Tallinna Ülikool

Informaatika Instituut

Eestikeelsetes tekstides asesõnade asendamise algoritm

Bakalaureusetöö

Autor: Stanislav Gaistruk

Juhendaja: Erika Matsak

Autor: ..... „.....„ 2012

Juhendaja: ..... „.....„ 2012

Instituudi direktor: ..... „.....„ 2012

Tallinn 2012

Autorideklaratsioon

Deklareerin, et käesolev bakalaureusetöö on minu töö tulemus ja seda ei ole kellegi teise poolt varem kaitsmisele esitatud. Kõik töö koostamisel kasutatud teiste autorite tööd, olulised seisukohad, kirjandusallikatest ja mujalt pärinevad andmed on viidatud.

....
(kuupäev) (autor)

Sisukord

Sissejuhatus

Suheldes igapäevaelus sõprade, kolleegide, koolikaaslastega ja teiste inimestega meie ümber kasutame asesõnu ning laseme dialoogis kasutatud asesõnade tähenduse arvata vestluskaaslasel. Samal ajal loome enesele märkamatult seoseid vestluspartneri poolt öeldud asesõnade ja nende abil viidatava vahel. Seoste loomise kiirus sõltub muu hulgas suuresti ka sellest kui hästi tajume konteksti ning seda on lihtne teha kui asesõna tähendus peitub asesõnaga samas lauses. Mida suuremaks ja keerulisemaks(või abstraktsemaks) muutub kontekst seda keerulisem on neid seoseid luua – võib juhtuda, et oleme sunnitud küsima vestluspartnerilt täpsustamist öeldu osas.

Samalaadset selgitamist kujutab endas asesõnade käsitsi asendamine tekstis enne selle töötlemist dialoogsüsteemis DST, mis on loodud teksti transformeerimiseks predikaat-loogika valemite tasemele(Matsak, 2005). Tulenevalt uuritavate tekstide mahust tekkis vajadus abivahendi järele, mis aitaks need asendused võimalikult suurel määral automatiseerida(Matsak, 2009).

2010. aasta lõpuks valmis autori seminaritöö tulemusena asesõnade ja sünonüümide asendamiseks mõeldud programmi SPUM prototüüp, mis põhineb Erika Matsaku poolt kirjeldatud algoritmil(Lisa 1). Algoritm osutus pandora laekaks, mille avamine tõi autori vastamisi probleemidega millele lahendust esmapilgul ei paistnud. Ajakulu kasutaja tegevustele ning programmi poolt asendustel tehtavatele päringutele seadis küsimuse alla olemasoleva prototüübi otstarbekuse - seda just mahukamate ning keerukamate tekstide puhul. Põhjusprobleemide detailsem sõnastamine ja võimalike lahenduste otsimine oli liialt mahukas, et seda kirjeldada seminaritöö raames ning töö jätkub bakalaureusetöö vormis.

Käesoleva töö eesmärgiks on täiustada eelmainitud algoritmi, vähendada kasutaja rolli asenduste teostamisel ning arendada edasi SPUM-i lähtudes dialoogsüsteemi DST vajadustest.

Teema valikul motiveerisid varasem huvi tehisintellektiga seotu vastu (ka futuristlikud filmid ja spekulatsioonid), võimalus läbi bakalaureusetöö protsessi saada aimu sellest, miks tänase seisuga ei eksisteeri inimese kõne täiel määral korrektselt interpreteerivat süsteemi ning lootus jätta oma panus teemaga seonduvale teadusharule.

Selleks, et käesolevat tööd oleks lihtsam jälgida on järgnevalt välja toodud mõned tekstis kasutatud ja/või töö teemaga seonduvad terminid ning nende seletused:

Asesõnad - tüüpjuhul iseseisvad mittetäistähenduslikud sõnad, mis muutuvad käändes ja arvus ning käituvad lauses nimi-, omadus- või arvsõnade taoliselt, kuid on nendega võrreldes „tühjema", abstraktsema sisuga, nt *mina, tema, see, niisugune, iga, mitu*.

Antetsedent – eellane; sõna või fraas mida asesõna esindab. Näiteks lauses *„Tudeng sööb nuudleid ja need on maitsvad"* on antetsedendiks sõna „nuudlid*"* mida esindab teises osalauses sõna „need*"*.

Morfoloogia – valdkonnana uurib sõnade struktuuri ja ehitust; morfoloogilise analüsaatori kohta tuleb juttu teises peatükis.

Dialoogsüsteem - arvutiprogrammi, mis suudab inimesega teatud teema(de)l suhelda, kusjuures suhtlus toimub loomulikus keeles ning kõne või teksti vahendusel(Koit, 2003).

1. Olemasolev teave ja rakendused

Selles peatükis tutvustame SPUM programmi arendusega seotuid keeletehnoloogilist tarkvara. Nimetatud rakenduste funktsionaalsuste kombineerimine ja SPUM-i integreerimine võimaldab tõsta SPUM-i kiirust ning vähendada kasutaja sekkumise vajadust.

1.1 Integreeritud rakendused

Eesti Keeletehnoloogia Sihtprogrammi „Keeletehnoloogiavahendid eesti keele jaoks“ raames(allikas) on valminud hulk rakendusi, mille funktsionaalsuste kombineerimine ja SPUM-i integreerimine võimaldas seda edasi arendada. Kirjeldatud rakenduste seas on ka väljaspool eelmainitud sihtprogrammi valminud süntaksianalüsaator EstCG Parser.

1.1.1 ELA - Eesti keele lausete eraldaja tekstist (testversioon).

Lausestaja võtab sisendiks tekstifaili ning faili töötlemise järel on väljundfailis on iga sõna omaette real, lause algus on tähistatud märgendiga *LA* ja lause lõpp märgendiga *LL*. Sulud, jutumärgid, komad, lauselõpu punktuatsioon jms on tõstetud sõnast lahku ning samuti erinevatel ridadel. Lausestaja väljundiks oleva tekstifaili sisu on vormingu poolest morfoloogilise analüsaatori sisendiks sobiv(Vaino, 1999).

1.1.2 ESTMORF - Eesti keele morfoloogiline analüsaator ja süntesaator

Morfoloogiline analüsaator on käesoleva töö ja rakenduse SPUM üks alustalasid. See on programm, mis sõna vormist lähtudes määrab selle sõna struktuuri (nt. tüvi, järelliide, lõpp), sõnaliigi ja käände või pöörde. Morfoloogiline süntesaator teeb vastupidist sellele, mida teeb analüsaator. Lähtudes sõna algvormist ja grammatilisest infost – sõnaliigist, käändest või pöördest - määrab süntesaator selle sõna konkreetse muutevormi(Kaalep, 1999).

1.1.3 TAHMM - Morfoloogiline ühtestaja (beetaversioon)

Morfoloogiline ühestaja on programm, mis morfoloogilise analüsaatori väljundist valib iga sõna analüüsidest just selle, mis antud konkreetses kontekstis on õige. Enne ühtestaja kasutamist tuleb analüsaatori väljundfail töödelda ühtestajale sobivaks, milleks kasutatakse ühtestajaga kaasas olevat konverteerijat. Rakenduse kasutusjuhendis on välja toodud ka ühtestaja täpsus - 93%(Kaalep, 1999).

1.2 EstCG Parser – Süntaktsianalüsaator

Loomuliku keele süntaksianalüsaator on programm, mis saab sisendiks morfoloogiliselt analüüsitud teksti (s.t on leitud iga sõna tüvi, lõpud, sõnaliik, kääne või pööre jms) ning väljastab süntaktiliselt analüüsitud teksti (leitud on igas lauses alus, öeldis, sihitis jt lauseliikmed). Enamasti esitatakse süntaktiline kirjeldus märgendite abil, s.t iga sõnavormi juurde kirjutatakse selle sõnavormi morfoloogilisi ja süntaktilisi omadusi kirjeldav märgend või märgendite kombinatsioon(Müürisep, 2002).

1.3 Comverse Inc. patenteeritud meetod ja süsteem asesõnade antetsedentide leidmiseks

Asesõnade mitmetähenduslikkuse vähendamise meetod ja süsteem on Ühendriikides patenteeritud ettevõttele Comverse Inc.

Illustreeriva näitena on patendis kirjeldatud fiktiivset juhtumit, kus süsteemi rakendatakse telefoniraamatuna, mille kasutaja võib näiteks peale automaatvastajale jäetud kõne lõppu öelda suuliselt süsteemile: „*Call him back“*(tõlge: „helista talle tagasi“) ning peale kontekstiga sobiva vaste leidmist andmebaasist ning kasutajapoolset kinnitust toimub automaatne kõne algatus. Olemasolevaid seosied saab muuta nii süsteem kui ka kasutaja

vastavalt nende kasutuse ajaloole. Seoste lisamine, päring ning haldus on algoritmilistelt kirjeldatud(Lisa 2 ja 3).

Patendis kirjeldatakse asesõnade ning nende poolt viidatavate andmete(nimisõnad, nimisõnafraasid vms) vaheliste seoste haldamise ja kasutamise süsteemi. Kasutatakse kaht andmebaasi, kus näiteks uue kirje lisamisel sisestatakse kasutajaga koostöös vastav info ka asesõnade andmebaasi(nimisõnafraasi „John Smith“ puhul oleks selleks „he“). Süsteem võib lisamise teha ka ise näiteks kui viidatava objekti puhul on nime asemel telefoninumber(eelmainitud näite puhul), ID, vms. Nii viidataval sõnal kui ka asesõnal võib andmebaasis olla mitu vastet(nimisõna puhul erinevad asesõnad ja asesõna puhul erinevad nimisõnad). Tegemist on õppiva süsteemiga, mille realiseeritud instants saaks toimida mingi kindla eesmärgi täitmisel, näiteks eelmainitud telefoniraamat(Panttaya, 2006).

2. Dialoogsüsteemi DST tutvustus

Selles peatükis kirjeldatakse Erika Matsaku poolt arendatavat dialoogsüsteemi DST. Tegemist on süsteemiga, mis kasutajaga koostöös on suuteline transformeerima loomuliku keele teksti predikaat-loogika valemite tasemele - selleks on süsteemil kaks režiimi.

Esimene režiim on oma olemuselt programmi õpetamine, kus kasutaja transformeerib lause samm-sammult, viies seda järjest lähemale kujule kus seda on võimalik vaadelda kui loogilist avaldist ning transformatsioon lõppeb lausest formaalse loogilise konstruktsiooni moodustamisega.

“Mina sõidan uue autoga”→ “Mina sõidan autoga ja auto on uus”→ $P_1(q_1, x_1)\&P_2(x_1)$.

Transformatsiooni käigus jäetakse meelde nii sõnade järjekord kui ka selle muutumine läbitud sammude jooksul. Lisaks läbib iga lause ka morfoloogilise analüüsi(Filosofti arendatud eesti keele HTML-i morfoloogilise analüsaatori abil), mille tulemuseks on loogilis-morfoloogiliste tähiste skeem.

Teises režiimis toimub transformeerimine automaatselt tuginedes sellele, et võrdsete morfoloogiliste skeemide puhul on tulemuseks ka võrdsed loogilised konstruktsioonid formaalsel kujul.

2.1 Ootused SPUM-i funktsionaalsusele lähtuvalt DST vajadustest

Oluline on enne transformeerimist tekst puhastada mitmetähenduslikest sõnadest, et vähendada transformeerimise käigus nõutud käsitsi sekkumist DST töösse. Asesõnad moodustavad neist suure osa. Esialgse algoritmi(Lisa 4) järgi arendatud moodulis tuli nende asendamiseks antetsedentidega iga uuritava asesõna kohta kirjutada eraldi lahtrisse selle poolt esindatav sõna või fraas, mille leidmise kiirus sõltus sellest, kui hästi kasutaja algteksti mõistis(antedetsendi leidmiseks pidi kasutaja selle SPUM-i poolt kuvatavast teksti osast üles otsima). Selline toimimisloogika ammendas end algteksti pikkuse kasvades ning efektiivsus oli võrredav asenduste tegemisega ilma abivahendita.

Tekstis leiduvate sõnade morfoloogiline analüüs võimaldab meil leida üles ka nimisõnad, mis tihtipeale ongi asesõnade antetsedentideks. Vaatamata sellele, et nimisõnad pole ainsad võimalikud eellased, saame siiski suurel määral vähendada asenduste ajakulu kui asendatava asesõna morfoloogilise(kääne, ainsus/mitmus jne) info alusel sünteesida sobival kujul nimisõna. Selleks tuleb välja selgitada kui suur peaks olema teksti osa(lausete arv), milles sisalduvad nimisõnad kasutajale asenduseks välja otsitakse. Kuna välistada ei saa ka erandjuhtumeid(näiteks pikema ajalooga asesõnad), peaks konteksti pikkus olema dünaamiline väärtus ja kasutaja poolt muudetav igal asendusel.

3. SPUM-i tutvustus ja funtsionaalsuse kirjeldus

Rakendus on kirjutatud Java programmeerimiskeeles, ning käesoleva töö tulemusena valminud versioonis on ohverdatud Java platvormisõltumatus, et integreerida eelnevalt kirjeldatud keeletehnoloogilised rakendused.

3.1 SPUM-i installeerimine ja teksti esmane töötlus

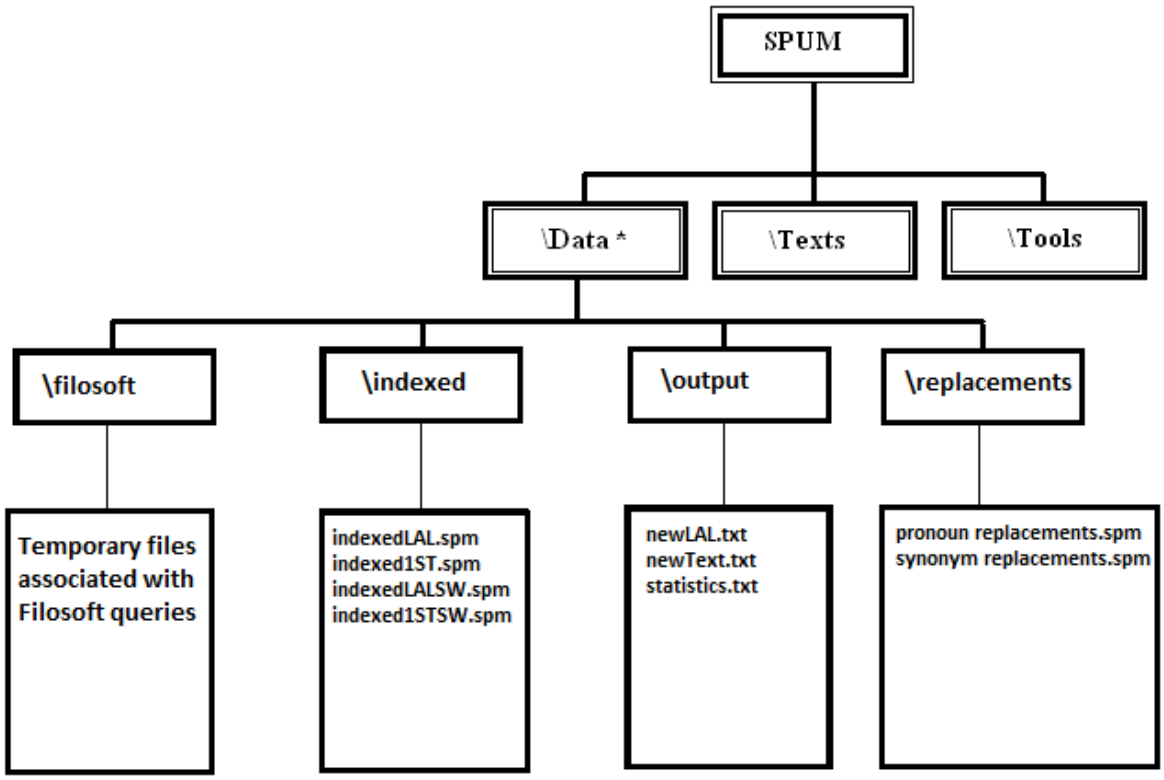
Programmi on edukalt käivitatud ning kasutatud järgmistel operatsioonisüsteemidel:

- Microsoft Windows XP Professional ja Home Edition
- Microsoft Windows Vista Ultimate
- Microsoft Windows 7 Enterprise

Installeerimisel ning algtöötlemisel teostatakse automaatselt järgnevad toimingud:

1. Vajalike kaustade ning failide loomine
2. Algteksti lausestamine
3. Lausestatud algteksti morfoloogiline analüüs
4. Analüüsi tulemuste ühtestamine
5. Sõnade indekseerimine nii lausestaja ja ühtestaja väljundis
6. Asesõnade kopeerimine eraldi faili

Alljärgneval joonisel on kujutatud SPUM-i faili- ning kaustastruktuur. Kaustas *Tools* asuvad Filosofti käsureapõhised rakendused ning kaustas *Texts* algtekstid. Kui kasutaja ei vali töötluseks ühtegi tekstifaili püüab SPUM laadida *Texts* kaustast faili *tekst.txt*.



Joonis 1. SPUM-i kaustastruktuur

* Kausta *Data* sisu(kaustad ja failid) genereeritakse installeerimisel automaatselt.

indexedLAL.spm – lausestaja väljund(indekseeritud)

indexed1ST.spm – ühtestaja väljund(indekseeritud)

indexedLALSW.spm – indekseeritud lausestaja väljund kus lisaks on iga sõna puhul kirjeldatud selle indeks lauses(SW – sentencewise).

indexed1STSW.spm – indekseeritud ühtestaja väljund kus lisaks on iga sõna puhul kirjeldatud selle indeks lauses.

newLAL.txt – lausestatud tekst kus asesõnad on asendatud nende antetsedentidega

newText.txt – algtekst asendustega

statistics.txt – asendustel automaatselt kogutava info kokkuvõte

pronoun replacements.spm – kasutaja teostatud asesõnade asendused(vt peatükk 4.3)

3.2 Indekseerimine

Selleks, et oleks võimalik pöörduda mingi kindla sõna või lause poole indekseeritakse lausestaja ja ühtestaja väljundis sõnad ja laused ning tulemused kirjutatakse kaustas *indexed* olevatesse failidesse. Indekseerimine toimub ka lause vaates, st iga sõna puhul kirjeldatakse tema positsioon lauses(tulemuseks *indexedLALSW.spm*).

LA-[lause nr]
...
[sõna nr lauses]_[sõna nr tekstis]_sõna
...
LL-[lause nr]

Näitena on allpool toodud välja üks täielikult indekseeritud lause falilist *indexedLALSW.spm* peale edukat installeerimist ning selle kirjeldus.

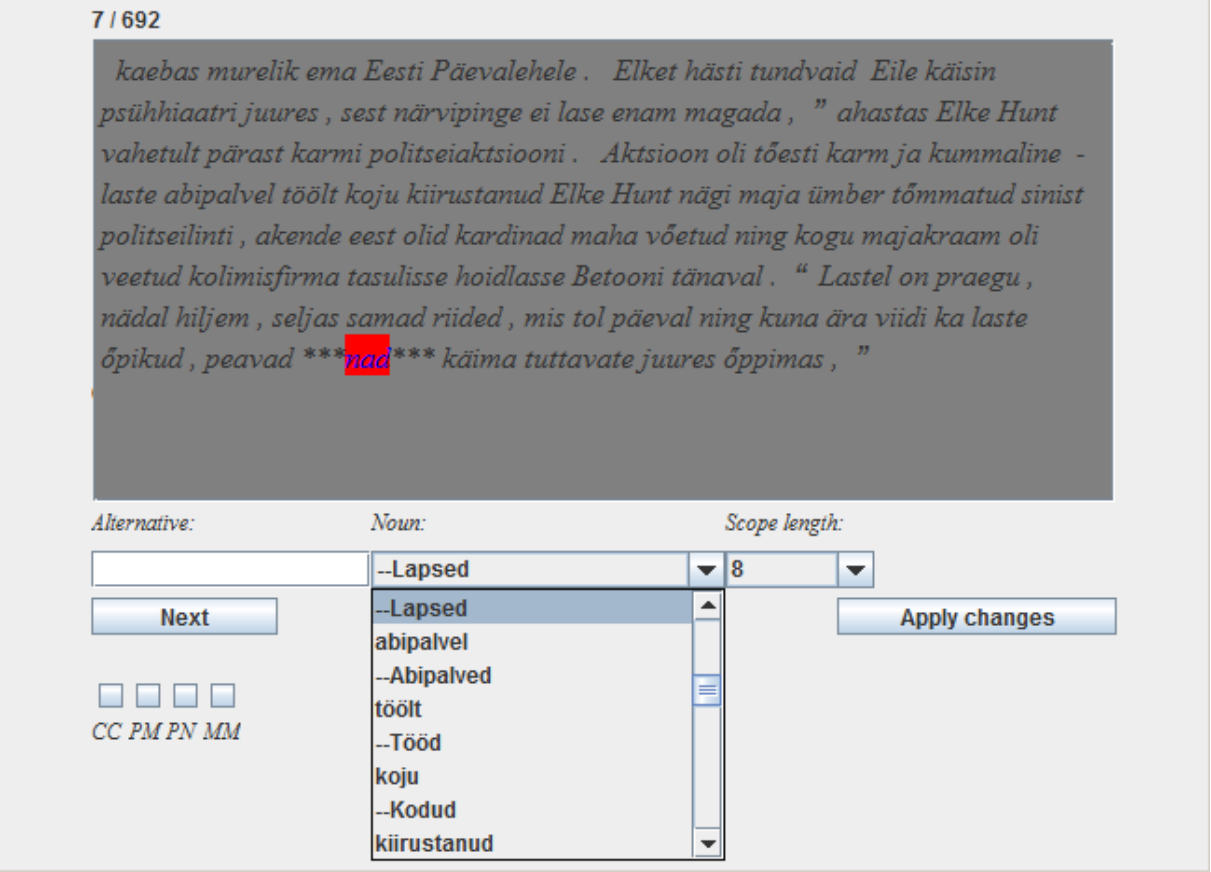
LA-943
1_8428_Merel
2_8429_polnud
3_8430_ei
4_8431_tualetti
,
5_8432_ei
6_8433_duširuumi
.
LL-943
Sõna positsiooni lauses näitavale numbrile järgneb sõna indeks kogu teksti vaates. Samasugune indekseerimine toimub ka *indexed1STSW.spm* puhul. Erinevalt indekseeritakse ainult asesõnadest koosnev fail, kus esimesena näidatakse, mitmenda asesõnaga tekstis

tegemist on. Juhul kui morfoloogiline ühtestamine ebaõnnestub(allpool olevas näites võib olla tegu nii mitmuse kui ka ainsuse vormiga) lisab SPUM asesõna indeks järele ’*’ märgi.

2*_70_kelle kes+0 //_P_ pl g, // kes+0 //_P_ sg g,
3_141_samad sama+d //_P_ pl n,

3.3 Asenduse teostamine SPUM-i abil

Kasutajale kuvatakse konteksti väljal valitud hulk lauseid, kus asesõna on värvitud punasega. Tekstis leitud nimisõnad on lisatud eraldi rippmenüüsse ning morfoloogilise info alusel sünteesitud nende vastavad vormid. Joonisel oleva näite puhul on asesõnaks „nad“, mis on mitmuse nimetavas vormis. Tekstis eespool leitud nimisõnad on samuti pandud mitmuse nimetavasse käändesse, sh ka asesõnaga samas lauses esimene sõna „Lastel“, mis peale sünteesimist on rippmenüüs uuel kujul: „Lapsed“. Enne asenduse teostamist nupuga „Next“ võib kasutaja vajadusel muuta algustähe kas suureks või väikeseks kasutadaes valikut „CC“(change case). Antud näite puhul on tarvis see valik teha, et tagada lause korrektsus(Joonis 2).



Joonis 2. Asenduse teostamine

Lisaks algustähe muutmise valikule on statistika kogumiseks võimalik lisada asendusele ka infot antetsedendi kohta:

PM- partial match(osaline vaste)

PN- phrase noun(nimisõnafraas)

MM- multiple match(võimalikke vasteid on rohkem kui üks)

Asenduse teostamine ei toimu reaalajas, selle asemel kogutakse asenduse kohta käiv informatsioon ning paigutatakse faili *pronoun replacements.spm* uuele reale. Ideaalkorras näeb üks asendus selles failis väja järgmiselt:

#pronoun -> 4SUBJ [20] #noun -> 1NN> [1] ***1154_nad -> ***2***106_lapsed #FIC, true# ((20 -> 19))***

1. ***#pronoun -> 4SUBJ [20]*** - asesõna oli lause neljas subjekt ja asus lauses 20. kohal.
2. ***#noun -> 1NN> [1]*** - nimisõna, millega ta asendati, oli lause esimene nimisõnast eestäiend ning kokkuvõttes lause 1. sõna.
3. ******1***154_nad -> ***2***106_lapsed*** - tekstis oli asesõna 154. ja nimisõna 106. sõna.
4. ***#FIC, true#*** - Found In Context(võib olla ka näiteks #NIC# ehk Not In Context) mis tähendab, et asesõna leiti antud juhul lähima 5 lause seas.
5. ***((20 -> 19))*** - asesõna oli 20. lauses ja nimisõna leiti 19. lausest. Siinkohal on tegu ilmselgelt veaga, mille põhjustab kasutusel oleva lausestaja suutmatus mõningatel juhtudel lause lõppu ning algust määrata.

3.4 Muudatuste kinnitamine ja statistika kogumine

Teostatud asendusi on võimalik igal ajahetkel ka algteksti üle kanda, muudetud kujul algtekst koos lausestaja väljundiga asuvad *\output* kaustas. Samas kaustas asub ka fail *statistics.txt,* kuhu muudatuste kinnitamisel kirjutatakse asendustel kogutud informatsiooni kokkuvõte.

Kokkuvõte sisaldab muu hulgas järgmist informatsiooni:

- Asesõna maksimaalset, minimaalset ning aritmeetilist keskmist kaugust nimisõnast
- Asenduste arv, kus asesõna ja nimisõna täitsid sama süntaktilist rolli
- Asenduste arv, kus asesõna ja nimisõna täitsid erinevat süntaktilist rolli
- Kõige sagedamini esinenud süntaktiliste rollide kombinatsioon

Samal ajal statistika kogumisega toimub ka asesõnade asendamine- esialgne lausestaja väljund kirjutatakse ümber faili *newLAL.txt*, kus iga realiseeritud asenduse puhul kirjutatakse asesõna asemel juba tema antetsedent. Sellest failist pannakse kokku juba muudetud kujul algtekst(*newText.txt*).

3.5 Jõudlus

Testimiseks kasutatud arvuti parameetrid:

- Operatsioonisüsteem: Windows 7 Enterprise 32-bit (6.1, Build 7600)
- Tootja: ASUSTeK Computer Inc.
- Mudel: N61Ja
- Protsessor: Intel(R) Core(TM) i5 CPU M 450 @ 2.40GHz (4 CPUs), ~2.4GHz
- Mälu: 4096MB RAM

Andmed teksti kohta:

- Sõnade arv: 209237
- Tekstifaili suurus: 2026 KB

Kiirus:

- Tekstifaili tükeldamine: 27s
- Algteksti lausestamine, morfoloogiline analüüs ja indekseerimine: 22s
- Igal asendusel kulub 1-2 sekundit võimalike antetsedentide leidmiseks ning sünteesimiseks.

Algteksti tükeldamine on vajalik, sest juhul kui algteksti pikkus ületab teatud piiri, ei suuda peatükis 2.1 nimetatud rakendused neid täies mahus töödelda ning tulemuseks on vigane väljund. Algtekst jaotatakse eraldi failidesse, kus iga fail sisaldab maksimaalselt 3000 rida teksti.

4. Andmekaeve algoritmi arendusvõimaluste väljaselgitamiseks

Käesolevas peatükis kirjeldatakse andmekaeve protsessi ning tulemusi. Selle läbiviimisel otsustas autor valida mitmekülgse ning rohkem asesõnu sisaldava teksti, mis lisaks võimaldaks paremini jälgida asesõnade kasutust. Valituks osutus ajakirja „Kroonika“ 16 artiklist koosnevat 2001-01-19 väljaanne, mis on kättesaadav Eesti keele segakorpusest. Korpusest alla laetud *.TEI* formaadis ja kodeeringus fail muudeti *.TXT* failiks kasutades programme Oxygem XML Editor ning Adobe Reader 9.

Andmekaeve eesmärk on hinnata SPUM-i efektiivsust ning leida uusi võimalusi asendustele kuluva aja vähendamiseks.

4.1 Andmekaeve käigus tehtavate asenduste liigitamine

Kuna oluline osa andmekaeve tulemuste analüüsist on ka SPUM-i efektiivsuse liigitame asendused järgmiselt:

- Arenduse tulemusena lihtsustatud - antetsedetsent on SPUM-i poolt pakutute seas ja õiges vormis.
- Arenduse tulemusena osaliselt lihtsustatud - antetsedetsent on SPUM-i poolt pakutute seas aga vales vormis.
- Probleemsed teostatavad – antetsedent ei ole SPUM-i poolt pakutute seas või on seal osaliselt
- Teostamatud probleemsed asendused

4.1.1 Arenduse tulemusena lihtsustunud asendused

Need on asendused mille puhul on täidetud järgmised kriteeriumid:

1. Asendamise teostamise järel säilib lause loomulikkus ning grammatilisus
2. Antetsedent on SPUM-i poolt pakutute seas olev nimisõna ning tekstiga sobivasse vormi sünteesitud

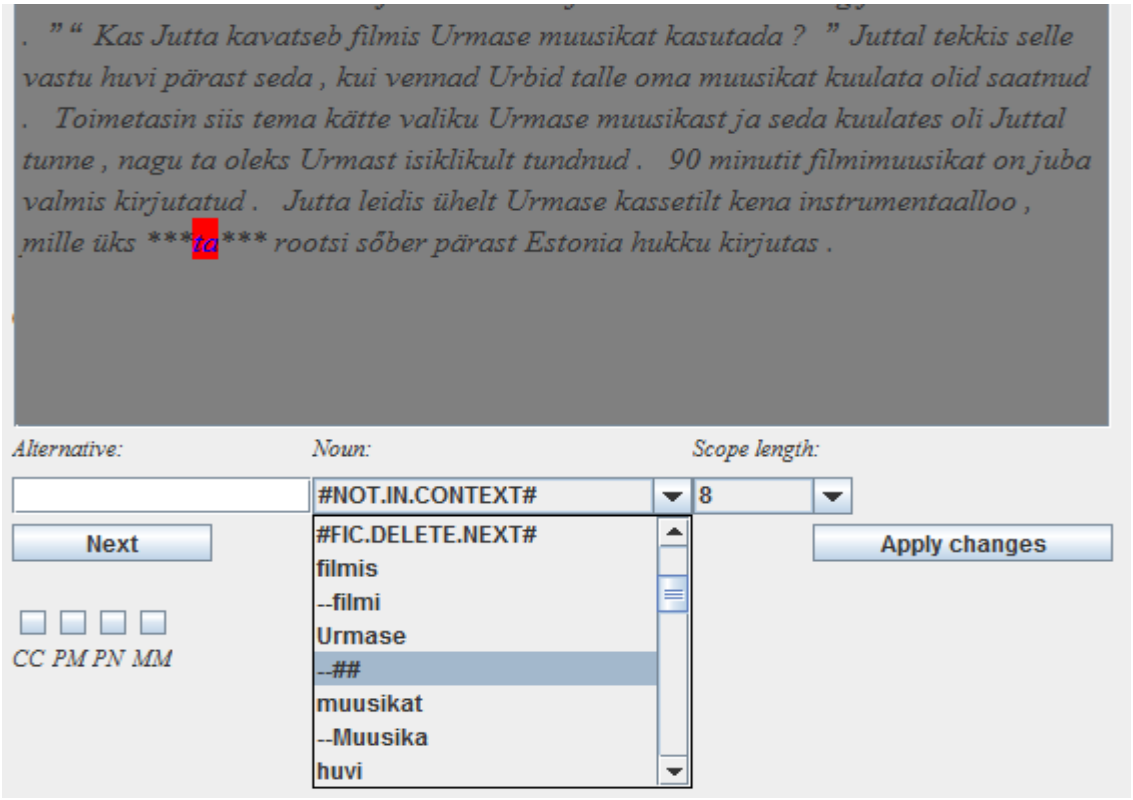
Need kriteeriumid võimaldasid hinnata SPUM-i võimekust ning välja tuua arendusvajadusi. Kasutaja ainsateks tegevusteks oli nimisõna valimine rippmenüüst ning vajadusel algustähe muutmine.

4.1.2 Arenduse tulemusena osaliselt lihtsustunud asendused

Need on asendused mille puhul on täidetud järgmised kriteeriumid:

1. Asendamise teostamise järel säilib lause loomulikkus ning grammatilisus
2. Antetsedent on SPUM-i poolt pakutute seas olev nimisõna aga tekstiga mittesobivas vormis.

Nende kriteeriumite seadmine võimaldas välja selgitada asenduste hulga, mille puhul saab SPUM koguda informatsiooni asesõna ja nimisõna kohta ka sellisel juhul kui pakutud nimisõna pole õiges vormis. Asenduse korrektseks teostamiseks tuli kasutajal valida rippmenüüst vastav nimisõna ning kirjutada antetsedendi õige või täielik vorm vastavasse lahtrisse. Selliseid juhtumeid tuli kõige sagedamini ette kui SPUM andis morfoloogilisele analüsaatorile vale või ebakorrektse info. Alltoodud näite puhul on ebaõnnestunud sünteesida „Urmase“ ainsuse nimetav vorm.



Joonis 3. Ebaõnnestunud nimisõna süntees.

4.1.3 Teostatavad probleemsed asendused

Probleemsete asenduste all peab autor silmas juhtumeid, mis vastavad järgmistele kriteeriumitele:

- Asendamise teostamise järel säilib lause loomulikkus ning grammatilisus
- Antetsedent puudub SPUM-i poolt pakutud nimekirjas või on seal osaliselt

Paljud asendused, mille autor SPUM-i testimisel vahele jättis olid teostatavad, kuid nende puhul ei leidunud antetsedent nimisõnade valikus või oli seal osaliselt. Näited kriteeriumitele vastavates juhtumitest on järnevad:

1. Nimi- või tegusõnafraas – „*Alles siis, kui lavastaja Penny Marshall talle filmis “Renaissance Man” kõrvalosa pakkus,* ***avastas*** ***Mark, milles peitub tema***

***tõeline anne**. Pärast **seda** ei ole ta enam elukutseid vahetanud, mikrofon on nüüdseks nurka visatud ja modellilepingud aegunud.“*

2. Asesõna mitmuse vormis - *„Kuigi **Andrei ja Leonid** on pühendunud spordile, pole **nad** enese harimist unustanud.“*

3. Arvud jms - *Sõltumata sellest, millisest lepingust Elke Hunt parajasti räägib, kinnitab ta ajakirjanduses, et **maksis Tiit Reinfeldile 650 000 krooni**. Kui jutt käib eellepingust, nimetab Elke **seda** käsirahaks, ostu-müügilepingu puhul aga esimeseks sissemakseks.*

4.1.4 Teostamatud probleemsed asendused

Arvestusväärne hulk asendusi on praeguse asendusloogikaga teostamatud kuna need erandlid eeldavad asenduse läbiviimiseks lause struktuuri muutmist. Järgnevalt on välja toodud näited reeglipäratute asesõnadega(Erelt, Erelt, Ross, 1997):

- Näitavad asesõnad: *„Pealegi - **kes** siis Narva elu edendab, kui kõik minema jooksevad... ”*
- Küsivad-siduvad asesõnad: *„Noor naine, **kelle** tagakiusamisest päevalehed nädal otsa rääkinud, on Elke Hunt."*
- Isikulised asesõnad: *„**Mina** oleksin küll olnud valmis balilasena sündima.“*
- Määratlevad asesõnad: *"...Reinfeld tegeles selle majaga **ise**..."*
- Liitasesõnad: *„...kusjuures **igaüks** neist kasutab...“*
- Umbmäärased asesõnad: *„**Keegi** on alati kõrval olnud.“*

Siinkohal on oluline mõista, et tegemist on eranditega ning on käesoleva uurimuse käigus on teostatud asendusi ka reeglipäratute asesõnade puhul.

4.2 Andmekaeve tulemuste kokkuvõte

Kokku leidis SPUM algtekstist 8567 sõna, millest 824 osutusid asesõnadeks(9,6% kogu sõnade arvust).

Kõikidest asesõnadest 249 oli neid, mille antetsedent oli asenduse ajal ka SPUM-i poolt pakutud nimekirjas(nende asendused vastasid petaükkides 5.1.1 või 5.1.2 kirjeldatud kriteeriumitele). Nende asendamisel oli SPUM-il võimalik koguda informatsiooni, mille lühikokkuvõte on järgmine:

- Asesõna ja antetsedendi vaheline kaugus oli -3 kuni 64 sõna ning 0 kuni 4 lauset(negatiivse väärtuse põhjuseks oli antetsedendi paiknemine lauses pärast asesõna).
- Asesõna keskmine kaugus antetsedendist oli 16 sõna(lauseindeksite keskmine vahe oli 1,5).
- 70 asenduse puhul täitsid asesõna ja nimisõna lauses sarnast süntaktilist rolli ning ülejäänud 179 puhul erinevat.
- Populaarseimaks süntaktiliste rollide kombinatsiooniks oli SUBJ -> NN (asesõna oli subjekti ning antetsedent nimisõnalise täiendi rollis). Kui esmapilgul pole selles midagi üllatavat(SPUM otsib ju kasutajale välja tekstis leiduvad nimisõnad) siis tasuks arvestada, et selline kombinatsioon esines kõigest 24 korral 249-st.

5. Ettepanekud asenduste automatiseerituse tõstmiseks

Käesolevas peatükis toob autor välja võimalused asenduste veelgi suuremal määral automatiseerimiseks tuginedes andmekaeve käigus ning selle tulemustest kogunenud informatsioonile.

5.1 Nimisõnafraaside lisamine antetsedentide loetelusse

Üldiselt osutusid kiireimateks asendused, mille puhul antetsedent oli SPUM-i poolt pakutute seas ning õiges vormis, mis tähendab seda, et esmajoones tuleks selliste asenduste osakaalu suurendada. Eesti Keeletehnoloogia Sihtprogrammi raames on valminud muu hulgas ka tarkvara EstNPTool – eestikeelsete nimisõnafraaside filtreerija(Koit, Roosmaa, 1999), mille integreerimine SPUM-i võimaldaks täiustada pakutud antetsedentide loetelu. Kahjuks puudub kirjeldatud rakendusel funktsionaalsus, mis lubaks automaatselt iga SPUM-is tehtava asenduse puhul võtta mingi kindla hulga lauseid ning need filtreerijale sisendiks anda(näiteks hetkel SPUM-i integreeritud rakenduste puhul on üheks parameetriks sisendfaili asukoht kettal). Lühidalt - puudub võimalus käivitamisel seada parameetrina sisendfaili.

5.2 Teostamatute asenduste edasiuurimine

On lauseid, mille puhul asendamisega kaoks lause loomulikkus, kuid oleks sellest hoolimata DST abil loogiliseks avaldiseks transformeeritav, näitekst peatükis 5.1.4 väljatoodud juhtumitest esimese kolme puhul: näitavad, küsivad-siduvad ja mõned isikulised asesõnad.

Edasine automatiseerimine oleks võimalik läbi siduvate asesõnade uurimise- näiteks laused, kus esimesena mainitakse antetsedenti, ning temale järgneb siduv asesõna.

„Noor naine, kelle...“

Siduva asesõna *kelle* asemel saaks nimisõnafraasi aktsepteeriva süntesaatori olemasolul automaatselt kasutaja sekkumiseta kirjutada *noore naise*. Arendusvõimalused tänase seisuga lubaksid automatiseerida selliseid asendusi juhul, kui antetsedent pole fraas. Sarnaselt saab tõsta automatiseeritust ka erineval kujul lauses esinevaid siduvaid asesõnu, kuid see eeldab paremat arusaamist osalausetest ning DST toimimisest nende puhul.

5.3 Täiustatud algoritm asesõnade asendamiseks

Algoritmi(Lisa 5) täiustamine toimus paralleelselt mooduli arendusega. Algoritmi kirjapanekul on autor püüdnud olla detailsem kuid seda vaid kindla piirini(tagades algoritmi kerge loetavuse). Järgnevalt kirjeldatakse täpsemalt samme, mida SPUM-i arendamisel oli keerulisem realiseeerida:

1. *Sentence parsing* – Lausestamine. Kui algoritm realiseerimisel kasutada SPUM-i integreeritud keeletehnoloogilist tarkvara tuleb arvestada, et lausestaja ei suuda teatud mahtu ületavat tekstifaili täielikult töödelda. Selle ületamiseks võib algteksti faili tükeldada umbes 60KB suurusteks failideks ning need eraldi lausestada, analüüsida, ühtestada ning seejärel uuesti ühendada enne indekseerimist.

2. *Indexing* – Lausestaja ning ühtestaja väljundi indekseerimine. SPUM lisab indeksi ainult nende ridade ette, mis algavad tähega, et mitte indekseerida kirjavahemärke.

3. *Get and display context* – Asesõnale eelneva tekstiosa kuvamine kasutajale. SPUM moodulis on selleks säilitatud sõna *globaalne indeks* ka asesõnade failis. Indeksi abil otsitakse asesõna üles indekseeritud lausestaja väljundfailist, eelnevalt hoides puhvris mingi hulga asesõnale eelnenud lauseid. Lausete hulk määratakse konteksti pikkusega.

4. *Get and list potential antecedents* – Võimalike antetsedentide kuvamine kasutajale. Moodulis on see realiseeritud sarnaselt asesnõnale eelneva tekstiosa kuvamisega, kuid lausestaja väljundi asemel läbitakse ühtestaja väljund ning korjatakse konteksti raadiuses välja sõnad mille morfoloogilise analüüsi tulemus sisaldab endas kirjet

_S_(nimisõna), *_Y_(lühend)*, või *_H_(pärisnimi)*. Võimalikud antetsedendid paigutatakse rippmenüüsse, kust kasutajal on võimalik valida sõna kas esialgsel või morfoloogilise süntesaatori abil saavutatud kujul.

5. *Apply replacements in original textfile* – asenduste ülekandmine lähteteksti. Ülekandmisel tuleks kõigepealt asendused realiseerida lausestaja indekseeritud väljundfailis. Seejärel eemaldada lausestaja lisatud lauselõpu ja –alguse märgendid ning indeksid. Kuna lausestajaga kaovad uut lõiku märgistavad reavahetused võib need vajadusel lähtetekstis eelnevalt märgistada.

Kokkuvõte

Käesoleva töö tulemusena valmis asesõnade asendamise täiustatud algoritm ning paralleelselt sellega on moodul SPUM viidud asesõnade asendamiseks sobivale kujule. Kasutaja rolli ning ajakulu asendustele on tunduvalt vähendatud läbi antetsedentide leidmise lihtsustamise ning rakenduse ergonoomilisuse tõstmise. Käesoleva töö käigus uuritud teksti puhul võimaldas moodul automaatselt koguda informatsiooni asesõnade ja nende antetsedentide kohta(sõnaliik, süntaktiline roll, asukoht tekstis ja lauses) ligikaudu 30%-l juhtudest. Asendustelt kogutav info toob esile seoseid asesõnade ja nende antetsedentide vahel mis omakorda võimaldavad veelgi suuremal määral vähendada kasutaja rolli asendustel.

Töö keerukuse tagas olemasoleva metoodika ning teabe puudus. Samas võimaldas tänane seis läheneda uurimisprobleemile loovamalt, mis ei tähendnud sugugi vabanemist empiirilise uurimuse vajadusest algoritmi arendamisel. Sissejuhatuses mainitud pandora laegas pole kaugeltki suletud kuna potentsiaali kasutaja rolli vähendamiseks asendustel on veelgi, kuid rakenduse keerukuse kasvades tuleb kindlasti jälgida selle loetavust ja rakendatavust ka teistes keeltes.

Inimestega suheldes mõistame kõige paremini oma lähedasi ja sõpru vaatamata sellele, et suhtlustasand on üldjuhul kaugel formaalsest ning mida vähem tunneme oma vestluspartnerit seda rohkem sõltume standardiseeritud ning üldlevinumast kõnepruugist. Loomulikku keelt aktsepteerivate süsteemide puhul on tänase seisuga võimalused suhteliselt piiratud, pannes kasutaja olukorda kus efektiivne suhtlus toimub vaeses ning tugevalt reglementeeritud keeles. Olukorra muutumiseks peaks süsteemid olema suutelised kasutaja poolt suulist või kirjalikku informatsiooni töötlema juba sügavamal tasandil ning seda ka vabama kõnepruugi korral. Kes meist ei tahaks paluda masinal meie eest vanematele helistada või koguni lõputöö kokkuvõtte valmis kirjutada?

Summary

This work aimed at developing an applicable algorithm for effectively replacing pronouns with their antecedents in Estonian language texts. Secondary objective was to implement the algorithm through further development of Synonym-Pronoun Unification Module - SPUM.

Lack of a suiting algorithm and ripe methodology for this specific task allowed to approach the subject with more creativity - nevertheless, importance of empirical research was not overlooked as it guaranteed reusable and veritable results.

An algorithm for replacing pronouns was developed along with a usable version of beforementioned module for replacing pronouns in texts. Though applicable, both the algorithm and module still allow for further development as there is still untapped potential to speed up replacement process and further reduce the need for user interaction. Main breakthrough described in this work was including a method for finding and synthesizing(using the tense/form/etc of the pronoun) possible antecedents for user to choose from. Hopefully after extended use the logging capabilities of SPUM will provide provide access to more interesting insight about connections between pronouns and their antecedents.

When communicating with eachother we often understand our friends and people close to us better despite the usual lack of formality in communication. The less we know about our conversation partner the more we rely on(at least in some way) standardized choice of words. Systems capable of interacting with the user at this point are somewhat limited in their functionality and force a level of formality and transparency that feels unnatural- as we often find ourselves interacting with kids. Every small step forward in Natural Language Processing field will allow us to use more ambiguous words and simpler expressions when interacting with a system that is making use of the latest advances in the field, taking us closer to the point where we could ask a machine to let thr professor know our thesis is going to be late.

Kasutatud kirjandus

Muischnek, K., Orav, H., Kaalep, Heiki-Jaan, Õim, H. *(*2003). *Eesti keele tehnoloogilised ressursid ja vahendid: Arvutikorpused, arvutisõnastikud, keeletehnoloogiline tarkvara.* Tartu: Eesti Keele Sihtasutus.

Panttaja, E.M., Comverse Inc. (2006) *Method And System For Pronoun Disambiguation.* U.S. Pat. 7,085,709 B2.

Erelt, M., Erelt, T., Ross, K. (1997). *Eesti keele käsiraamat*. Tallinn: Eesti Keele Sihtasutus.

Koit, M. (2003). *Dialoog arvutiga eesti keeles.* Arvutimaailm, 6, 48 - 51.

Matsak, E. (2005*).* Dialogue System for Extracting Logic Constructions in Natural Language Texts. *The 2005 International Conference on Artificial Intelligence.* Las Vegas, USA.

Matsak, Erika (2009). *Loogiliste konstruktsioonide avastamine eesti laste keelest.,* Tallinna Tehnikaülikool, Infotehnoloogia teaduskond, Arvutitehnika instituut) Tallinn: TTU Press

Gaistruk, Stanislav (2011). *SPUM- moodul dialoogisüsteemile DST asesõnade ja sünonüümide töötlemiseks* (Seminaritöö). Tallinna Ülikool, Tallinn.

Müürisep, K.. (2003). *Eesti keele süntaksianalüsaatori märgenditest*. URL http://kodu.ut.ee/~kaili/papers/myyriseprakling03final.pdf (Viimati vaadatud 03.01.2012)

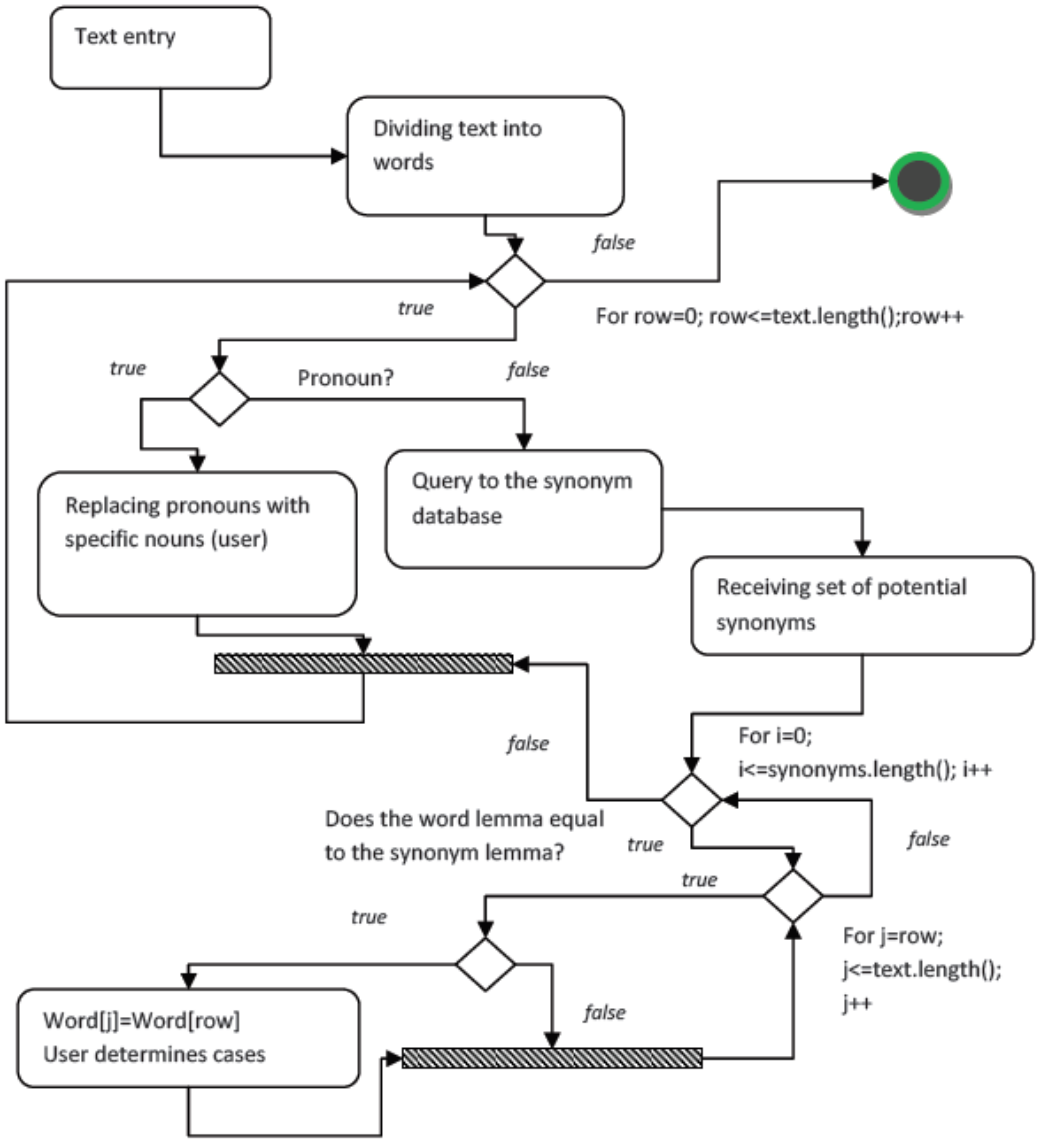
Koit, M., Roosmaa, T. (1999). *Eestikeelsete nimisõnafraaside filtreerija.* URL http://www.eki.ee/keeletehnoloogia/projektid/EstNPTool (Viimati vaadatud 03.01.2012)

Kaalep, H.-J. (1999). *ESTMORF, eesti keele morfoloogiline analüsaator ja süntesaator*. URL http://www.eki.ee/keeletehnoloogia/projektid/estmorf/ (Viimati vaadatud 03.01.2012)

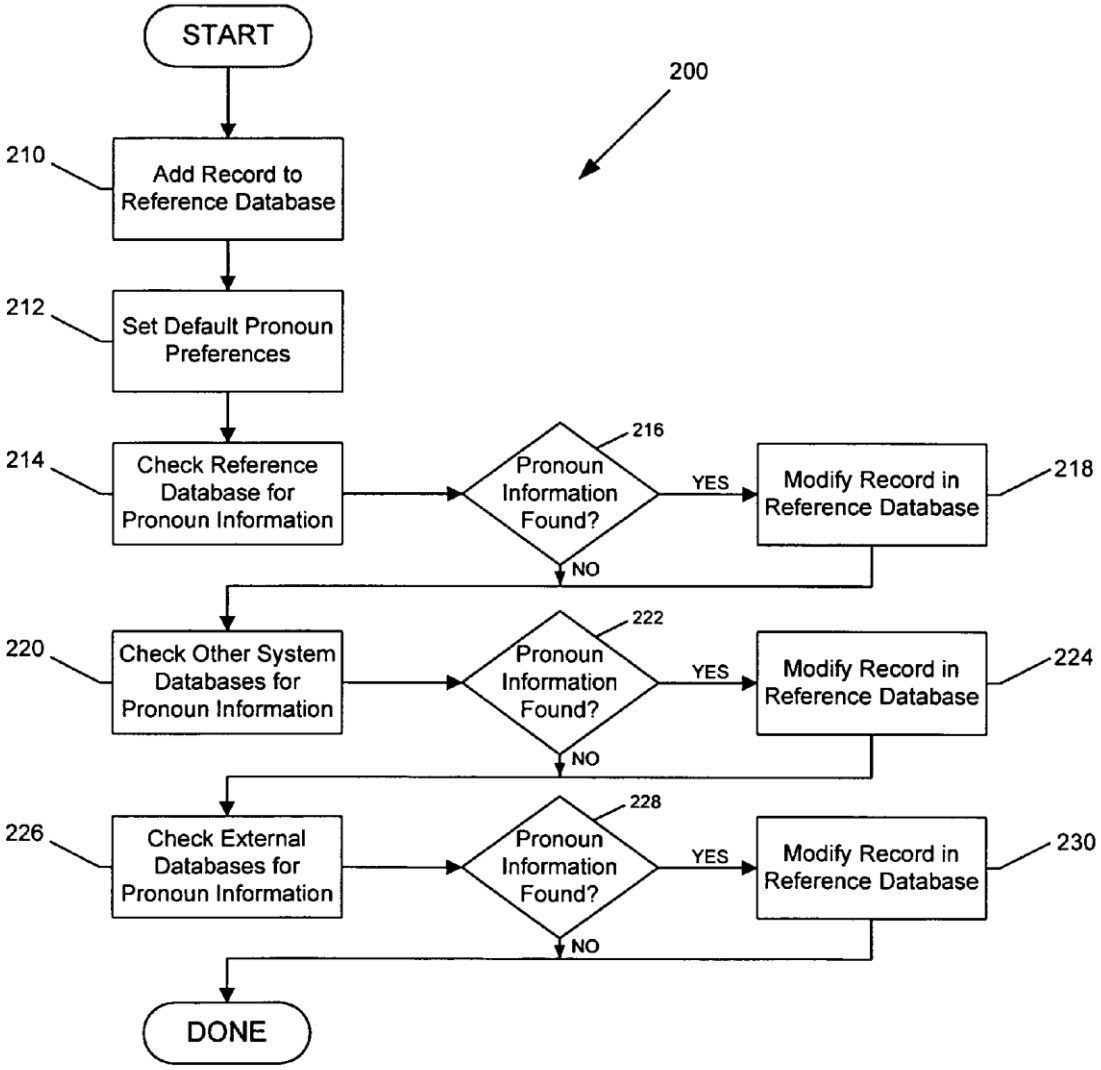
Kaalep, H.-J. (1999) TAHMM, ESTMORFi tulemuste ühestaja. *URL* http://www.eki.ee/k*eeletehnoloogia/projektid/tahmm/ (Viimati vaadatud 03.01.2012)*

Vaino, T. (1999). *ELA – Eesti keele lausestaja.* URL http://www.eki.ee/keeletehnoloogia/projektid/ela/ela.html (Viimati vaadatud 03.01.2012)

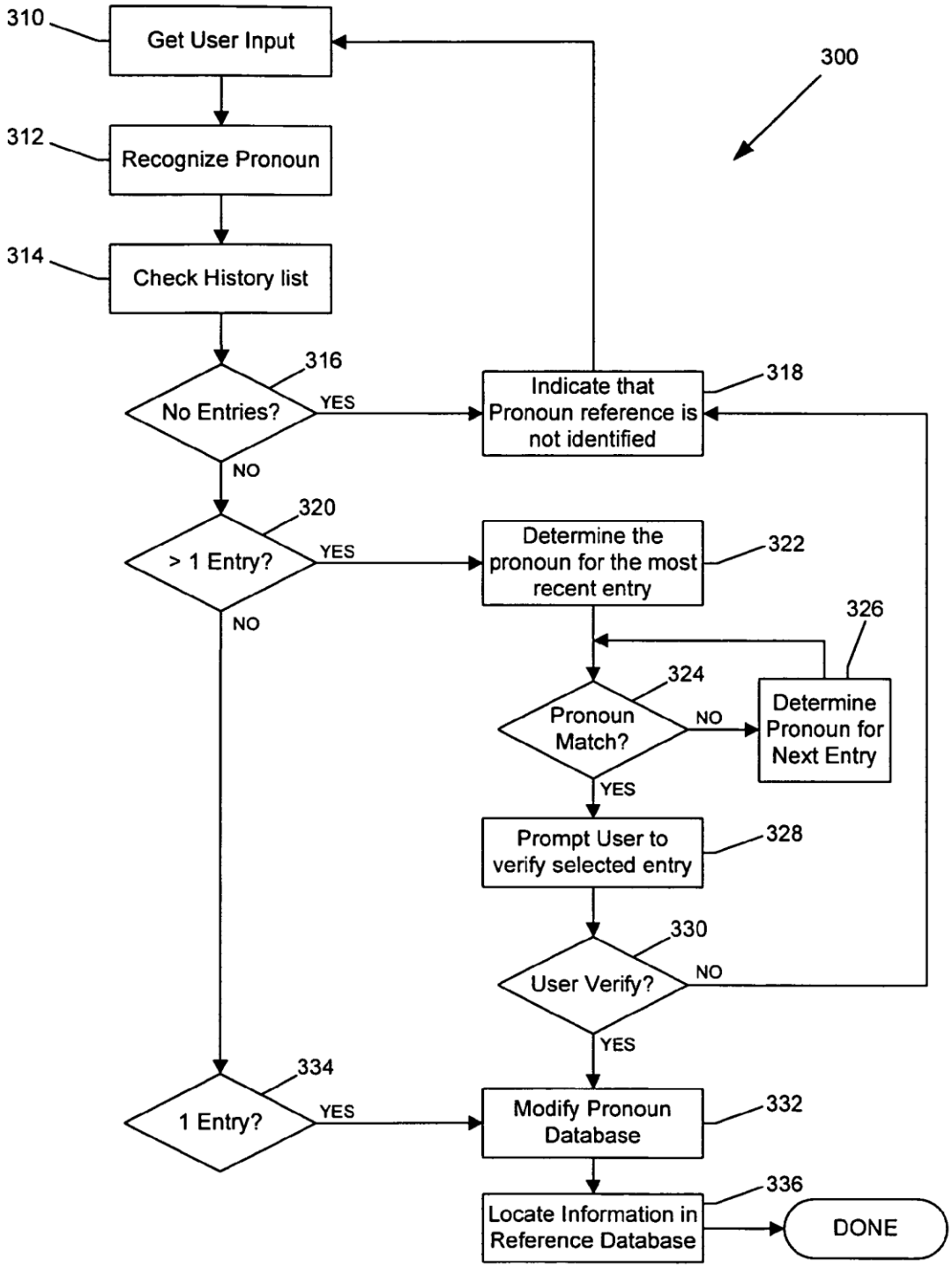
LISAD



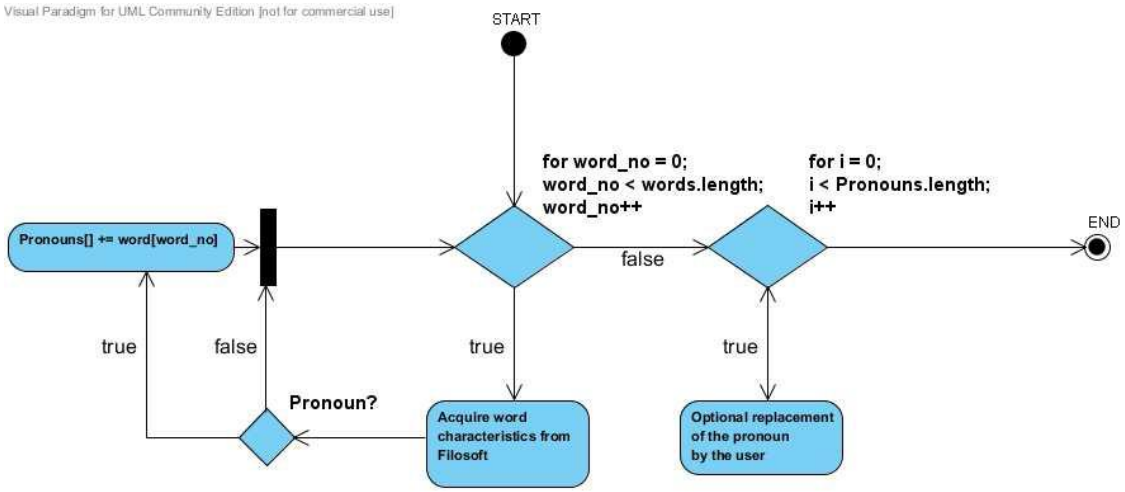
Lisa 1. Asesõnade ja sünonüümide asendamise algoritm



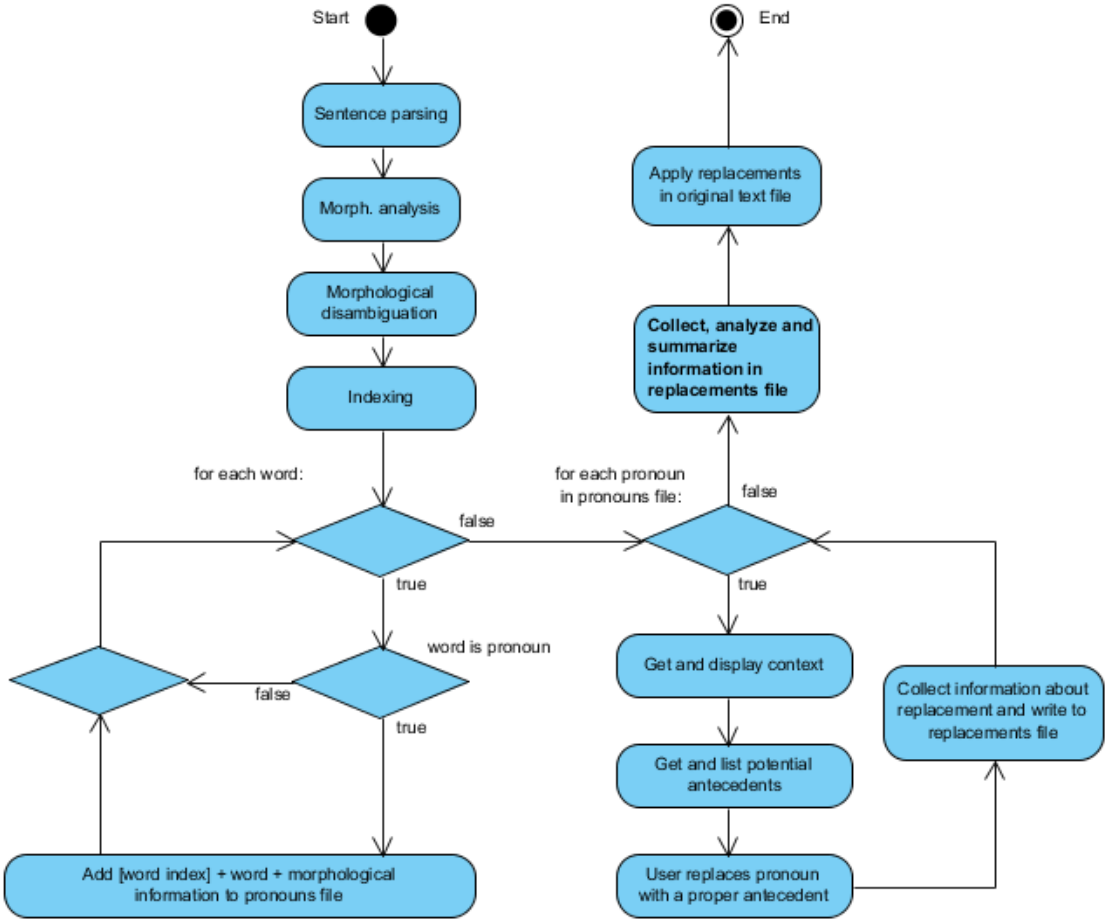
Lisa 2. Asesõnale tähenduse määramine.



Lisa 3. Kasutaja sisendi analüüs ja süsteemipoolne kasutajaga suhtlemine.



Lisa 4. Mooduli SPUM esimese versiooni loomisel kasutatud algoritm



Lisa 5. Täiustatud algoritm asesõnade asendamiseks.