD ning RDF. Klienditarkvara on tehniliselt realiseeritud kui Firefox brauseri laiendus (extension).

2. Sõnastikusüsteem EELex

2.1. Eellugu

Alates 1978. aastast on EKI-s loodud paarkümmend elektroonilist sõnastikku, mis oma teostuselt on üsna erinevad: tehnilised võimalused on ligi 30 aastaga palju muutunud (ülevaadet EKI elektroonilistest sõnastikest vt nt Loopmann et al 2006, varem on sellest kirjutanud Ülle Viks (1990)).

EKI sõnastikes on aja jooksul kasutatud erinevaid märgendusi, alates tüpograafilisest (elemente on võimalik omavahel eristada kirjastiilide järgi) ja kirjeldavast (deskriptiivsest) märgendusest – kus eri struktuuriüksusi tähistavad eri sümbolid ning iga uus üksus lõpetab automaatselt eelneva – kuni üldise, SGML märgenduskeeleni (Langemets 2000).

Esimene elektrooniline sõnastik EKI-s oli "Õigekeelsussõnaraamat" (ÕS 1976), mis tipiti arvutisse tabelina, kus sõnaartikli iga struktuurielemendi jaoks oli ette nähtud kindel positsioon reas, nt positsioonides 1-35 oli märksõna, positsioonides 36-37 muuttüübi number jne. Analoogiline süsteem pisut teisel kujul oli kasutusel paaris väiksemas sõnastikus, kus sõnaartikli elemente eraldas tühik ja elemendi tähendus oli määratud tema järjekorraga reas.

Mitmed suuremad sõnaraamatud (nt "Väike murdesõnastik" I-II (1982–1989), "Vene-eesti sõnaraamat" I–IV (1984–1994) jt) sisestati arvutisse polügraafilise märgendusega, selleks et neid fotolao vahendusel trükki anda: eraldi koodidega tähistati kirjastiilid, taanded, suurtähed jne.

80-ndate aastate teisel poolel võeti kasutusele sisuline deskriptiivne märgendus, kus iga struktuurielement oli varustatud prefiksilaadse tähtkoodiga, nt märksõna ees oli kood *m+*, tähenduse ees *t+*, grammatilise info ees *g+* jne. Suurem osa EKI sõnastikke ongi märgistatud nii, sõltuvalt sõnastikust võis koodide valik olla erinev. Selline märgendus võimaldab edukalt kirjeldada sõnaartikli struktuuri – juhul kui elemendid paiknevad lineaarselt: uue elemendi

algus on ühtlasi eelmise lõpp. Hierarhilist paigutust, kus üks element paikneb teise sees, niiviisi edasi anda ei saa. Deskriptiivse märgenduse suurim puudus on aga struktuurikontrolli puudumine. Sõnaartiklite struktuuri saab analüüsida alles tagantjärele, võrreldes artiklite koodijadasid omavahel ja otsides nende hulgast ebatüüpilisi (ehk vigaseid). Kuid just korrektne struktuur on see, mis tagab sõnastiku hilisema arvutitöötluse korrektsuse, nii keeletehnoloogilistes rakendustes kui ka sõnastiku enda edasiarendamisel.

Esimene katse siduda sõnastiku sisestamisega struktuurikontroll tehti EKI-s 90-ndate aastate alguses, kui oli valminud Eesti-vene sõnaraamatu (EVS 1997, 2000, 2004) 1. köite käsikiri. Koostöös TPI-ga loodi tarkvarasüsteem, mis sõnaartikli sisestamise igal sammul andis ette just selle valiku struktuurielemente, mis on antud kontekstis võimalikud. Nt pärast märksõna sisseviimist sai valida kas tähendusnumbri või seletuse või vaste, aga vaste järel võis tulla ainult vene grammatika jne. Süsteemi väljundiks oli tekstifail, kus iga struktuurielemendi ees oli numbriline kood (sisuliselt sama mis tähtkood plussiga). Artiklite struktuur oli küll kontrollitud ja vastas etteantud kirjeldusele, kuid see struktuur oli ikkagi lineaarne: element, mille sees juhtus olema mõni teine element, lõhuti tükkideks. Nt kui näite tõlke keskel sattus olema rektsiooniküsimus, siis tekkis kahe sisulise elemendi asemel kolme elemendi jada: "tõlge", "rektsioon", "tõlke jätk":

```
<näide>mängib tunnetega <tõlge>игр"ает <rektsioon>чьими <tõlke jätk>ч"увствами
```

EVS jäi selle süsteemi ainsaks rakenduseks – tema kohandamine teiste sõnastikega oleks nõudnud ulatuslikku programmeerimistööd. Selle süsteemiga on sisestatud EVS-i 3 esimest köidet.

2.2. EELexi ülesehitus

Käesolevaks ajaks (mai 2007) on EELex täielikult üle viidud nn AJAX („Asynchronous JavaScript And XML") tehnoloogia kasutamisele. AJAX all mõeldakse veebirakenduste loomise tehnoloogiat, mille eesmärgiks on veebilehed muuta interaktiivsemaks, kiiremaks ja võimalusterohkemaks. AJAX iseenesest ei ole uus tehnoloogia, vaid märgib terve grupi tehnoloogiate, eelkõige (X)HTML („HyperText Markup Language"), CSS („Cascaded Style Sheets"), DOM („Document Object Model") andmemudelite, XMLHttpRequest objektide ning JavaScript või VBScript programmeerimiskeelte ühiskasutust. AJAX kasutuse korral toimub infovahetus serveriga kasutajale märkamatult ning päringute korral ei ole vajadust kogu veebilehte uuendada (AJAX kohta vt nt WikipediA).

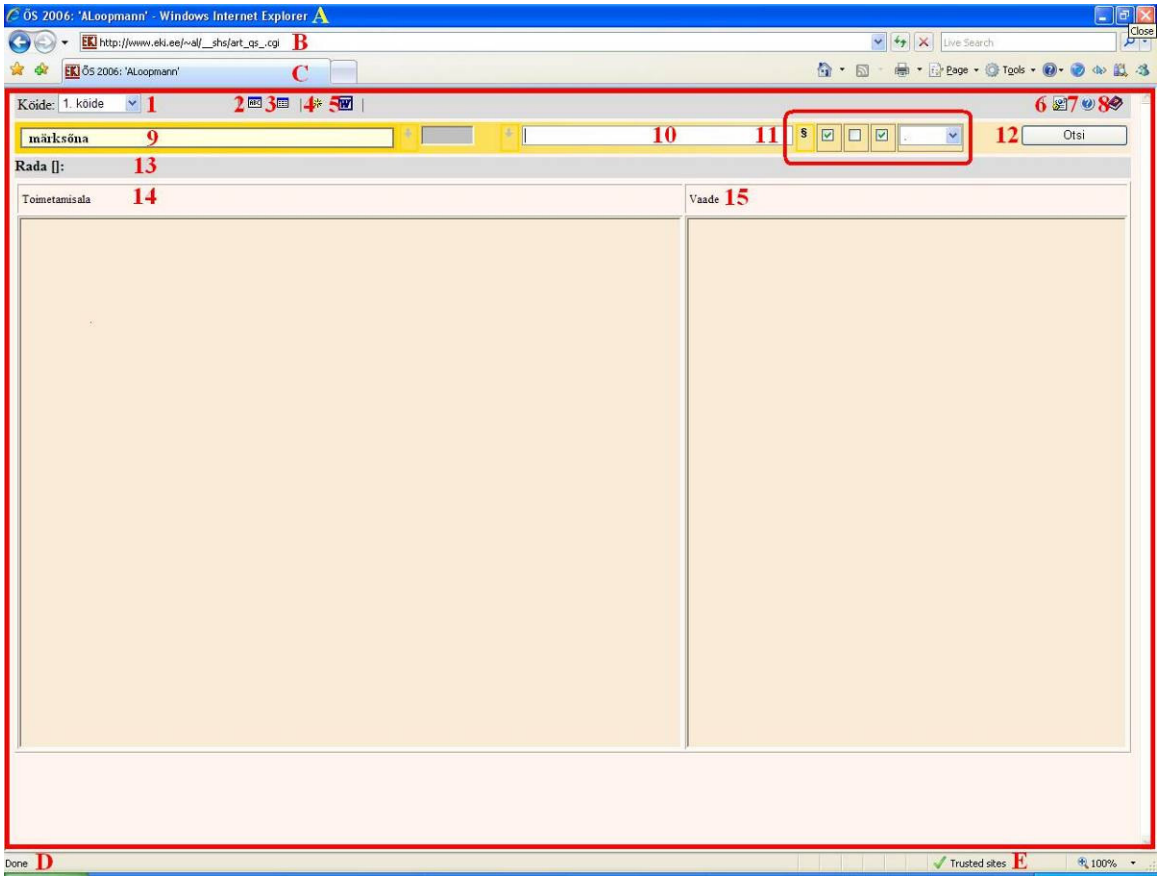
EELex on nn klient-server tüüpi rakendus. Sõnastikuandmed paiknevad tsentraalses veebiserveris XML vormingu kujul. Serveris paiknev veebiserveri tarkvara serveerib päringu tulemused tööjaamale veebilehel. Tööserveri konfiguratsioon on praegu järgmine: Pentium IV 2 GHz, 1,5 GB RAM, operatsioonisüsteemiks FreeBSD Unix v4.8, veebiserveriks Apache v1.3.33, XML parseriks XML::LibXML v1.58 ja XML::LibXSLT v1.57 Perl moodulid. Seega kõik serveri komponendid on vabavaralised. Perli versioon tööserveris on 5.8.7 ning kõik serveri protseduurid on tehtud Perlis. EELex süsteemi tööpõhimõtteid on tutvustatud ka EURALEX-i kongressil (Ü. Viks, M. Langemets, A. Loopmann – vt Langemets et al 2006a).

Sõnastikus koostatavate artiklite struktuur on määratud XSD skeemiga. XSD skeemi abil on võimalik kirjeldada XML faili struktuuri, määrata elemendi korduvust oma vanemelemendis, määrata elementide andmetüüpi (järjend, ajahetk, täisarv, elemendi väärtuste võimalikud loendid), defineerida uusi andmetüüpe jm.

Andmevahetus veebiserveri ja tööjaama vahel toimub samuti XML vormingus andmete abil. Tööjaama tarkvara valimisel on algusest peale lähtutud sellest, et rõhuv enamus inimesi kasutab arvutis MS Windows operatsioonisüsteemi. Nõudeks oli ja on, et tarkvara peab olema standardne ning et mingeid täiendavaid tarkvara installeerimisi poleks vaja teha. Seetõttu on tööjaama tarkvaraks valitud MS Windowsiga kaasas olev IE7 („Windows

Internet Explorer", praegu v7.0) veebilehitseja. Siinkohal võib ka märkida, et eri lehitsejates (nt Internet Explorer ja Mozilla Firefox) on DOM, CSS jt tehnoloogiate teostus ja W3C („World Wide Web Consortium") standarditele vastavus erinev ning seega ei ole eri lehitsejad omavahel täielikult ühilduvad. XML parseriks tööjaamas on MSXML4.

EELexi sisenetakse kasutajanime ja parooli abil, misjärel avaneb süsteemi avaleht (joonis 1). Joonisel 1 on kujutatud EELexi avaleht sisenemisel ÕS 2006 sõnaraamatusse. ÕS 2006-s on ca 50 000 artiklit ning füüsiliselt on tema maht XML-failina ca 20 MB. (A) kõrval on näha Windows programmiakna – antud juhul IE7 – nimi ja andmed: milline sõnaraamat, kasutajanimi ning märge, kas töötatakse testandmebaasis või mitte. (B) on IE7 aadressi lahter, (C) on antud internetiaadressi sakk (kaart). EELexi kasutusliidese ala on märgitud ära punase ristkülikuga, sellest väljaspool olevad elemendid kuuluvad IE7 kasutajaliidesele, seespool olevad elemendid sõnastikusüsteemile. (D) on IE7 staatusriba, sinna kirjutavad olulist informatsiooni nii IE7 kui ka sõnastikusüsteem. Kohas (E) näitab IE7, millisesse internetitsooni antud internetiaadress kuulub.



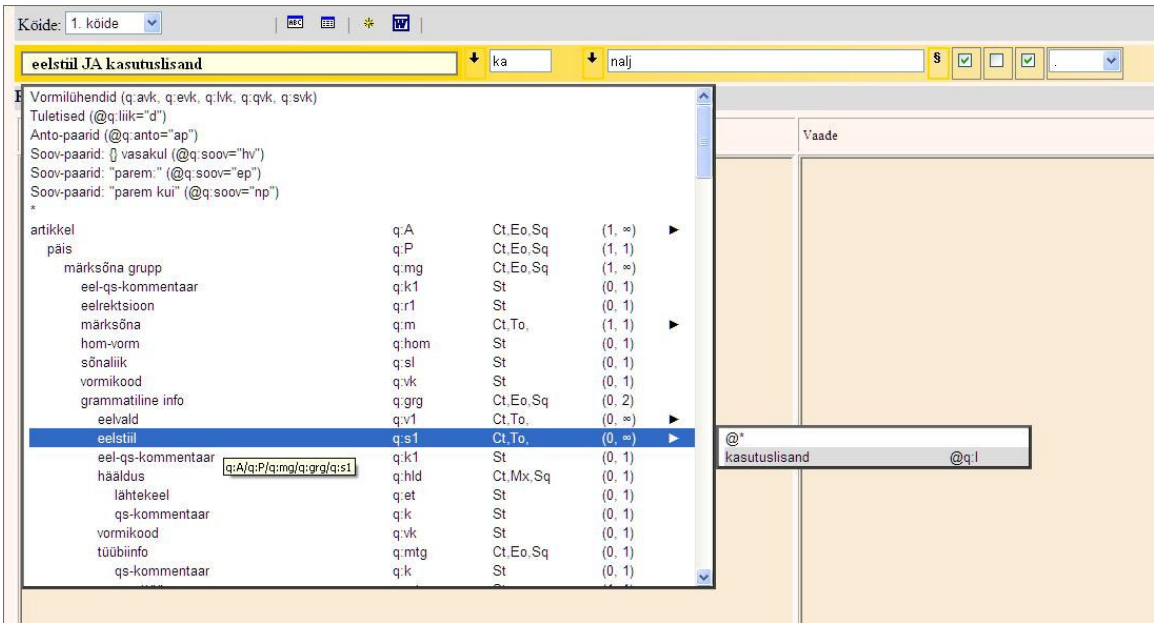
Joonis 1. Süsteemi avaleht: ÕS 2006

Tabelis 1 on toodud sõnastikusüsteemi kasutajaliidese nuppude otstarve:

Tabel 1. EELexi nuppude otstarve (vt joonis 1)

1	Sõnastiku köite valik
2	Artikli kuvamine
3	Otsingutulemuste loetelu kuvamine
4	Uue artikli lisamine
5	Sõnastiku küljenduskujul eksport MS Wordi
6	Sõnastiku skeemi (struktuuri) kuvamine
7	Lühispikrite kuvamine
8	Kasutusjuhendi kuvamine
9	Otsitava elemendi valik
10	Otsitav tekst antud elemendis
11	Päringu sooritamise parameetrid: a) laiendatud võimalustega päringuakna kuvamine, b) tõstutundlikkuse linnuke, c) sümbolitundlikkuse linnuke, d) globaalse-lokaalse otsingu linnuke, e) otsinguala valik
12	Otsingu käivitamine
13	Rada, millel kuvatakse artikli tekstiosa unikaalne identifikaator
14	Toimetamisala
15	Vaade

Nupud (1), (9), (10) ja (11) määravad otsingu viisi: millisest köitest, millise elemendi tekstist ja kuidas. Otsitav element määratakse nupu (9) abil: esile tuleb artikli struktuurile vastav menüü (vt joonis 2):



Joonis 2. Otsitava elemendi valik

Joonisel 2 on kõigepealt menüüs näha nn salvestatud päringud: (kõik) vormilühendid, (kõik) tuletised, (kõik) antonüümipaarid jt. Neile järgnevad artikli struktuurile vastavad elemendid. Elementide kohta käiv info on jaotatud 4 veergu: elemendi nimi, elemendi XML-silt, elemendi tüüp ning elemendi korduvus. Kui elemendil on ka tunnuseid (XML-atribuute), siis kuvatakse ka tunnuste menüü ning võimalik on otsingut sooritada ka nende järgi. Antud näites on otsitavaks stiilimärgend „ka nalj“: stiilimärgendi kasutuslisandi tunnus peab võrduma „ka“ ning stiilimärgendi tekst peab võrduma „nalj“. Selle otsingu tulemuseks on 6 artiklit, nt „eliksiir“, „kondipuru“ ja „rotisaba“. Kui tunnuse või elemendi väärtused on määratud loendiga, tekib tunnuse või elemendi kõrvale allapoole suunatud noolekujuline nupp, mille abil saab loendi kuvada ja vajadusel sellest ka valiku sooritada. Lühispikris (kollase taustaga kirjas) kuvatakse elemendi hierarhiline asukoht artikli struktuuris.

Peale elemendi nime – antud juhul „eelstiil“ – ja elemendi sildi – antud juhul „q:s1“ – näidatakse menüüs ära ka elemendi tüübi veerg, antud juhul „Ct,To“. „St“ (simpletype) tähistab elemendi lihttüüpi, mille sisuks saab olla ainult tekst; „Ct“ (complextype) komplekstüüpi. Komplekstüüpi element võib omada tunnuseid ning tema sisuks võivad peale teksti olla ka teised elemendid. Järgmine üksus tüübi veerus näitabki, milline sisu antud elemendil võib olla. „To“ (textonly) tüübi sisuks võib olla ainult tekst, „Eo“ (elementonly)

tüübi sisuks võivad olla ainult teised elemendid, „Mx“ (mixed) tüübi sisuks saab olla nii tekst kui ka teised elemendid. Viimane üksus tüübi veerus näitab, milline andmemudel kehtib elemendis sisalduvate teiste elementide jaoks. „Sq“ (sequence) tähistab kindlaksmääratud järjekorda, „All“: elemente maksimaalselt üks kord suvalises järjestuses, „Ch“ (choice): ühte elementi loetelust ning „Any“: suvalised elemendid suvalises järjestuses.

Elemendi kohta käiva info menüüs lõpetab korduvuse veerg: esimene arv sulgudes näitab, mitu korda minimaalselt element peab oma vanemelemendis sisalduma, teine arv: mitu korda maksimaalselt ta vanemelemendis võib sisalduda. Antud näite korral võib eelstiil grammatilises infos olla suvaline arv kordi: (0,∞).

Nuppude (11) abil (vt joonis 1) määratakse, kuidas otsing sooritatakse. Tõstutundlikkuse linnuke määrab, kas otsingul eristatakse suur- ja väiketähti või mitte. Sümbolitundlikkuse linnukese abil määratakse, kas otsingul arvestatakse elemendi tekstis sisalduda võivaid „mittetähti“ või mitte. Nt ÕS 2006 märksõnade tekstis on peale tähtede võimalikeks sümboliteks ka liitsõnaeraldajad (/), palatalisatsioonimärk (’) (apostroof), vältemärk (.) (punkt) jt märgid. Globaalse-lokaalse otsingu linnuke määrab, kas teksti otsitakse kogu artiklis sisalduda võivatest samanimelistest elementidest või ainult selle elemendi konkreetses asukohas. Antud näite korral tähistaks globaalne otsing, et stiilimärgendit „nalj“ otsitakse kõikidest artiklis sisalduvatest eelstiilidest q:s1; lokaalse otsingu korral otsitaks stiilimärgendit „nalj“ ainult päise märksõna grupi grammatilisest infost. Lõpuks: otsinguala valik määrab, millistest elemendi all olevatest üksustest teksti otsitakse. Võimalik on otsida kõikidest elemendi all olevatest tekstiüksustest eraldi (algväärtus), kõikidest tekstiüksustest kui ühest tervikust ning ainult elemendi külge kuuluvatest tekstidest (omab tähendust segaelementide – Mx-tüüpi – elementide korral).

Kui otsingutingimused on määratud, võib nupuga (12) otsingu käivitada. Otsingutingimustest moodustatakse kindla struktuuriga XML-fragment, mis edastatakse serverile. Tööjaama ja serveri vahelise XML-fragmendi transpordi eest hoolitseb XMLHttpRequest objekt. XML-fragment edastatakse serverile tavaliselt asünkroonselt, st peale info saatmist serverisse jätkab süsteem kohe oma tööd ning vastuse ja selle tõlgendamise eest hoolitseb

XMLHttpRequest objekt: nii kui vastus serverist on käes, informeeritakse sellest süsteemi. Vahepealse aja jooksul on süsteemil võimalus teha muid tegevusi, nt informeerida kasutajat visuaalselt protsessi käigust, vastata teistele kasutaja tegevustele vms.

Enne serverisse saatmist peab süsteem otsingutingimused „tõlkima“ XPath süntaksisse, et serveris töötav XML parser saaks otsingu täita. Antud näite korral on XPath ligikaudu selline:

```
q:A[.//q:s[translate(., concat('NALJ!#$%&()+,-./0123456789:;<=>?@[\]^_`
{|}~ ¡¢£¤¥¦§¨©ª«¬-®¯°±²³´µ¶·¸¹º»¼½¾¿`´^¨¸', "'", '"'), 'nalj') = 'nalj']
[@q:l = 'ka']]
```

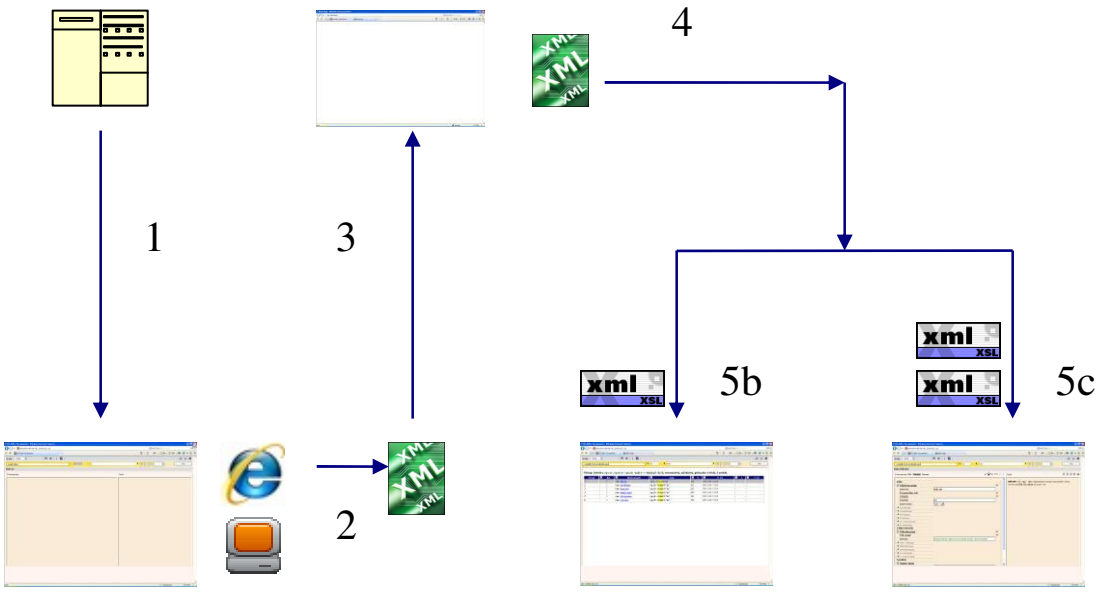
Kuna otsing oli antud juhul tõstutundetu ja sümboleid tekstis ei arvestatud, tuleb otsingul sõnastiku köite XML faili tekste teisendada nii, et suurtähed oleksid võrdsed väiketähtedega ning et sümbolid arvesse ei tuleks. Selle eest hoolitseb XPath funktsioon „translate()“, mille teine argument määrab teisenduse lähtetähed ja kolmas argument sihttähed ning kõik sümbolid teises argumendis, millele kolmandas vastet ei ole, jäetakse arvestamata. Funktsiooni esimene argument määrab otsinguala: kuna otsiti elemendi all olevast tekstist tervikuna, on väärtuseks „.“, mis on „self::node()“ sünonüüm. Kõikidest tekstidest eraldi otsingu korral oleks väärtus „.//text()“ ning ainult elemendi juurde kuuluvatest tekstidest otsimise korral „text()“. Kuna päring oli ka globaalne, algab artikli q:A kohta käiv XPath tingimuse rada kohe stiilimärgendi sildist „.//q:s“, lokaalse otsingu korral tuleb antud stiilimärgendi asukoht struktuuris täielikult välja kirjutada.

Otsingu sooritamisel edastatakse serverile XML-fragmendis käsu kood (antud näites otsing), sõnastiku köite number, päringu kohta käiv info (nt kuvamiseks logifailis, otsinguleidude võõpamiseks jm) ning otsingu XPath. Tööjaama suhtlemise serveriga võib jagada järgmisteks etappideks:

1) Tööjaamas avatakse süsteemi avaleht.
2) Tööjaamas moodustatakse käsu parameetrite järgi XML-fragment.
3) Tööjaam edastab serveris paiknevale protseduurilehele XMLHttpRequest objekti abil XML-fragmendi.

4) Protseduurileht täidab käsu ning tagastab tööjaamale vastuse XML-fragmendi, mille võtab vastu XMLHttpRequest objekt.
    a. Kui tegemist oli artikli salvestamise, kustutamise või lisamisega, tagastatakse info, kas käsk oli edukas.
    b. Kui tegemist oli otsinguga ja tulemuseks on rohkem kui üks artikkel, tagastatakse XML-fragmendis sõnastiku märksõnade loend.
    c. Kui tegemist oli otsinguga ja tulemuseks on ainult üks artikkel, tagastatakse XML-fragmendis ka kogu artikkel.
5) Vastuse XML-fragmendi analüüs tööjaamas.
    a. Salvestamise, kustutamise ja lisamise korral kuvatakse ainult operatsiooni info lehitseja staatusereal (joonis 1 (D)).
    b. Märksõnade loendi korral esitatakse loend XSLT teisenduse abil lehitsejas tabelina ning staatusereal kuvatakse info päringu kohta.
    c. Ühe artikli korral esitatakse artikkel toimetamisalas ühe XSLT teisendusega ja vaate alas teise XSLT teisendusega ning staatusereal kuvatakse info päringu kohta.

Skemaatiliselt näeb infovahetus tööjaama ja serveri vahel otsingu korral välja järgmine:



Joonis 3. Infovahetus tööjaama ja serveri vahel

Iga sõnaraamatu kõik tegevused säilitatakse antud sõnaraamatu logiraamatus. Logiraamatusse kantakse kuupäev, kellaaeg, kasutajanimi, arvuti võrguaadress, tegevuse kood, köite nr, tegevusele kulunud serveri aeg, tulemuse kood, tagastatud kirjete arv, artikli info ja päringu XPath süntaks. Näiteks:

```
2007-02-03 18:15:10 YViks "193.40.113.36" BrowseRead qs_1 5 Success
1 hälbima q:A[q:G='8d966963-95f6-41b1-a3dc-877ea8f96e91']

2007-02-03 18:14:52 YViks "193.40.113.36" ClientRead qs_1 7 Success
5 - q:A[q:P/q:mg/q:grg/q:k1]

2007-02-03 18:06:14 ALoopmann "62.65.218.82" ClientRead qs_1 9
Success 147 - q:A[q:P[.//q:k or .//q:k1 or .//q:k2]]
```

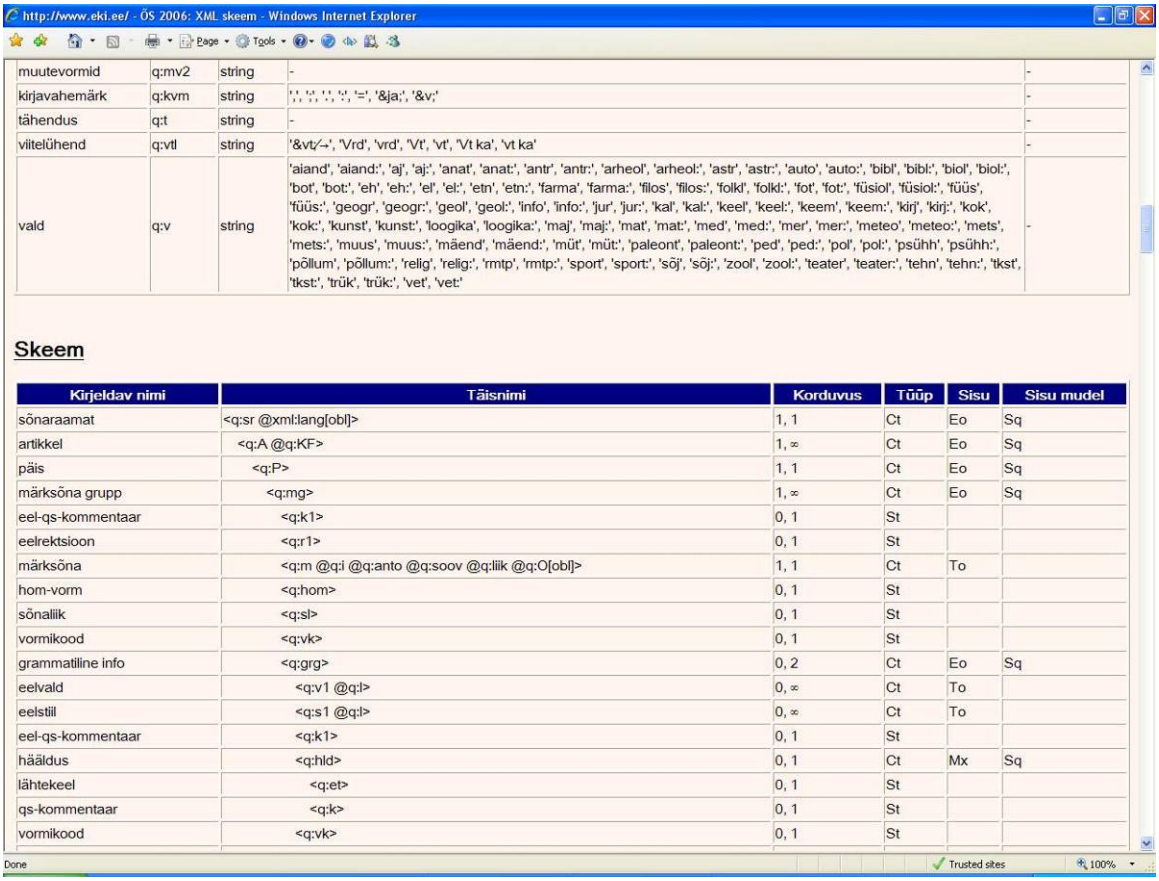
Iga tegevuse korral kuvatakse tööjaama lehitseja staatusereal sellele kulunud aeg. Stiilimärgendi „ka nalj“ otsingu näite korral kulus 6 artikli leidmiseks 9 sek.

2.3. Sõnastikud EELexis

EELexi sõnastikusüsteemi on praeguseks üle viidud Eesti-vene sõnaraamat (EVS), Õigekeelsussõnaraamat ÕS 2006, Kohanimevalimik (ÕS 2006 lisa), Eesti kirjakeele seletussõnaraamatu uute märksõnade ja uute tähenduste sõnastik (LEKS-baas) ning Silvi Vare Sõnapered. EELexi abil on lõpetatud, küljendatud ja ilmunud Eesti-vene sõnaraamatu 4. köide (2006) ning Õigekeelsussõnaraamat ÕS 2006 koos Kohanimevalimikuga. Koostamis- ja toimetamisjärgus on Eesti-vene sõnaraamatu 5. köide, S. Vare Sõnapered, LEKS-baas ning alustatud on eesti-läti ja eesti-leedu sõnaraamatute koostamist. Koostamisel on ühtne mitmekeelsete sõnaraamatute eesti keele poole prototüüp.

2.4. Artiklite struktuur ja järjestus sõnastikufailis

Lisaks otsitava elemendi valiku menüüle (vt joonis 1, nupp 9) on skeemi võimalik vaadelda eraldi aknas sõnastiku skeemi (struktuuri) kuvava nupu abil (vt joonis 1, nupp 6). Avanevas aknas kirjeldatakse sõnastikus kasutusel olevad nimeruumid, atribuudid, elemendid, artikli hierarhiline struktuur ning sõnastikus kasutusel olevad andmetüübid. Lõik ÕS 2006 skeemi esitusest näeb välja järgmine:



Joonis 4. Lõik ÕS 2006 skeemist

Artiklite struktuurikirjeldust on sõnastikes püütud ühtlustada. Kõikides sõnastikes jaotub artikkel nn suurteks üksusteks ja nende all olevateks elementideks ja gruppideks. Tabelis 2 on esitatud artiklite struktuuri põhijaotised (nn suurte üksuste ühised sildid on tähistatud suurtähtedega):

Tabel 2. Artiklite põhijaotised

Kirjeldav nimi	Ingliskeelne nimetus	Ühine silt
artikkel	entry	**A**
päis	head	**P**
sisu	body	**S**
tähendused	senses	**Z**
moodustusplokid	word-formation blocks	**B**
näited	examples	**N**
fraseoloogia	phraseology	**F**
viited	cross-references	**VT**
kommentaarid	comments	**KOM**
GUID	GUID	**G**
artikli koostaja	created by	**K**
koostamise algus	date of creation	**KA**
koostamise lõpp	end of creation	**KL**
artikli toimetaja	modified by	**T**
toimetamise algus	date of modification	**TA**
toimetamise lõpp	end of modification	**TL**

Lisaks sõnastikuelementidele – nagu päis, sisu, fraseoloogia jt – kasutatakse EELexis iga sõnastiku kõikides artiklites ka logielemente, mille abil on võimalik sõnastiku tööprotsessi jälgida. Nendeks on artikli koostamise aeg, koostaja nimi, toimetaja nimi ja viimase toimetamise aeg. Samuti on igal artiklil oma unikaalne identifikaator: GUID. GUID on 16-baidine unikaalne täisarv, tema esitus järjendina kuueteistkümnendsüsteemi arvudes võtab enda alla 32 märki. Nt suurtähega algavate märksõnade otsingust pärit artikli märksõnaga „Achilleuse kand“ GUID on

DDD11553-69F6-493D-A807-CB4A1D890450.

GUIDi kasutatakse süsteemis nt artikli leidmisel otsingutulemuste nimekirjast, artikli salvestamisel ja kustutamisel.

Igas sõnastikus on artiklite järjestamise reeglid erinevad. Nt mõnes võidakse mitmesõnalises märksõnas tühikut järjestamisel arvestada, teistes mitte. Lisaks võivad märksõnas kasutusel olevad sümbolid olla erinevad. Nt ÕS 2006-s on liitsõnaeraldajaks (/), LEKS-baasis (|), teistes (+). ÕS 2006-s lisatakse märksõna teksti ka vältemärk (.), palatalisatsiooni märk (’) jt. Polüseemsete märksõnade teksti eristamiseks võidakse kasutada allkriipsu (_), lisaks peavad

väiksema homonüüminumbriga märksõnad olema sõnastikus eespool suurema homonüüminumbriga märksõnadest.

Sõnastiku XML fail on kogu aeg füüsiliselt õiges artiklite järjestuses. Iga uue artikli lisamisel või märksõnade teksti muutmisel järjestatakse sõnastik uuesti. Järjestamisel kasutatakse serveri XSLT teisenduse „`xsl:sort`“ funktsiooni.

2.5. Otsingud

Artiklite otsimiseks on võimalik kasutada nn tavaotsingut, otsingut XPath süntaksi järgi ning otsingut regulaaravaldiste süntaksiga. Praegu eristatakse neid otsingu liike otsitava teksti lahtris (vt joonis 1, nr 10) oleva teksti esisümbolite järgi: kui tekst algab kahe paragrahvi sümboliga (§§), siis kasutatakse regulaaravaldiste süntaksit; kui tekst algab ühe paragrahvi sümboliga (§), siis kasutatakse XPath süntaksit; ülejäänud juhtudel tavaotsingut. Lähimaks eesmärgiks on need otsingu liigid koondada ühte, nn laiendatud võimalustega otsinguekraani, milles oleksid ühendatud kõik variandid ning kus XPath süntaksi tundmine ei oleks vajalik.

2.5.1. Tavaotsing

Tavaotsingul kasutatakse tekstijärjendeid ja metasümboleid. Metasümboliteks tavaotsingus on tärn (*) ja allkriips (_). Tärn (*) tähistab suvalist järjendit, allkriips (_) üksikut tähte. Tavaotsingu näiteid on esitatud tabelis 3, liitsõnapiiri tähistab püstkriips (|).

Tabel 3. Valik tavaotsingu näiteid (LEKS-baasist)

Otsinguväli	**Otsitav tekst**	**Kommentaar**	**Vastus**
märksõna	**abikäsi**	TÄPNE OTSING: leitakse kõik märksõnad "abikäsi"	abi|käsi
näiterühm: stiil	**kõnek**	TÄPNE OTSING: leitakse kõik näiterühmades esinevad stiilimärgendid "kõnek"	breik .. *Breiki tegema, panema.* kõnek.
märksõna	**abi***	SÕNA ALGUSE OTSING: leitakse kõik märksõnad, mis algavad järjendiga "abi"	abi|aine, abielu|akt, abielulahutus|akt ...
märksõna	***abi**	SÕNA LÕPU OTSING: leitakse kõik märksõnad, mis lõpevad järjendiga "abi"	humanitaar|abi
märksõna	***ndus***	OTSING SÕNA SEEST: leitakse kõik märksõnad, mis sisaldavad järjendit	agraar|-industriaalne, ajakirjandus|vabadus, asendus|teenistus,

		"ndus"	auto|ärandus ...
märksõna	**a*v**	SÕNA ALGUSE JA LÕPU OTSING: leitakse kõik märksõnad, mis algavad a-ga ja lõpevad v-ga	anne|päev
sagedus	**(*lahter jäetakse tühjaks*)**	NB! Kuna salvestamisel kustutatakse alati tühjad väljad, siis tavaliselt tulemus puudub	
liitsõnade plokk	**=NULL**	PUUDUVA VÄLJA OTSING: leitakse artiklid, kus valitud otsinguväli puudub	aaderdama, aaderdus, aasta|intress ...
märksõna	*****	KÕIKIDE SÕNADE OTSING: leitakse kõik märksõnad (kui ei ületata otsingutulemuste limiiti)	(*sõnade loend*)
märksõna	***|***	leitakse märksõnad, milles sisaldub liitsõnapiir '|'	aasta|intress …
märksõna	***|*|***	leitakse märksõnad, milles on kaks liitsõnapiiri '|'	halogeen|(hõõg)|lamp, kõne|post|kast
märksõna	**b__t**	leitakse neljatähelised märksõnad, mis algavad tähega 'b' ja lõpevad tähega 't'	bait, balt
märksõna	**__**	leitakse ainult kahetähelised märksõnad	ca, CD, CV …
märksõna	**__b__**	leitakse viietähelised märksõnad, mille kolmas täht on 'b'	album, kebab

Metasümbolite kasutamisel kasutatakse XPath süntaksi saamiseks XPath funktsioone „contains()“, “string-length()”, „substring-after()“ jt.

2.5.2. Otsing regulaaravaldistega

Regulaaravaldiste kasutamine tekstiotsinguis võimaldab määrata üksikute tähtede või tervete järjendite korduvust; märkida otsingus täheklasse nagu kirjatäht/mittekirjatäht, suurtäht/väiketäht; otsida sümboleid loetelust; märkida sõnapiiri, teksti algust, lõppu jne. Kuigi W3C XPath 1.0 standardi järgi ei ole regulaaravaldiste kasutamine XPath avaldistes

otseselt võimalik, on võimalik kasutada XSLT parserites olevaid võimalusi skriptide või funktsioonide kasutamiseks. EELexis kasutatakse XML::LibXSLT Perl mooduli „`register_function()`“ protseduuri, mille abil saab XSLT teisendusefailides igas XPath päringus kasutada Perl funktsioone. Nt otsides ÕS 2006-st suurtähega algavaid märksõnu, oleks serverisse saadetav XPath ligikaudu selline:

```
q:A[.//q:m[al_p:rex(self::node(), '(^\p{Lu})') > 0]]
```

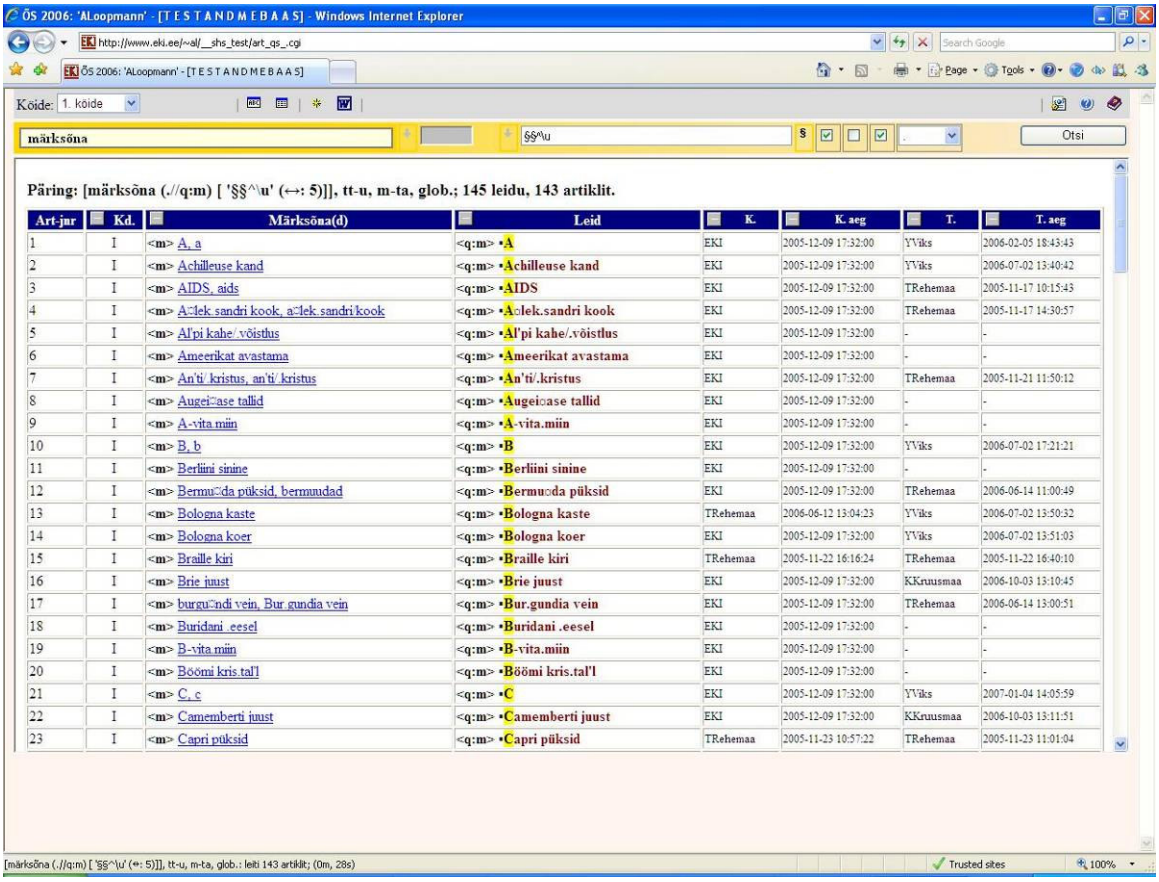
Funktsioon „rex“ (eesliitega „al_p“) on siin serveri protseduurilehe Perl funktsioon, mitte XPath funktsioon.

Kasutusel on kõik standardsed regulaaravaldiste märgid ja (Unicode) täheklassid, lisaks ka süsteemis defineeritud täheklassid, nt täis- ja kaashäälikud. Mõned näited:

- ^ - teksti algus; $ - teksti lõpp
- \u – suurtäht
- \l – väiketäht
- \v – vokaal, täishäälik
- \k – konsonant, kaashäälik
- \b – sõnapiir
- [xyz] – vähemalt üks märk loetelust
- [A-Z] – vahemik A kuni Z
- \p{L} – suvaline Unicode täht suvalises keeles/kultuuris
- \p{Lu} – suvaline Unicode suurtäht suvalises keeles/kultuuris
- \p{N} – suvaline numbrimärk suvalises keeles/kultuuris
- …

2.5.3. Otsingutulemuste esitamine

Kui otsingu tingimustele vastavaid artikleid on rohkem kui üks, kuvatakse otsinguleiud tabelina. Näiteks suurtähega algavate märksõnade otsingu korral ÕS 2006-st (vt joonis 5):

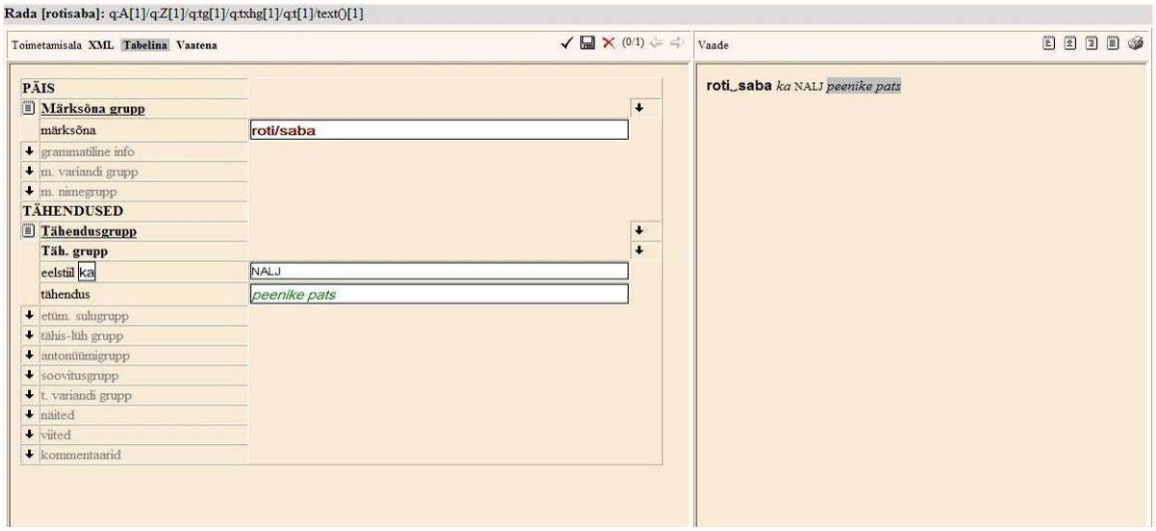


Joonis 5. Otsingutingimustele vastavate leidude tabel (ÕS 2006)

Leidude tabelis kuvatakse artikli järjenumber, köite number, märksõna elemendi silt ja märksõna, leiuelemendi silt ja leid, artikli koostaja nimi, koostamise aeg, artikli toimetaja nimi, viimase toimetamise aeg. Leidude veerus on otsitav järjend kollase taustaga võõbatud. Iga veerg tabelis on sorteeritav, lisaks on võimalik kuvada igast veerust kas ainult erinevad väärtused või kõik väärtused. Märksõna tekst märksõnade veerus on sinise värviga ning temal klõpsates avaneb vastav artikkel. Tabelist artikli avamisel otsitakse artikkel üles tema unikaalse identifikaatori – GUID – kaudu. Otsingutulemuste tabeli ja artikli vahel on võimalik liikuda nuppude (2) ja (3) abil (vt joonis 1 ja tabel 1).

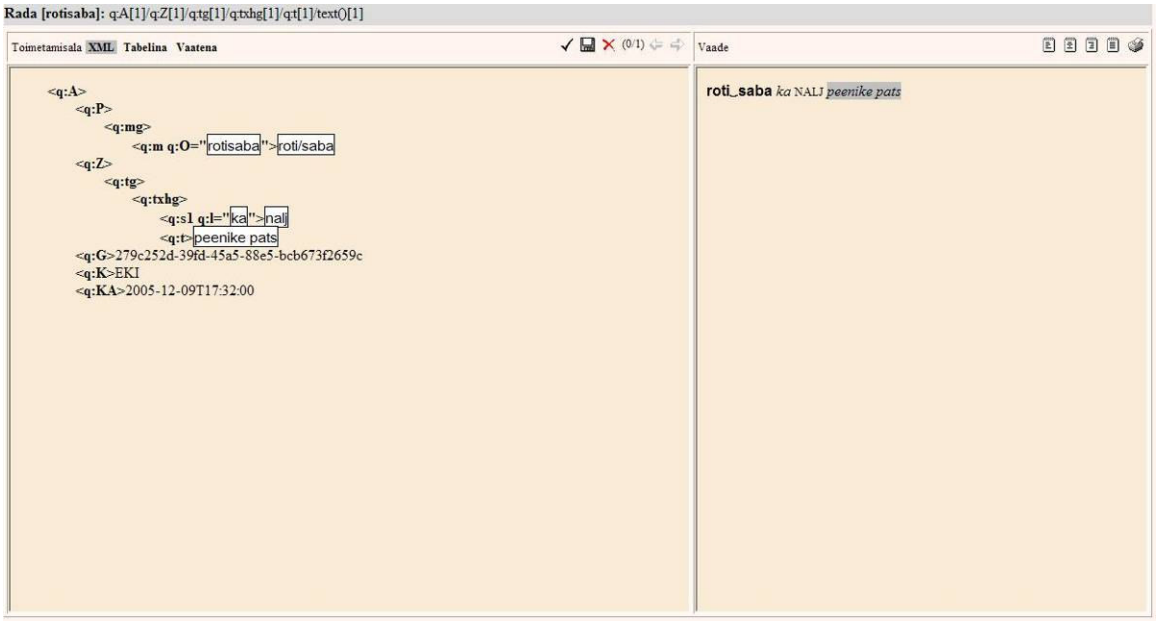
2.6. Artiklite toimetamine

Artikli avamisel avaneb vasakul poolel toimetamisala, paremal artikli küljendusvaade. Näitena toome siin esimesest otsingust – eelstiil võrdub „ka nalj“ – pärit artikli „rotisaba“ (vt joonis 6).

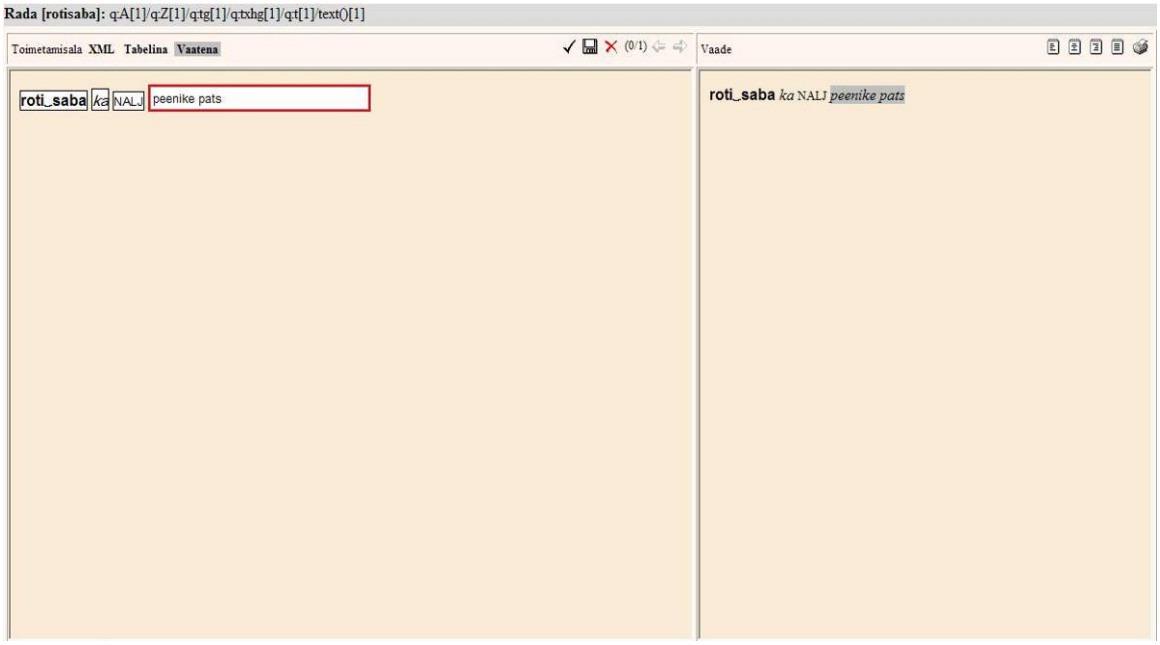


Joonis 6. Sõnaartikli ekraanipilt (ÕS 2006: "rotisaba")

Artiklit saab parandada toimetamisalas. Toimetamisalas saab ka kasutada erinevaid vaateid, algväärtuseks on vaade tabelina. Artiklit on võimalik parandada ka XML-vaates (joonis 7) ning küljendusvaates (joonis 8).



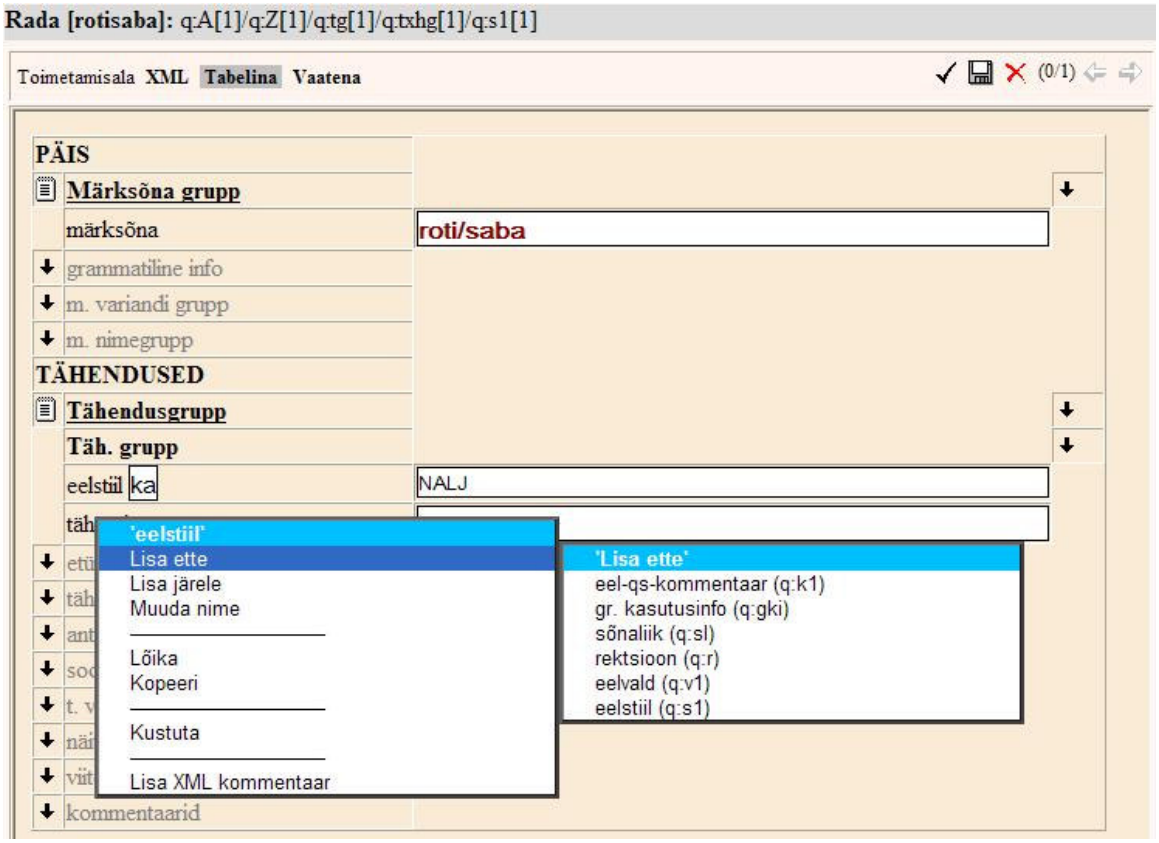
Joonis 7. Toimetamisala XML-vaade (ÕS 2006: "rotisaba")



Joonis 8. Toimetamisala küljendusvaade (ÕS 2006: "rotisaba")

Igale artikli üksusele (XML elemendile) on võimalik esitada tema kontekstmenüü. Kontekstmenüü tekib, kui klõpsata mingit elementi parema hiireklahviga. Nagu eespool märgitud, kasutatakse tööjaamas MSXML4 parserit XML-andmete haldamiseks. MSXML parseris on realiseeritud ka skeemi mudel SOM („Schema Object Model“). Kogu

kontekstipõhine info – millised elemendid võivad antud elemendi ees, sees või taga asetseda, kas antud elemendil on olemas tunnused jms – võetakse SOM-ist. Tänu sellele on võimalik kontekstmenüüde kaudu esitada ainult need tegevused, mis antud artikli osas on skeemi järgi lubatud. Nt artiklis „rotisaba“ on kontekstmenüü sisu eelstiili valimise korral järgmine (vt joonis 9):

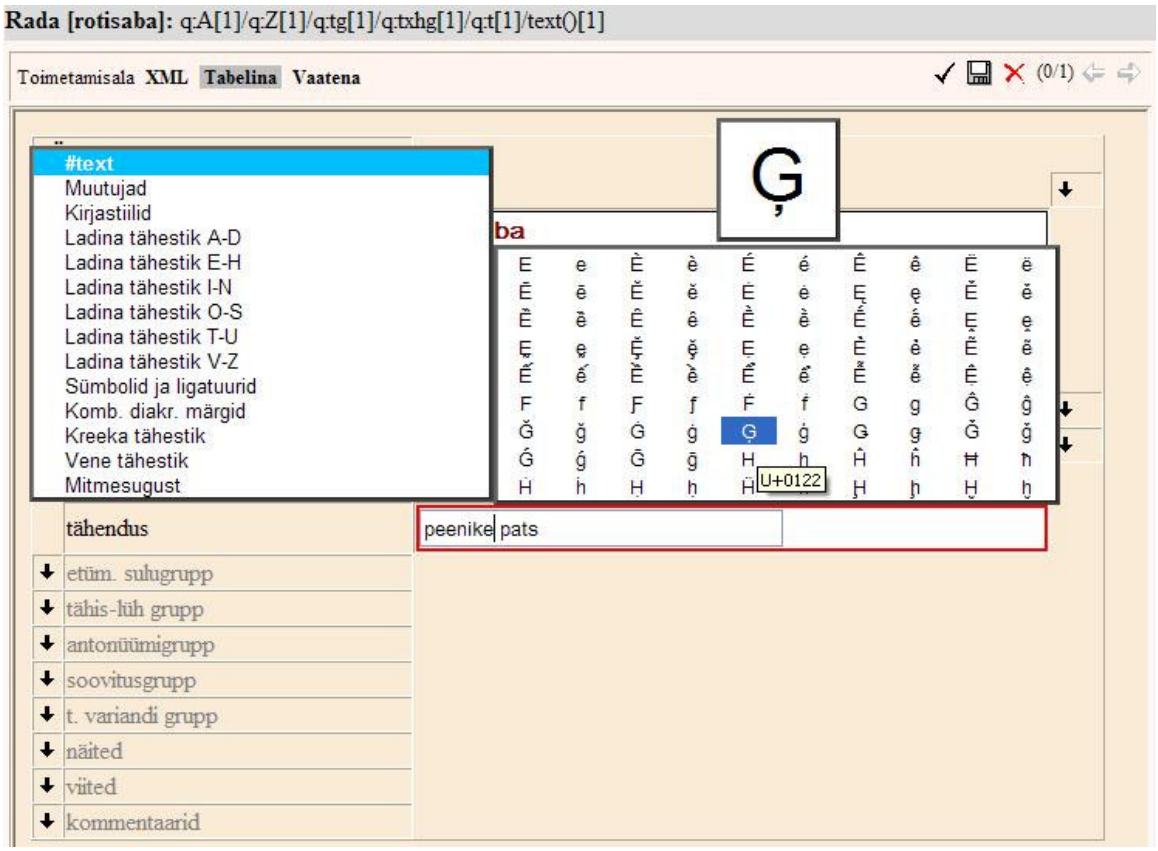


Joonis 9. Valitud elemendi kontekstmenüü (ÕS 2006)

Eelstiili ette on võimalik lisada ainult kindlaid elemente, nagu kontekstmenüüst näha. Lisaks on kontekstmenüü kaudu võimalik lisada elemente antud elemendi järele; elemendi sisse – kui sisu on veel tühi – ; lisada elemendile tunnuseid, kui elemendil on need skeemi järgi olemas; muuta elemendi silti (ja nime), kui skeemi järgi on see lubatud; kopeerida („Copy“), lõigata („Cut“) või kleepida („Paste“) elementi või tervet gruppi; kustutada element, kui see on lubatud ning lisada XML-kommentaare.

Eraldi kontekstmenüü on olemas ka tekstilahtrite sees: ka seal on võimalikud ainult kindlad, kontekstile vastavad tegevused. Kui tegemist on segatüüpi (Mx-tüüpi) elemendiga, kuvatakse

selles kontekstmenüüs samuti „Lisa ette“, „Lisa järele“ jt menüüpunktid, tavalise tekstivälja korral kuvatakse ainult sümbolite ja kirjastiilide menüüpunktid. Joonisel 10 on kujutatud kontekstmenüü tähenduse sees paremklõpsu tehes:



Joonis 10. Tekstilahtri kontekstmenüü (ÕS 2006)

Ladina tähestiku kontekstmenüüsse on lisatud kõik laiendatud ladina tähestiku tähed, kombineerivate diakriitiliste märkide abil on võimalik märkida rõhkusid, võimalik on lisada mitmesuguseid matemaatilisi jm sümboleid. Kontekstmenüüs liikudes kuvatakse kontekstmenüü kohal ka lisatava tähe suurem kujutis.

Tekstilahtri sisu muutmisel kontrollitakse sisu vastavust andmetüübile ja skeemi seatud piirangutele, soovi korral on võimalik kasutada tööjaama MS Word õigekirjakontrolli.

SOM infot kasutades on võimalik luua ka protseduurid tervete gruppide lisamiseks. Gruppide ja elementide lisamiseks on tabeli vaates olemas eraldi protseduurinupud. Elementide nimest vasakul olevad valge taustaga ikoonid tähistavad uue sarnase grupi lisamist, mustad allapoole

suunatud noolekujulised ikoonid uue grupi loomist. Tabelivaate paremas ääres olevad noolekujulised ikoonid lubavad ühekorraga lisada kõik antud grupis veel mitte olemas olevad elemendid.

Toimetamisala tabelivaade esitatakse tööjaamas XSLT-teisendusena. Vastavalt artikli seisule peab teisendus arvesse võtma olemasolevad ja puuduvad grupid; arvesse võtma igas grupis, millised elemendid on olemas, millised mitte; arvestama, kas elemendi tunnused on täidetud; esitama elementide ja gruppide nimed vastavas kujunduses ning sellele vastavalt esitama artikli. Kuna „käsitsi“ oleks sellise XSLT-teisenduse kirjutamine äärmiselt töömahukas, on tabelivaate kujundamine automatiseeritud ning see genereeritakse automaatselt vastavalt sõnaraamatu skeemile.

Toimetamisala ja küljenduse vaated on omavahel seotud: klõpsates ühel neist mingile elemendile, avatakse või näidatakse taustaga võõbatult sama element ka teises.

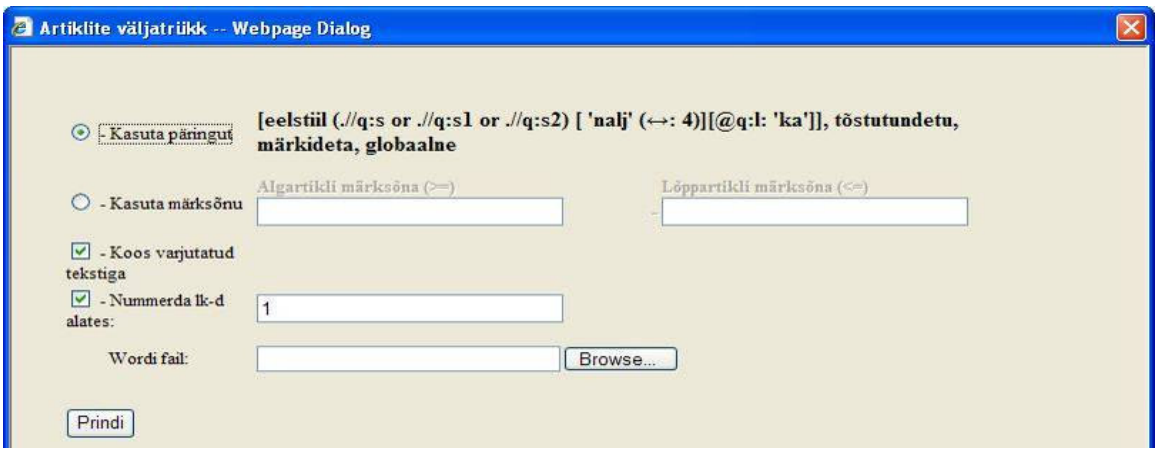
Parandamisel on võimalik kasutada tühista/taasta („Undo/Redo“) käske, vastavad noolekujulised nupud asuvad toimetamisala üleval paremas ääres. Numbrid noolenuppudest vasakul tähistavad paranduste jooksvat puhvrit ja puhvrite koguarvu.

2.7. Artiklite salvestamine

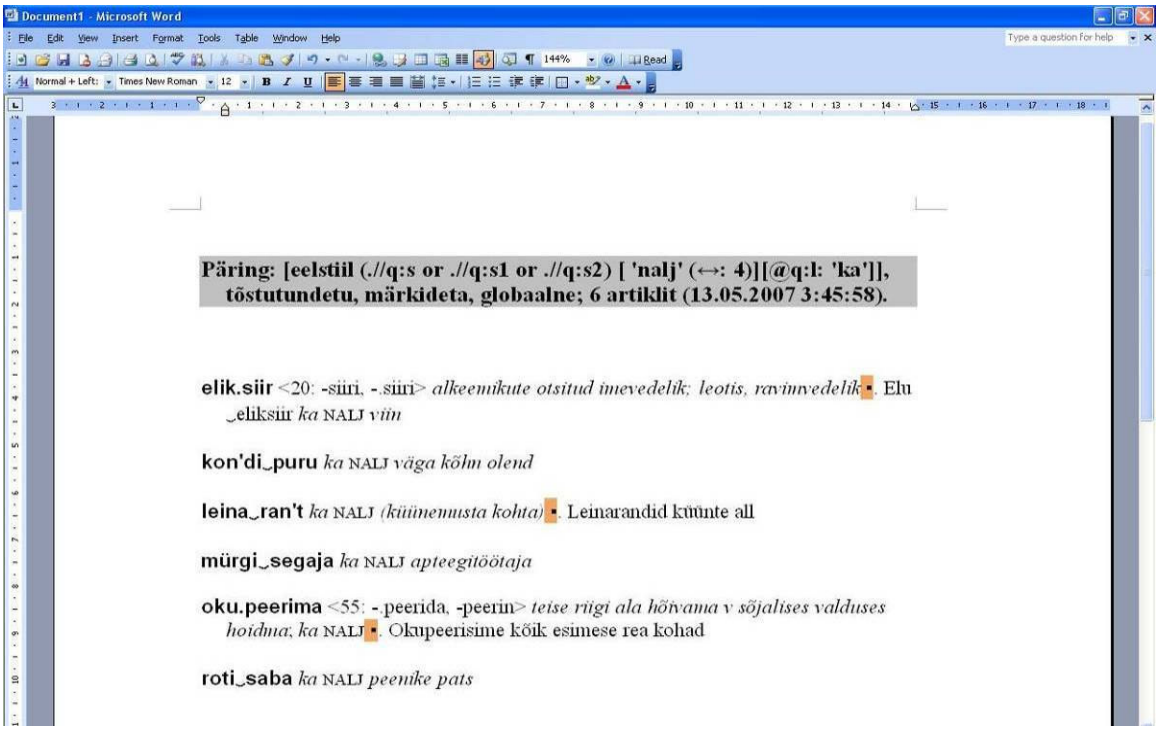
Artikli salvestamisel kustutatakse kõik tühjad tekstiväljad ja grupid ning kontrollitakse artikli vastavust skeemile. Kui artikkel skeemile ei vasta, informeeritakse kasutajat veateatega ning salvestamise protsess katkeb. Kui artikli struktuuri ja sisuga on kõik korras, formeeritakse XML-fragment salvestuse käsuga, lisatakse artikli sisu fragmendile ning fragment saadetakse serverisse. Salvestamine toimub toimetamisala ülal paremas ääres asuva flopikujulise nupu abil. Artikli sisu on võimalik kontrollida ka ilma salvestamata igal hetkel, vastav linnukesekujuline nupp asub kohe salvestamisnupu kõrval.

2.8. Sõnastiku küljendamine

Sõnastiku küljendatud kuju esitatakse tööjaamas MS Wordis. Küljenduse väljatrüki ekraan, kasutades ÕS 2006 eelstiili „ka nalj“ otsingut ning väljatrükk Wordis on esitatud joonistel 11 ja 12:



Joonis 11. MS Word väljatrüki ekraan

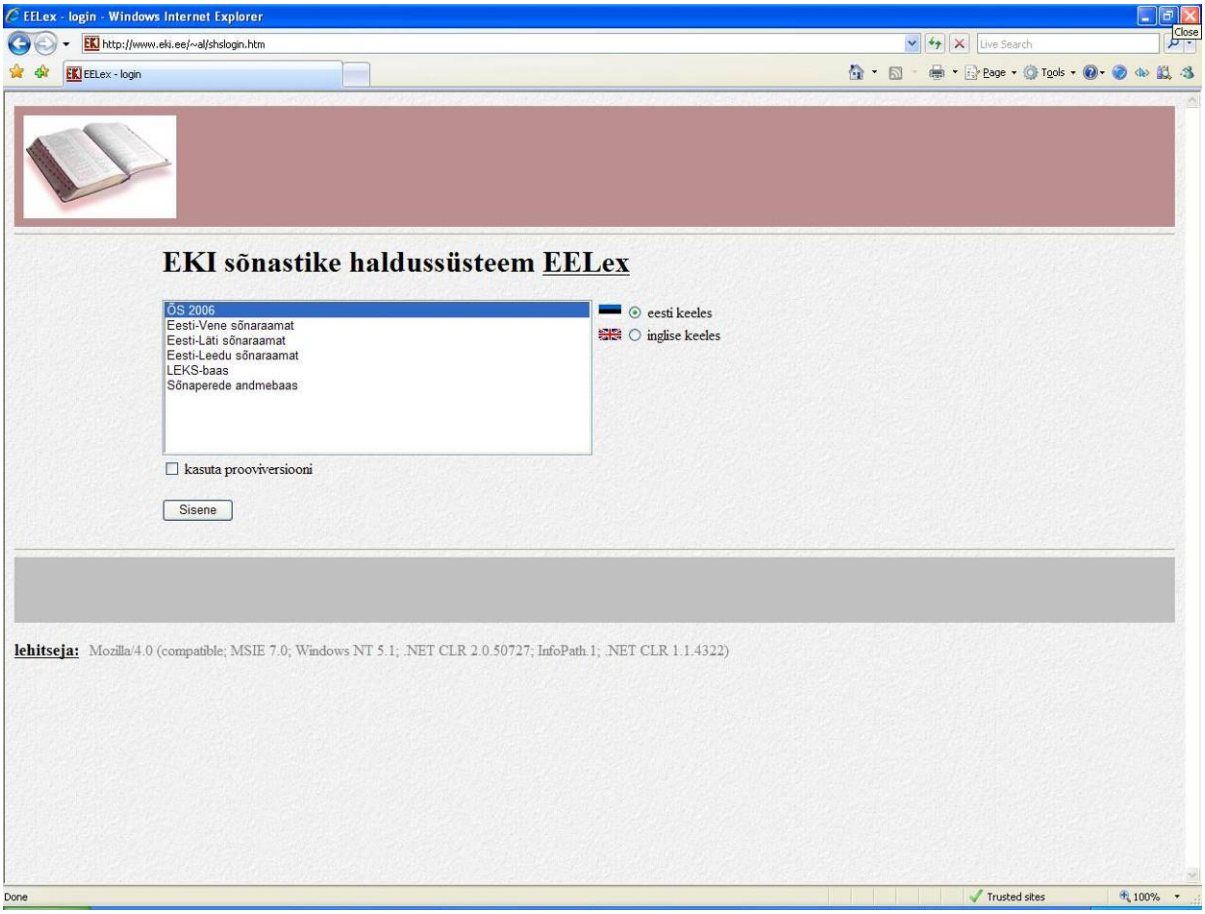


Joonis 12. Sõnaraamatu väljatrükk MS Wordis

Artikleid on võimalik Wordi eksportida kas päringu järgi või märksõnast märksõnani. Wordi dokumendi leheküljed on võimalik nummerdada alates soovitud väärtusest ning on võimalik valida, kas küljendatud kujusse tulevad ka vaate taustaga elemendid. Võimalik on valida ka Word dokumendi nimi ja asukoht.

2.9. EELexi kasutajaliides

EELexi on võimalik kasutada mitmes töökeeles. On võimalik eristada kasutajaliidese töökeelt ja sõnaraamatu töökeelt. Töökeele tervikuna määrab EELexi sisenemise parameeter, mis määratakse ühisel sisselogimise lehel (vt joonis 13).



Joonis 13. EELexi sõnastike sisselogimise ekraan

Sõnaraamatu töökeele valimisel kasutatakse keeleinfot sõnaraamatu artikli skeemis. Näide sõnaliigi definitsioonist skeemis:

```
<xs:element name="sl" type="d:sl_tyyp">
	<xs:annotation>
		<xs:documentation xml:lang="et">sõnaliik</xs:documentation>
		<xs:documentation xml:lang="en">part of speech</xs:documentation>
	</xs:annotation>
</xs:element>
```

KOKKUVÕTTEKS

Antud töö esimeses peatükis vaadeldi sõnastikusüsteeme üldiselt: milline funktsionaalsus peaks tänapäeval sõnastikusüsteemidel olema ja millised on mõned olemasolevad sõnastikusüsteemid. Eksisteerivaist sõnastikusüsteemidest on märgitud TshwaneLexi, Kuuba koolisõnastikku, Jibiki ning DEBII platvormi.

Teises peatükis kirjeldati Eesti Keele Instituudis välja töötatud sõnastike haldussüsteemi EELex. Teise peatüki alguses on lühidalt vaadeldud elektrooniliste sõnastike ajalugu EKI-s, EELexi ülesehitust ning millised sõnastikud on praegu EELexi hallata. On antud EELexi põhitoimingute – artiklite otsing, toimetamine, salvestamine ja küljendamine – kirjeldus ja tööpõhimõte.

EELexi iseloomustab AJAX-laadne suhtlus tööarvuti ja serveri vahel ning sõnastikuandmete säilitamine serveris XML vormingus. Artiklite XML sisu esitatakse tööjaama internetilehitsejas HTML veebilehtedena XSLT teisenduste kaudu. Iga sõnastiku struktuur vastab tema XSD skeemile, toimetamisoperatsioonid artiklis on kontekstipõhised ning salvestamisel kontrollitakse artiklite vastavust skeemile. Sõnastiku küljendus esitatakse tööjaamas MS Word tekstitöötlusprogrammis.

KIRJANDUS

Alegria, Iñaki; Arregi, Xabier; Artola, Xabier; Astiz, Mikel; Miyares, Leonel Ruiz 2006. A Dictionary Content Management System. – Corino, Elisa; Marello, Carla; Onesti, Cristina (eds.) 2006. Proceedings of the XII Euralex International Congress. Vol. I-II. Alessandria: Edizioni dell'Orso, 105–110.

Crystal, David 1986. The ideal dictionary, lexicographer and user. – *Lexicography: An emerging international profession*, ed. by Robert Ilson. Manchester: Manchester University Press in association with the Fulbright Commission, London, 72–81.

De Schryver, G-M (ed by) 2006. DWS 2006: Proceedings of the Fourth International Workshop on Dictionary Writing Systems, Turin, Italy, Turin University.

De Schryver, Gilles-Maurice 2003. Lexicographers' Dreams in the Electronic-Dictionary Age. – *International Journal of Lexicography* Vol. 16, No. 2, 143–199.

Erelt, Tiiu 2007. Õigekeelsussõnaraamatud läbi sajandi. – ÕSi lätted. Õigekeelsussõnaraamatud läbi sajandi. Eesti keele õigekirjutuse-sõnaraamat 1918 [faksiimiletrükk]. Eesti Keele Instituut. Koostanud Urmas Sutrop. Eesti Keele Sihtasutus, 5–34.

EVS = Eesti-vene sõnaraamat I–IV (V). Tallinn, Eesti Keele Sihtasutus. 1997–.

Hovy, E.; Ide, N., Frederking, R.; Mariani J.; Zampolli A. (eds.) 1999. Multilingual information management: Current Levels and future abilities.

Langemets, Margit 2000. Sõnaraamatu arvutilingvistiline analüüs. Väitekiri magistrikraadi taotlemiseks. Käsikiri Eesti Keele Instituudis.

Langemets, Margit 2003. Kas ükskeelne või kakskeelne sõnaraamat? – Toimiv keel I. Töid rakenduslingvistika alalt. Eesti Keele Instituudi toimetised 12. Tallinna Pedagoogikaülikool, Eesti Keele Instituut. Tallinn: Eesti Keele Sihtasutus, 151–177.

Langemets, Margit; Loopmann, Andres; Viks, Ülle 2006a. The IEL Dictionary Management System of Estonian. – de Schryver, Gilles-Maurice (ed.) 2006. DWS 2006: Proceedings of the Fourth International Workshop on Dictionary Writing Systems. Turin: Turin University, 11–16.

Langemets, Margit; Loopmann, Andres; Veldi, Enn 2006b. Märkmeid Torino leksikograafiakongressilt Euralex 2006. – Keel ja Kirjandus 12, 1012–1016.

LEKS-baas = Eesti keele leksikaalsemantiline andmebaas (esialgu sisaldab uusi sõnu ja tähendusi "Eesti kirjakeele seletussõnaraamatu" jaoks). http://www.eki.ee/~al/eexlogin.cgi

Loopmann, Andres; Sein, Kati; Viks, Ülle 2006. Sõnastike haldussüsteem Eesti Keele Instituudis. – Koit, Mare; Pajusalu, Renate; Õim, Haldur (toim). Keel ja arvuti. Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 6. Tartu: Tartu Ülikooli Kirjastus, 246–258.

Mangeot, Mathieu 2006. Dictionary Building with the Jibiki Platform. – Proceedings of the XII Euralex International Congress, 185–188.

Pala, Karel; Horák, Aleš 2006. From Web Pages to Dictionary: a Language-Independent Dictionary Writing System. – Proceedings of the XII Euralex International Congress, 199–204.

Sõnapered = Vare, Silvi. Eesti keele sõnapered. Käsikiri. Elektrooniline andmebaas Eesti Keele Instituudi sisevõrgus.

Viks, Ülle 1990. Sõnastike andmebaas: milleks, mis ja kuidas. – Ross, Jaan (toim). Arvutuslingvistika sektori aastaraamat 1988. Tallinn: Keele ja Kirjanduse Instituut, 167–175.

ÕS 2006 = Eesti Õigekeelsussõnaraamat ÕS 2006. Toim T. Erelt. Tallinn: Eesti Keele Sihtasutus.

Elektroonilised viited

Chinese Character Bible: http://www.globechinese.com/

Oxford Advanced Learner's Dictionary website:
http://www.oup.com/elt/catalogue/teachersites/oald7/?cc=gb

TshwaneLex. 2002-2006. http://www.tshwanedje.com/tshwanelex/

WikipediA. *The Free Encyclopedia.* http://www.wikipedia.org/